UTS UNIVERSITY OF TECHNOLOGY SYDNEY

Causal Model for Recommendation

by Dianer Yu

Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

under the supervision of Professor. Guandong Xu and Dr. Qian Li

University of Technology Sydney
Faculty of Engineering and Information Technology

July 2025

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, *Dianer Yu*, declare that this thesis is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, 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 is supported by the Australian Government Research Training Program.

Production Note:
SIGNATURE: Signature removed prior to publication.

DATE: 19/July/2025

LOCATION: Sydney, Australia

ACKNOWLEDGMENTS

Looking back on my years at UTS—from the Diploma and Bachelor, to the Master's and PhD—I want to thank myself for never giving up, for pushing forward through every challenge, setback, and moment of doubt.

Above all, my deepest and most heartfelt gratitude goes to my wife, Jinlu Li. Your love and unwavering support have been the foundation of my strength. Your quiet sacrifices, gentle encouragement, and belief in me carried me through my most difficult days. To my dear mother-in-law, who traveled from China to Sydney during my PhD years to help care for our children—you have given me the gift of focus and peace of mind, and I am forever indebted to your kindness and selflessness. This accomplishment is not mine alone—it is the result of our shared resilience, love, and determination as a family.

I was so lucky to meet my supervisors, Prof. Guandong Xu and Dr. Qian Li, who provided thoughtful guidance, patience, and encouragement throughout my PhD journey. Coming from a non-research background, I often felt uncertain, but your wisdom and belief in my potential allowed me to grow beyond my own expectations. I am equally grateful to Prof. Angela Huo, my teaching mentor, whose thoughtful guidance and encouragement helped me gain confidence and joy in teaching. Her support has greatly enriched my academic journey and played an essential role in my development as an educator.

Thank you all for walking this journey with me.

ABSTRACT

Recommender systems (RSs) are central to digital ecosystems, offering personalized item suggestions by analyzing users' historical interactions. However, a major challenge in RSs is the prevalence of spurious correlations—misleading patterns in interaction data caused by confounding effects or external factors such as exposure mechanisms—which can obscure true user preferences and degrade recommendation quality. Fortunately, causal inference provides a principled statistical framework for modeling causal relationships between variables while accounting for confounding effects, helping identify true causal factors influencing recommendation outcomes. This thesis introduces a comprehensive causal-based framework to address spurious correlations in RSs through three core contributions: (1) the development of foundational causal models to reduce confounding biases for unbiased recommendations, (2) advanced causal techniques that improve model robustness in complex recommendation scenarios, and (3) methods leveraging causal insights to enhance recommendation explainability via counterfactual reasoning. Extensive experiments on real-world datasets demonstrate that the proposed causal models consistently outperform state-of-the-art baselines in both accuracy and interpretability. These findings validate the effectiveness of causal inference in distinguishing genuine user preferences from spurious correlations, advancing the robustness and transparency of modern recommender systems.

LIST OF PUBLICATIONS

PUBLICATIONS:

1. **Counterfactual Explainable Conversational Recommendation**
   Dianer Yu, Qian Li, Xiangmeng Wang, Qing Li, Guandong Xu.
   *IEEE Transactions on Knowledge and Data Engineering **(TKDE)**, IF 8.9, Q1, Core A*, CCF A.*

2. **A Causal-Based Attribute Selection Strategy for Conversational Recommender Systems**
   Dianer Yu, Qian Li, Xiangmeng Wang, Guandong Xu.
   *IEEE Transactions on Knowledge and Data Engineering **(TKDE)**, IF 8.9, Q1, Core A*, CCF A.*

3. **Breaking The Loop: Causal Learning To Mitigate Echo Chambers In Social Networks**
   Dianer Yu, Qian Li, Huan Huo, Guandong Xu.
   *ACM Transactions on Information Systems **(TOIS)**, Q1, Core A*, CCF A*

4. **LLM-enhanced Dual Propensity Score Estimation for Sequential Recommendation**
   Dianer Yu, Qian Li, Sirui Huang, Guandong Xu.
   *IEEE Transactions on Knowledge and Data Engineering **(TKDE)**, IF 8.9, Q1, Core A*, CCF A.*

5. **Causality-guided Graph Learning for Session-based Recommendation**
   Dianer Yu, Qian Li, Hongzhi Yin, Guandong Xu.
   *ACM International Conference on Information and Knowledge Management **(CIKM 2023)**, Core A, CCF B.*

6. **Deconfounded Recommendation via Causal Intervention**
   Dianer Yu, Qian Li, Xiangmeng Wang, Guandong Xu.
   ***Neurocomputing**, IF 6.0, Q1.*

7. **Semantics-Guided Disentangled Learning for Recommendation**
   Dianer Yu, Qian Li, Xiangmeng Wang, Zhichao Wang, Yanan Cao, Guandong Xu.
   *The Pacific-Asia Conference on Knowledge Discovery and Data Mining **(PAKDD 2022)**, Core A.*

8. **A Survey of Causal Learning in Graph-Based Recommender Systems**
   Dianer Yu, Qian Li, Balamurugan Soundararaj, Christopher Pettit, Guandong Xu.
   Under Review of *ACM Computing Survey **(CSUR)**, IF 28.0, Q1, Core A*, CCF A.*

9. **MGPolicy: Meta Graph Enhanced Off-policy Learning for Recommendations**
   Xiangmeng Wang, Qian Li, **Dianer Yu**, Guandong Xu.
   *ACM SIGIR Conference on Research and Development in Information Retrieval **(SIGIR 2022)**, Core A*, CCF A.*

10. **Off-policy Learning over Heterogeneous Information for Recommendation**
    Xiangmeng Wang, Qian Li, **Dianer Yu**, Guandong Xu.
    *ACM The Web Conference **(WWW 2022)**, Core A*, CCF A.*

11. **Counterfactual Debiasing for Multi-behavior Recommendations**
    Sirui Huang, Xiangmeng Wang, Qian Li, **Dianer Yu**, Guandong Xu, Qing Li.
    *International Conference on Database Systems for Advanced Applications **(DASFAA 2024)**, Core B, CCF B.*

12. **Causal Disentanglement for Semantics-Aware Intent Learning**
    Xiangmeng Wang, Qian Li, **Dianer Yu**, Peng Cui, Guandong Xu.
    *IEEE Transactions on Knowledge and Data Engineering **(TKDE)**, IF 8.9, Q1, Core A*, CCF A.*

13. **Reinforced Path Reasoning for Counterfactual Explainable Recommendation**
    Xiangmeng Wang, Qian Li, **Dianer Yu**, Qing Li, Guandong Xu.
    *IEEE Transactions on Knowledge and Data Engineering **(TKDE)**, IF 8.9, Q1, Core A*, CCF A.*

14. **Causal Time-aware News Recommendations with Large Language Models**
    Sirui Huang, Qian Li, Haoran Yang, **Dianer Yu**, Qing Li, Guandong Xu.
    *ACM Transactions on Information Systems **(TOIS)**, Q1, Core A*, CCF A*

15. **Counterfactual Explanation for Fairness in Recommendation**
    Xiangmeng Wang, Qian Li, **Dianer Yu**, Qing Li, Guandong Xu.
    *ACM Transactions on Information Systems **(TOIS)**, Q1, Core A*, CCF A.*

16. **Neural Causal Graph Collaborative Filtering**
    Xiangmeng Wang, Qian Li, **Dianer Yu**, Qing Li, Guandong Xu.
    ***Information Sciences**, Q1, IF 8.1.*

17. **Constrained Off-policy Learning over Heterogeneous Information for Fairness-aware Recommendation**
    Xiangmeng Wang, Qian Li, **Dianer Yu**, Qing Li, Guandong Xu.
    *ACM Transactions on Recommender Systems **(TORS)**.*

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE **Page**

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Background

Recommender systems (RSs) play a crucial role in the current digital age of information explosion, helping users navigate through vast content options to discover their preferred items and enabling content providers to optimize the item exposure [131, 104, 218, 229]. Traditional RSs typically rely on historical interaction data, such as purchasing, liking, or viewing behaviors, to infer user preferences and generate relevant recommendations, as shown in Figure 1.1. However, a fundamental challenge in these systems lies in their heavy reliance on observed correlational patterns within interaction data, which often contain spurious correlations to obscure users' true preferences and degrade system explainability [129]. Spurious correlations refer to misleading statistical relationships between variables, where observed user-item interactions may be influenced by external factors rather than reflecting users' true preferences [50, 26, 198, 100]. For instance, an item displayed in a prominent position may attract more clicks regardless of its relevance to users, creating misleading patterns that distort preference modelling. Such misleading spurious correlations can simultaneously degrade recommendation performance by promoting irrelevant items to users, and reduce system explainability by obscuring the true causal factors driving user behaviors. Therefore, effective identification and mitigation of spurious correlations is crucial for more robust recommendations.

Most of the existing approaches address the problem of spurious correlations in recommender systems through two main directions. The first direction is to leverage auxil-

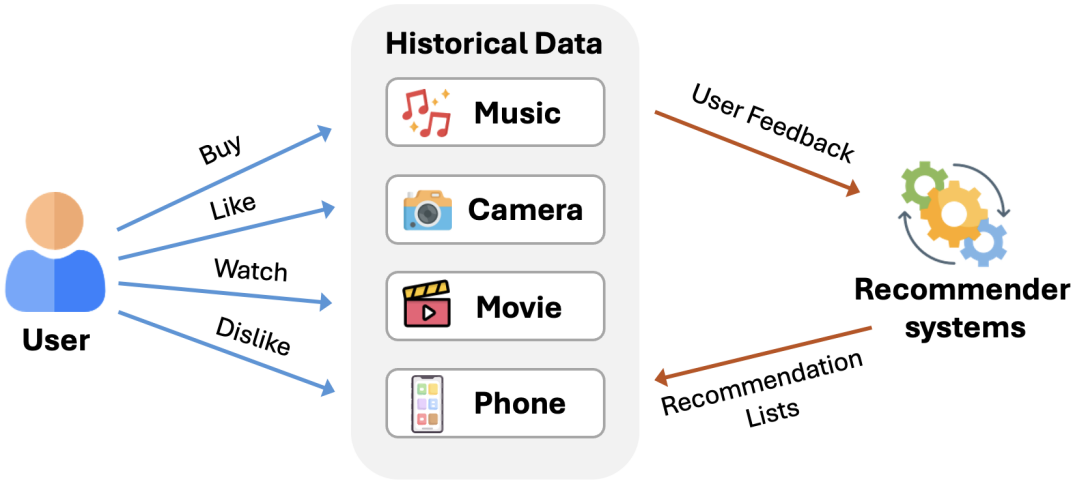


Figure 1.1: The general workflow of recommender systems.

iary information, incorporating supplementary data such as user demographics and contextual signals to enhance the modelling of user preferences [124, 161, 230]. While this approach aims to mitigate spurious patterns by providing additional sources of information, it remains vulnerable to biases inherent in the auxiliary data itself, which can inadvertently introduce new spurious correlations rather than eliminate them. The second direction focuses on employing advanced modelling techniques, particularly deep learning and graph neural networks, to capture complex non-linear relationships between users and items [181, 190, 244, 214, 66]. Although these advanced models demonstrate strong capabilities in learning intricate interaction patterns, they fundamentally rely on observed correlations in historical data to generate recommendations. This dependence makes them prone to learning and amplifying existing spurious correlations, rather than effectively distinguishing between genuine user preferences and misleading spurious patterns [104, 56]. Given the persistent limitations of existing approaches, it is imperative to investigate a principled framework to fundamentally and systematically address the root causes of spurious correlation towards more robust recommendations.

Fortunately, causal inference provides a principled statistical framework for modelling the causal relationships between variables, helping to identify and isolate the true factors that influence outcomes [129, 229]. As shown in Figure 1.2, the overall goal of causal inference is to identify the true root cause from multiple potential factors (e.g., smoking, age, diet etc.) that affect the outcome (e.g., lung cancer). Generally, through rigorous causal techniques such as causal intervention and propensity scoring, causal inference systematically evaluates the impact of different factors by comparing outcomes between treatment

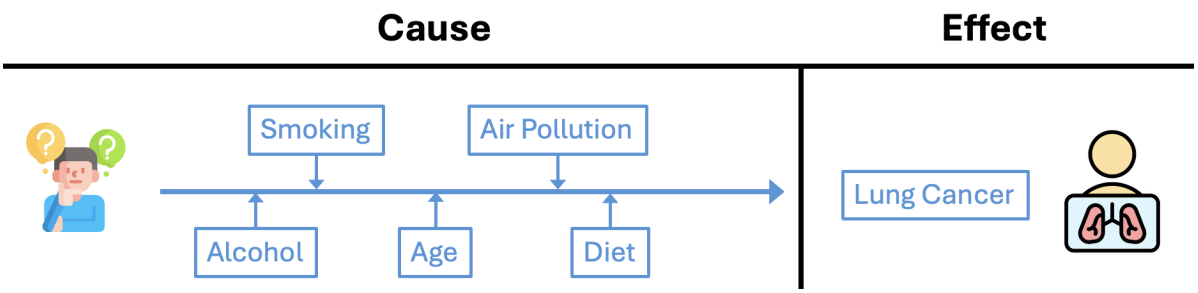


Figure 1.2: A toy example of causal analysis in which multiple factors impact the outcome, highlighting the need to identify the true causality for effective treatment and prevention.

and control groups under carefully controlled conditions [129, 192]. This controlled comparison approach enables the isolation of each factor's individual effect while accounting for other influencing variables, similar to how clinical trials assess drug effectiveness by comparing treatment groups against placebo. Through this methodical process, we can precisely measure true cause-effect relationships among variables without distortion from external influences that may confound observational studies. In the context of recommendations, causal inference can differentiate true causality (e.g., user interest leads to clicks) from spurious correlation (e.g., item location leads to clicks) by treating item features or recommendation strategies as treatments and user behaviors as outcomes. Technically, causal inference provides several complementary analytical tools for systematically identifying the true causality behind interactions: causal graphs for visually mapping relationships among variables and identifying potential confounders [51, 131]; causal intervention for estimating the causal effect of treatments on user behaviors or system outcomes through causal graphs [191, 186, 239]; do-calculus for providing mathematical rules for valid causal estimation among variables on causal graphs [34, 164]; and counterfactual reasoning for enabling explorations of hypothetical scenarios [158, 243]. Together, these causal techniques fundamentally address the problem of spurious correlations by identifying true causal relationships between variables, leading to more robust recommendations.

This thesis aims to develop novel causal-based recommender models to fundamentally address the issues caused by spurious correlations in recommendations. Leveraging causal inference's unique strengths in identifying true cause-effect relationships among variables, we conduct in-depth research across three interconnected aspects. The first aspect focuses on developing causality-based models to eliminate confounding biases introduced by spurious correlations, enabling more accurate estimation of user preferences. Building upon this foundation, we design advanced causal approaches to enhance recommendation robustness in complex real-world relational scenarios, where higher-order interaction pat-

terns make spurious correlations particularly challenging to address. Finally, we extend our designed causality-based models to improve the explainability of the system by utilizing causal reasoning to generate transparent and interpretable recommendations. Our extensive experiments on multiple real-world datasets demonstrate that the systematic integration of causal inference successfully achieves more accurate, unbiased, and explainable recommendations. By establishing the causal-guided framework, our work advances the state-of-the-art recommendations and provides a solid foundation for future developments in the field.

1.2 Existing Research and Limitations

Traditional recommendation techniques, such as collaborative filtering, content-based methods, and hybrid approaches, have been widely deployed across diverse domains to infer user preferences based on historical interactions [147, 163, 26, 198]. However, these methods primarily rely on observed correlations between users and items, making them susceptible to spurious patterns that do not necessarily reflect true user interests [126, 121, 149, 78, 117]. Such correlation-based reliance can lead to biased preference estimations, as user interactions are often shaped by external factors, such as exposure bias, popularity effects, and temporal dynamics, which may obscure genuine preferences rather than revealing them [74, 90, 64]. Consequently, traditional models struggle to distinguish whether an observed user-item interaction stems from actual user interest or external influences, leading to suboptimal recommendations that reinforce existing biases rather than capturing users' true needs. This limitation becomes even more pronounced in real-world recommendation scenarios, where contextual and social factors interact in complex ways, making it increasingly difficult to disentangle genuine user preferences from spurious correlations and uncover the causal mechanisms underlying user behavior.

Advanced techniques, particularly deep learning and graph neural networks (GNNs), have emerged as sophisticated approaches for handling the spurious correlations in the modern complex recommendation scenarios [223, 228, 214, 249, 190, 175]. Generally, these advanced techniques enhance recommendation performance by modeling user-item relationships through non-linear transformations and structural representations [198, 39, 210]. Technically, deep learning architectures automatically extract hierarchical features of user interests from massive interaction data, capturing intricate patterns beyond simple linear correlations [244]. Graph neural networks model higher-order connections and information flow across the entire interaction network through structured information propaga-

tion, capturing complex dependencies overlooked by traditional methods [206]. However, despite their improved preference modelling capabilities, they still fundamentally rely on observed correlations for inference rather than causality, failing to eliminate the problem of spurious correlations at its root.

Beyond accuracy, spurious correlations pose a significant challenge to the interpretability of recommender systems by obscuring the true factors driving user behaviors [126, 1]. These misleading interaction patterns make it difficult for systems to generate accurate and meaningful explanations, reducing transparency and limiting the reliability of recommendation outputs [118, 222]. In e-commerce scenarios, for example, high user engagement with certain items may be mistakenly attributed to genuine interest, while the actual driving factors could be external influences such as prominent placement or seasonal trends, leading to misinterpretation of user behavior. To improve recommendation interpretability, existing research has explored various strategies, such as explanation based on user comments [65, 21, 61], explanation based on feature importance [63, 22, 209, 103] or reasoning over knowledge graphs [28, 252, 208, 189]. However, most of these methods rely on correlation-based analysis, which only captures co-occurring patterns rather than uncovering true causal mechanisms driving user behaviors. This limitation makes it difficult to determine if a user's interaction with an item stems from true interest or from confounding factors like position bias. Therefore, it is crucial to develop methods that systematically address spurious correlations at their root, enabling recommender systems to provide accurate and meaningful explanations based on true causal mechanisms rather than misleading spurious correlations.

Recently, causal inference has emerged as a principled statistical framework to systematically identify and understand true causal relationships among variables [99, 51, 91] Unlike traditional and advanced techniques that rely on correlational patterns to improve performance, causal inference aims to establish explicit causal structures between multiple variables to effectively differentiate correlation from causation [129, 191]. By systematically analyzing cause-effect relationships, causal inference can effectively mitigate spurious correlations by identifying genuine influencing factors while controlling for confounding biases that create misleading patterns in the data. Instead of merely leveraging observed associations, it ensures that recommendations are based on causal mechanisms, leading to more robust interpretable outcomes. Technically, the comprehensive framework of causal inference can be broadly divided into several main techniques: (1) causal graphs [51] serve as the foundation by providing a formal representation of hypothesized causal relationships between users, items, and contextual factors in recommendations; (2) building upon

causal graphs, causal interventions [186] assess how specific changes in the system affect user behavior by simulating controlled experiments; (3) counterfactual reasoning [158] extends this intervention capability by enabling the evaluation of hypothetical scenarios without requiring actual experimentation, helping understand potential outcomes under different conditions; (4) do-calculus [34] provides the mathematical framework necessary for rigorously estimating causal effects from observational data, translating theoretical causal relationships into practical estimation strategies; and (5) propensity score [250] mitigates selection and exposure biases through balancing covariate distributions across treatment groups, ensuring more reliable causal estimates from observational data. Together, these interrelated techniques form a comprehensive causal framework that enables more robust estimation of the causal factors driving user behaviors in recommendations.

Despite the growing interest and recent progress in causal-based recommendation, current approaches exhibit important limitations. For instance, early causal recommender models [18, 242, 227] introduced the use of propensity score adjustment and back-door correction to address exposure bias, but these solutions often assume that all relevant confounders are observed and correctly specified, which rarely holds in real-world scenarios. Other causal models [5, 191, 68] leverage counterfactual reasoning to simulate interventions, but are limited to binary treatments and may not handle multi-faceted or sequential interventions effectively. Recent works on causal graph neural networks [197, 44] have demonstrated the integration of causal reasoning into graph-based recommendation, but often require an accurate prior knowledge of the causal graph structure, which is usually partially observable in practice. Moreover, these methods are specialized for either debiasing or explainability, and seldom provide a unified framework capable of systematically addressing interpretability and performance in complex recommendation settings. Therefore, this thesis aims to bridge these research gaps by developing novel causal models that (i) account for both observed and hidden confounders, (ii) adapt to complex real-world recommendation scenarios including high-order and dynamic dependencies, and (iii) jointly optimize both recommendation robustness and interpretability through principled causal inference.

1.3 Research Questions

Spurious correlations in recommender systems—arising from confounding factors, algorithmic biases, and external influences—undermine both performance and explainability. These challenges weaken the robustness of recommendations. Through a systematic anal-

ysis, we identify three research questions that build on one another, aiming to address spurious correlations using causal inference principles.

The first challenge is to distinguish genuine user preferences from spurious correlations in interaction data. Although collaborative filtering, deep learning, and graph-based methods are widely used to model user preferences, they often overlook confounding effects, resulting in suboptimal recommendations. This motivates our first research question:

- **RQ1**: How to effectively mitigate confounding effects caused by spurious correlations in recommendations through causal inference?

By answering RQ1, we aim to develop foundational causal models that can identify and mitigate confounding biases in interaction data, thereby enabling more reliable estimation of users' true preferences.

Subsequently, real-world recommendation scenarios often involve complex structures in which multiple types of spurious correlations interact and reinforce each other. Such complex recommendation scenarios require more sophisticated causal models that can handle higher-order dependencies and complex confounding effects, which leads to our second research question:

- **RQ2**: How to enhance recommender system robustness against spurious correlations in complex recommendation scenarios through causal inference?

By answering RQ2, we aim to develop more advanced causal models to address the challenges posed by complex recommendation environments, where spurious correlations manifest in more intricate ways.

Beyond improving accuracy, the ability to provide transparent explanations for recommendation decisions becomes crucial as recommender systems increasingly influence user choices across various domains. Traditional correlation-based approaches struggle to provide meaningful explanations since they cannot distinguish whether recommendations stem from genuine user preferences or spurious patterns. By leveraging the causal insights from causal models, we investigate our third research question:

- **RQ3**: How to enhance the interpretability of recommender systems by disentangling spurious correlations from true user preferences through causal inference?

By answering RQ3, we aim to develop explainable recommendation mechanisms that not only provide accurate recommendations but also offer transparent rationales based on genuine causal factors rather than misleading correlations.

The above three research questions form a progressive framework that systematically addresses spurious correlations in recommender systems through causal inference. By first establishing foundational debiasing techniques (RQ1), then extending them to complex scenarios (RQ2), and finally leveraging these insights to improve explainability (RQ3), we aim to advance the field of recommender systems toward more robust and interpretable recommendations.

1.4 Thesis Contributions

This thesis makes three interconnected contributions that systematically advance recommender systems by addressing spurious correlations through causal inference. Each contribution directly maps to one of our research questions while building on previous insights:

1. First, we develop foundational causality-based models to mitigate confounding effects arising from spurious correlations in recommendations data (Chapter 3). By integrating causal adjustment and propensity scoring, we develop LDPE and CCR frameworks capable of distinguishing true user preferences from misleading spurious correlations. These causality-based models greatly improve recommendation performance over state-of-the-art baselines on multiple real-world datasets, verifying the effectiveness of causal inferences in generating unbiased recommendations.

2. Second, we extend causal inference to tackle spurious correlations in complex recommendation environments. Specifically, we develop three advanced causality-based models targeting distinct complex recommendation scenarios (Chapter 4): CGSR for blocking shortcut paths in complex user-item session graphs, GCRec for addressing multiple confounders in high-order interaction networks, and CEDA for mitigating echo chamber effects in complex social networks. By systematically integrating GNNs with causal interventions, these causality-based models effectively address spurious correlations in complex recommendation scenarios while maintaining computational efficiency.

3. Third, we leverage causal insights from causality-based models to improve the recommendation explainability (Chapter 5). Through causal reasoning, we develop SeDLR and CECR frameworks capable of uncovering the underlying causal mechanisms behind recommendations. These causality-based models effectively differentiate spurious correlations from true causal drivers of recommendation decisions, thus significantly improving the interpretability of recommendations.

In summary, each contribution directly addresses a key research question. The first contribution establishes foundational causal models to mitigate confounding effects in recommendation systems. The second contribution extends the causal inference principles to enhance model robustness in complex recommendation scenarios. The third contribution leverages the causal insights from causal models to improve recommendation explainability. Collectively, these contributions advance recommender systems by fundamentally addressing spurious correlations using through causal inference principles.

1.5 Thesis Structure

This thesis is organized into six chapters that progressively develop causality-based models to fundamentally address the issue of spurious correlations in recommender systems:

- Chapter 1 introduces the research background, research questions, and our contributions, providing a comprehensive overview of how this thesis advances the field of recommender systems through causal inference principles.
- Chapter 2 presents a thorough literature review on existing recommendation techniques, analyzing traditional methods, advanced approaches, and the emergence of causal inference as a principled solution for addressing spurious correlations in RSs.
- Chapter 3 addresses Research Question 1 by developing causality-based models to mitigate confounding effects caused by spurious correlations, ensuring unbiased and reliable recommendations.
- Chapter 4 addresses Research Question 2 by enhancing the robustness of the causality-based model to distinguish spurious correlations from users' true preferences in complex recommendation scenarios.
- Chapter 5 addresses Research Question 3 by developing interpretable causality-based models to uncover the causal mechanisms behind recommendations, enhancing interpretability and transparency in decision-making processes.
- Chapter 6 concludes the thesis by summarizing key findings, discussing practical implications, and outlining promising research directions in the field of recommender systems.

CHAPTER 2

LITERATURE REVIEW

This chapter provides a comprehensive literature review of existing recommendation techniques through three interconnected perspectives: traditional recommendation techniques, advanced recommendation techniques, and causal-based recommendation techniques. Specifically, the review begins by examining traditional recommendation techniques, such as collaborative filtering and content-based methods, highlighting their reliance on observed correlations and limitations in distinguishing true user preferences from spurious patterns. It then explores how modern advanced techniques, such as deep learning and graph-based models, have attempted to mitigate spurious correlations by capturing more complex user-item relationships, yet often remain constrained by their correlation-based nature. Finally, the chapter delves into causal inference techniques, which provide a principled framework for disentangling spurious correlations from true causal relationships, thereby enhancing both the robustness and interpretability of recommender systems. By establishing explicit cause-and-effect relationships, causal approaches offer a promising solution to the limitations identified in both traditional and advanced recommendation techniques. Through this structured analysis, the chapter establishes a clear connection between the historical development of RSs and the emerging role of causal learning as a promising solution to the challenges posed by spurious correlations.

2.1 Recommender Systems

Recommender systems (RSs) are information filtering mechanisms, which play a pivotal role in today's digital ecosystems by providing personalized and relevant content tailored to user preferences [176, 240, 198, 122, 247, 147]. They operate across diverse domains such as e-commerce, entertainment, and social media platforms, analyzing user-item interaction data (e.g., clicks, purchases, ratings) to generate recommendations that enhance user engagement and satisfaction. The business impact of these systems is substantial, with Netflix attributing 80% of its viewed content to recommendations [53] and Amazon reporting that 35% of its revenue stems from its recommendation engine [76]. A fundamental challenge in recommender systems is the presence of spurious correlations within historical interaction data. These misleading statistical patterns arise when observed user-item interactions are influenced by external factors like exposure mechanisms rather than reflecting users' true preferences [181, 67, 248]. For instance, items displayed in prominent positions may attract more clicks regardless of their relevance to users, creating a self-reinforcing feedback loop where the system continues promoting already-visible items while overlooking potentially better matches. Such spurious correlations can simultaneously degrade recommendation performance by promoting irrelevant items and reduce system explainability by obscuring the true factors driving user behaviors.

Traditional approaches to addressing this challenge have attempted to leverage auxiliary information or employ advanced modeling techniques like deep learning and graph neural networks. However, these methods fundamentally rely on observed correlations in historical data rather than identifying true causal relationships, making them vulnerable to perpetuating and even amplifying existing biases. Instead, causal inference offers a principled statistical framework, which can help systematically identify and mitigate spurious correlations by modeling the causal mechanisms underlying user preferences. This thesis proposes novel causal models that effectively address spurious correlations in recommender systems through three interconnected contributions: (1) developing foundational causal models to mitigate confounding effects, (2) enhancing recommendation robustness in complex scenarios through extended causal interventions, and (3) improving system explainability through counterfactual reasoning of causal inference. In the following, we will describe traditional recommendation approaches that established the field's foundation, advanced recommendation techniques that enhanced performance through sophisticated modeling, and causality-based recommendation techniques that fundamentally address spurious correlations through principled inference mechanisms.

2.2 Traditional Recommendation Techniques

Traditional recommendation techniques aim to mitigate spurious correlations in recommendations by leveraging user behavior patterns and item characteristics to generate unbiased recommendations [147, 42, 85]. Generally, traditional techniques can be broadly categorized into three main approaches: collaborative filtering, content-based methods, and hybrid methods [163, 86, 26]. Each approach demonstrates unique strengths and different limitations. In the following, we examine each of these approaches in detail, discussing their methodologies, applications, advantages, and limitations.

Collaborative Filtering (CF) operates on the assumption that users with similar past behaviors will share similar preferences in the future, which is one of the most widely adopted recommendations strategies [148, 125, 154]. As shown in Figure 2.1, User A and User B have the same historical data, both watched Movie 1 and Movie 2. Therefore, CF assumes that they have similar preferences, so if User B later watched Movie 3, then CF will recommend Movie 3 to User A. The CF typically encompasses two complementary approaches to capture and leverage the different user behavioral patterns. First, memory-based CF analyzes comprehensive user-item interaction histories to identify similar users and items, generating recommendations by considering the full scope of historical behaviors rather than isolated interactions [236]. By leveraging diverse interaction patterns across the user base, it can help mitigate spurious correlations to some extent by incorporating rich contextual information. However, its reliance on raw interaction data makes it prone to biases, as frequently interacted items tend to be recommended more often, potentially reinforcing popularity bias and exposure effects. Second, model-based CF employs sophisticated predictive models, particularly matrix factorization techniques, to learn latent user/item representations that capture underlying preference structures [2]. These learned representations enable the system to model complex relationships, allowing for more personalized recommendations while partially reducing spurious correlations by considering user preferences beyond explicit interactions. However, model-based CF still relies on observed user-item interactions, meaning that latent factors can inadvertently capture spurious associations present in the training data, making it difficult to differentiate true user preferences from spurious patterns. Despite its effectiveness in capturing user preferences, CF-based methods struggle with explainability, as they lack explicit reasoning mechanisms to justify why a recommendation was made. Moreover, data sparsity issues remain persistent challenges, limiting CF's applicability in scenarios with limited user interactions or new items.

Content-Based Filtering (CBF) uses item attributes to recommend other items that are

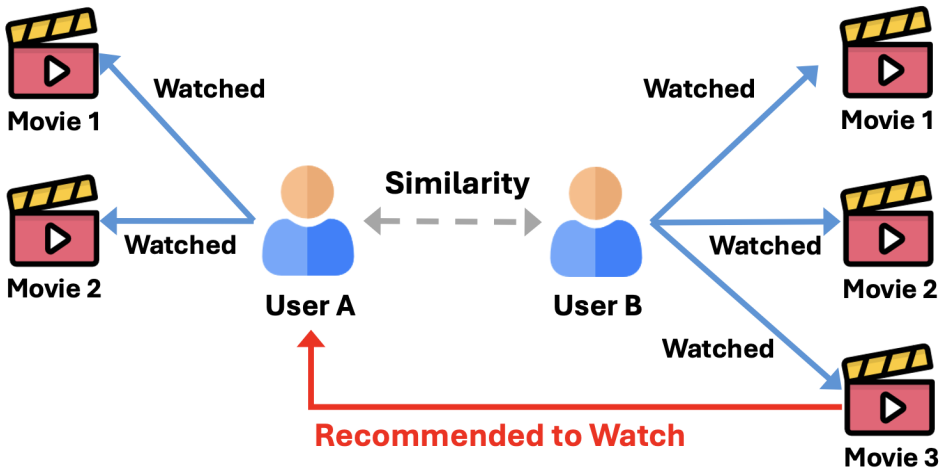


Figure 2.1: A toy example of collaborative filtering in recommendations.

similar to the user's previously preferred items, based on the user's previous actions and explicit feedback [128, 174, 152]. As shown in Figure 2.2, the CBF generates recommendations to users based on the similarity among items analyzed by the item attribute. This is particularly effective in domains with rich descriptive metadata, such as books, movies, and products. A key advantage of CBF is its ability to generate recommendations without requiring extensive user interaction histories, making it well-suited for addressing cold-start problems for new items [171, 81]. For example, if a user frequently reads science fiction books, a content-based system will recommend other books within the same genre, regardless of their past popularity. This approach enables highly personalized recommendations, as users receive suggestions based on specific attributes that align with their past preferences. However, CBF also presents several challenges. First, it tends to over-specialize recommendations, as it primarily suggests items that are similar to those the user has already interacted with, potentially limiting diversity and exploration. This can result in redundant recommendations, reinforcing established preferences while failing to introduce users to novel or diverse content. Second, content representations are crucial, and the effectiveness of CBF depends on the quality and availability of item metadata. In domains where item features are not well-defined or difficult to extract, content-based methods may struggle to generate meaningful recommendations. Finally, CBF remains correlation-based, meaning that it does not explicitly model causal relationships between user interests and item characteristics, making it susceptible to spurious associations. To overcome these limitations, more advanced approaches incorporate context-aware learning and causal inference techniques to enhance the robustness of content-based recommendations.

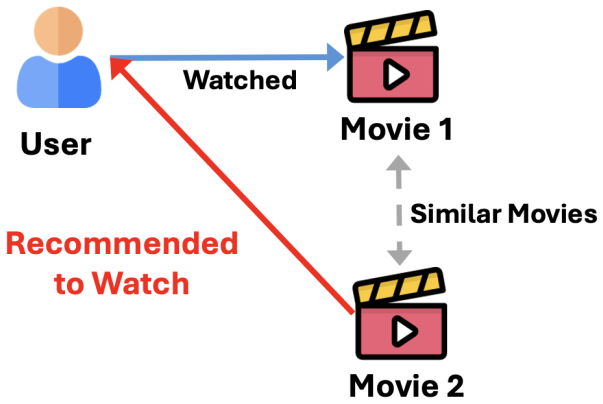


Figure 2.2: A toy example of content-based filtering in recommendations.

Hybrid recommendation strategies aim to overcome the limitations of collaborative filtering and content-based filtering by combining their strengths [17, 83, 8]. By integrating multiple techniques, hybrid models provide more robust, accurate, and adaptive recommendations, making them particularly useful for scenarios involving data sparsity, cold-start issues, and complex preference patterns. As shown in Figure 2.3, there are several common hybrid strategies: (1). Weighted Hybrid, which combines the outputs of multiple recommendation models by assigning dynamic or fixed weights [45, 14]. This approach balances collaborative filtering and content-based recommendations, allowing for more adaptive preference modeling. (2). Switching Hybrid, which dynamically selects the most suitable recommendation technique based on data availability or user behavior [13, 52]. For instance, in cold-start situations, the system may rely on content-based filtering, while in data-rich scenarios, it may favor collaborative filtering. (3). Feature Combination Hybrid, which merges multiple data sources, integrating user-item interaction patterns, content attributes, and contextual factors into a unified model [169, 109]. This method allows for richer representation learning and can improve both recommendation accuracy and adaptability. While hybrid methods enhance performance and reduce spurious correlations to some extent, they also introduce greater model complexity. Integrating multiple techniques requires careful tuning to balance different recommendation signals, and the added computational cost may impact scalability. Furthermore, while hybrid models combine different perspectives, they still rely on correlation-based learning, meaning that spurious associations can persist. Without explicitly modeling causal structures, hybrid approaches remain limited in disentangling true user preferences from confounding factors. To further improve recommendation quality, incorporating causal inference techniques within hybrid models can help mitigate spurious correlations, ensuring that recommen-

dations are based on genuine preference signals rather than misleading patterns.

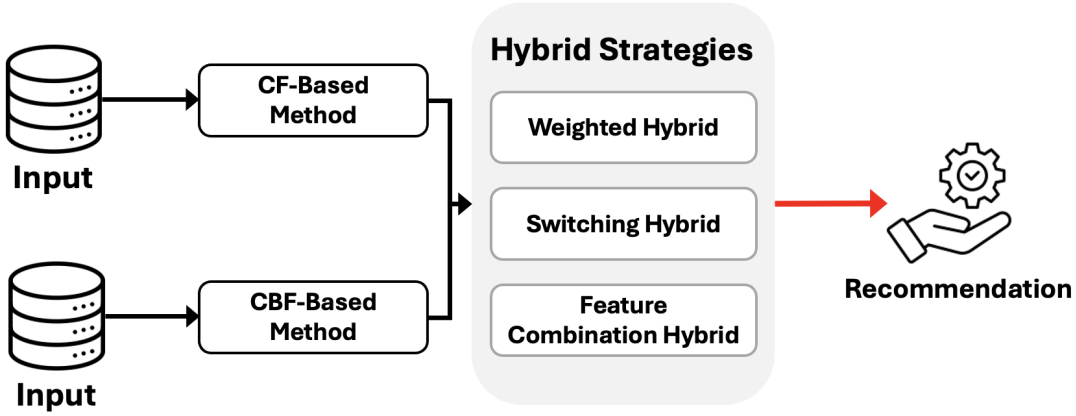


Figure 2.3: A toy example of hybrid strategy in recommendations.

In summary, traditional recommendation techniques have played a crucial role in improving recommendation accuracy by leveraging user behavior patterns, item attributes, and hybrid modeling approaches. Collaborative filtering captures preference similarities from user interactions, content-based filtering recommends items based on intrinsic attributes, and hybrid methods integrate multiple strategies to enhance robustness. However, these methods remain fundamentally correlation-based, making them susceptible to spurious correlations that distort recommendation quality. Despite improving performance, these approaches lack causal reasoning, making it difficult to distinguish true user preferences from confounding factors such as popularity bias, exposure effects, or data sparsity. Additionally, they struggle with explainability, as recommendations are often derived from latent factors or observed associations without explicit causal justifications. To address these challenges, more advanced techniques are needed to systematically identify and mitigate spurious correlations, ensuring that recommendations are not only accurate but also interpretable in complex real-world scenarios.

2.3 Advanced Recommendation Techniques

While traditional recommendation methods provide a strong foundation, they often struggle to address spurious correlations that arise in complex high-order interactions and relational structures [50, 249, 245]. To overcome these limitations, advanced recommendation techniques employ sophisticated modeling approaches that extend beyond simple user-item correlations to capture deeper dependencies in user preferences. These approaches

can be broadly categorized into five main directions: First, auxiliary information-based methods integrate contextual data such as user demographics and item attributes to enrich preference modeling, thereby addressing data sparsity challenges inherent in traditional approaches. Second, deep learning-based methods leverage neural networks to learn complex non-linear relationships from large-scale data, capturing intricate patterns that traditional methods might overlook. Third, graph-based models utilize network structures to represent high-order user-item relationships, enabling more comprehensive modeling of complex interaction patterns. Fourth, sequential and session-based recommendations capture temporal dynamics by modeling user preference evolution over time, accounting for the dynamic nature of user interests. Finally, conversational systems enable interactive multi-turn dialogues to refine user preferences in real-time, providing personalized recommendations through active user engagement. Each of these techniques provides different advantages in handling complex dependencies and user behaviors, contributing to more robust recommendation systems. The following sections examine these approaches in detail, discussing their methodologies, applications, advantages, and limitations.

Auxiliary information-based methods significantly enhance recommendation robustness by systematically incorporating additional data sources beyond simple user-item interactions, representing an advanced approach to addressing spurious correlations in complex recommendation scenarios [124, 161, 230]. As shown in Figure 2.4, these methods leverage diverse contextual information, such as temporal patterns, spatial relationships, user demographics, and social network connections, to address the fundamental data sparsity challenges inherent in traditional recommendation approaches. The integration of rich contextual data enables these methods to uncover more nuanced patterns in user preferences and item relationships that might be overlooked when relying solely on interaction data [3, 173]. For example, context-aware recommender systems analyze situational factors such as temporal and location data to generate recommendations that align more precisely with users' current contexts. Similarly, social recommendations exploit the wealth of information available in social networks, such as user-user relationships and influence patterns, to improve recommendation accuracy. In other words, by considering such a broader contextual framework, auxiliary information-based methods develop a more comprehensive understanding of user preferences, leading to recommendations that reflect not only historical interactions but also the complex environmental factors that influence user behavior [159, 114]. These methods can effectively identify and mitigate certain types of spurious correlations by introducing orthogonal information sources that are less susceptible to the same biases as interaction data. For instance, when popularity bias creates misleading cor-

relation patterns in historical interactions, demographic or social network data can provide alternative validation mechanisms to verify whether these patterns genuinely reflect user preferences or merely exposure effects. This multifaceted approach results in more relevant and reliable recommendations, ultimately enhancing both recommendation quality and user satisfaction through a more complete understanding of the recommendation context. However, while auxiliary information-based methods represent a significant improvement over traditional approaches, they may still be limited in their ability to systematically identify and address all forms of spurious correlations, as they fundamentally rely on correlative rather than causal analytical frameworks. This limitation motivates our exploration of causal inference as a more principled approach to disentangling true user preferences from misleading interaction patterns in subsequent chapters.

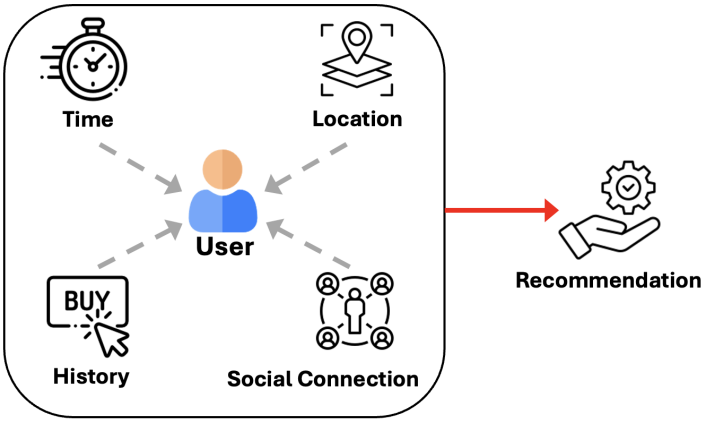


Figure 2.4: A toy example of auxiliary information-based methods in recommendations.

Deep learning-based recommendation methods have significantly improved the field by enabling automatic learning of complex, non-linear relationships from large-scale data, representing a huge advancement in addressing spurious correlations in recommendation systems [190, 244, 214, 175]. As shown in Figure 2.6, these methods overcome traditional techniques' limitations by employing sophisticated neural network architectures to learn rich latent representations of users and items that capture intricate interaction patterns. For instance, neural collaborative filtering models utilize deep learning architectures to uncover more nuanced and expressive representations compared to traditional matrix factorizations [66, 139]. Furthermore, deep learning techniques like Convolutional Neural Networks (CNNs), have the ability to effectively process unstructured data, including images and text, enabling content-aware recommendations in multimedia domains [104, 56]. A key advantage of deep learning models lies in their ability to automatically extract high-

level features from raw data [143, 123], eliminating the need for manual feature engineering that often introduces human bias and oversight. This automated feature learning capability enables deep learning-based recommenders to capture subtle patterns and complex relationships that might be missed by traditional handcrafted features, leading to substantial improvements across diverse application domains. The hierarchical representation learning paradigm inherent in deep architectures allows these models to progressively abstract meaningful features at multiple levels of granularity, from low-level interaction signals to high-level preference patterns. Despite these advancements, it is important to note that deep learning approaches still fundamentally rely on correlation-based learning paradigms. While they excel at discovering complex patterns in historical data, they may inadvertently amplify rather than mitigate existing spurious correlations if such patterns are prevalent in the training data. This limitation arises because deep models optimize for predictive accuracy on observed data without explicit mechanisms to distinguish between genuine causal relationships and spurious associations. Consequently, while deep learning significantly enhances the representation capacity of recommendation systems, it does not inherently solve the fundamental challenge of spurious correlations that this thesis aims to address through causal inference techniques. But, the ability to learn directly from vast amounts of heterogeneous information while automatically identifying complex patterns has established deep learning as a cornerstone of modern recommendation systems.

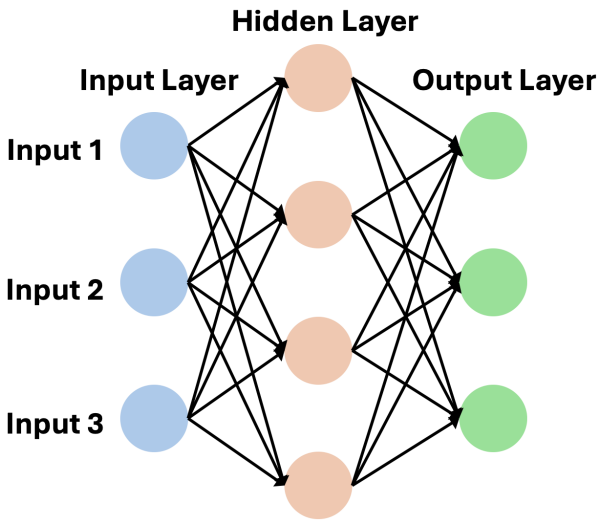


Figure 2.5: A toy example of deep learning-based methods in recommendations.

Graph-based recommendation models capture complex user-item relationships by representing interactions as graphs, revealing patterns beyond the reach of traditional meth-

ods [210, 223]. Graph Neural Networks (GNNs) have emerged as powerful tools for processing such structured data, enabling the identification of high-order connectivity and richer semantics through multi-hop message passing [206, 75, 48, 202]. As illustrated in Figure 2.6, each GNN layer aggregates information from a broader neighborhood, uncovering latent relationships and improving recommendation quality, especially in sparse settings. Despite these advances, GNNs still face the challenge of spurious correlations: by propagating observed interactions, they may reinforce biases from factors such as popularity, exposure, or external confounders. This correlation-driven process can amplify misleading associations, ultimately distorting true user preferences and recommendation outcomes. To address these challenges, recent research incorporates causal inferences into graph-based models to separate true user preferences from confounding influences [213, 197, 29]. While these methods significantly advance bias reduction and causal reasoning in graph-based recommendation, they rely on strong assumptions regarding the accurate modeling of confounders and the completeness of observed interactions. Additionally, they only focus on static settings or a single type of confounder, and seldom offer unified frameworks that address complex, dynamic, or multi-faceted recommendation scenarios. In contrast, the solutions developed in this thesis provide a more flexible and generalizable causal modeling paradigm. Our models are specifically designed to adapt to various confounding structures, support complex recommendation environments, and jointly enhance both robustness and interpretability through principled causal inference.

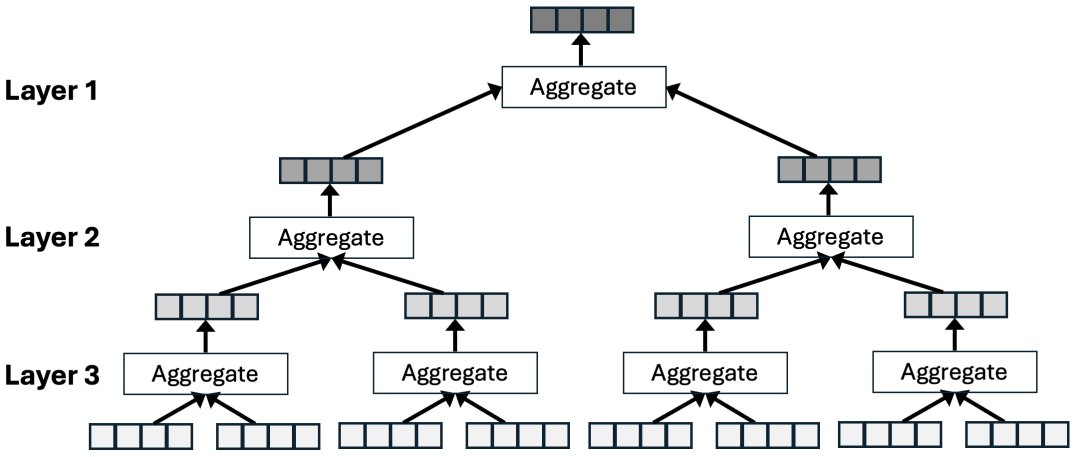


Figure 2.6: A toy example of GNNs in recommendations.

Sequential/session recommendations significantly enhance recommendation robustness by systematically modeling temporal dynamics and sequential patterns as session graphs, representing a crucial advancement in addressing time-dependent spurious corre-

lations [156, 47, 179, 105]. As shown in Figure 2.7, these methods operate on the fundamental understanding that user preferences and intentions evolve dynamically over time, with recent interactions often exerting substantial influence on current choices. At the architectural level, Recurrent Neural Networks (RNNs) and their advanced variants, particularly Long Short-Term Memory (LSTM) networks, demonstrate exceptional capability in modeling sequential data through their sophisticated internal state mechanisms that update with each interaction [237, 36]. Technically, these architectures incorporate specialized memory cells and gating mechanisms that selectively retain or forget information across the temporal sequence, enabling them to capture both short-term transitions and longer-term dependencies simultaneously. Such architectural design enables simultaneous modeling of both immediate user interests through recent interaction patterns and enduring preferences through persistent memory structures, creating a comprehensive representation of user behavior that adapts over time. The field has further evolved with the introduction of self-attention mechanisms and transformer models, which have significantly enhanced the ability to capture long-range dependencies in user behavior sequences, leading to more robust recommendations by understanding complex temporal relationships [84, 108]. Unlike traditional sequential models that process information linearly, attention-based approaches can directly model relationships between any two positions in a sequence, regardless of their distance. This capability allows the system to identify meaningful patterns across temporally distant interactions, recognizing when past behaviors become relevant again as user interests cycle or evolve in non-linear ways. However, despite their effectiveness in capturing sequential patterns, these approaches face a significant challenge with the cold-start problem for new users/items, as they rely heavily on historical sequential data to make accurate predictions.

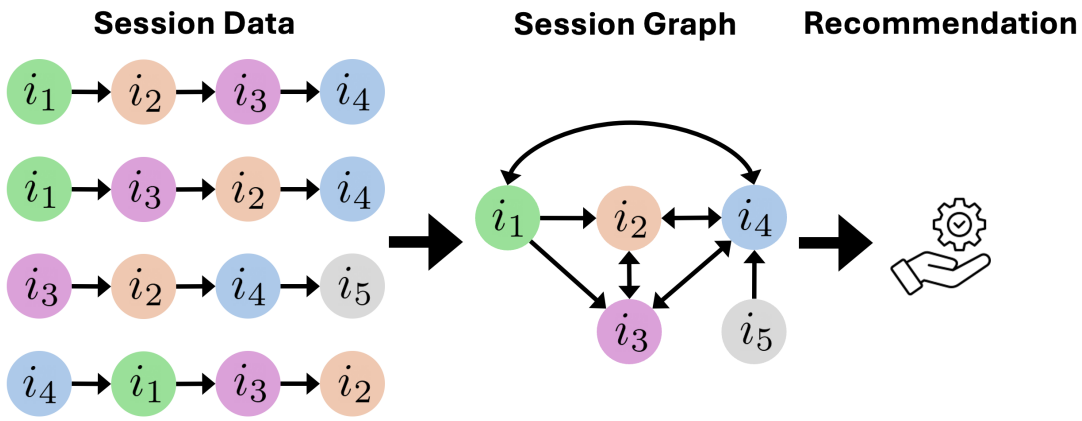


Figure 2.7: A toy example of session graph in sequential recommendations.

Conversational recommender systems (CRSs) represent another significant advancement in handling complex relational data by enabling dynamic preference refinement through interactive multi-turn conversations, marking a fundamental shift from static recommendation approaches [102, 80, 246, 92]. These systems overcome traditional limitations in capturing nuanced user preferences by establishing interactive conversations that progressively refine the understanding of user needs, as shown in Figure 2.8. Through the sophisticated integration of natural language processing techniques with recommendation algorithms, these systems demonstrate remarkable capability in understanding user intents, generating contextual explanations, and dynamically adjusting recommendations based on real-time feedback [136, 30]. The incorporation of reinforcement learning methods further enhances these systems by optimizing dialogue strategies to maximize long-term user satisfaction and personalization effectiveness [41, 40]. Despite their advantages, spurious correlations present a significant challenge in CRSs. Since user feedback in interactive systems is influenced by previous recommendations, spurious correlations can emerge when the system incorrectly infers user preferences based on biased interaction patterns. For instance, if a CRS repeatedly asks about certain attributes (e.g., price) due to prior interactions, it may reinforce misleading patterns rather than uncovering true user intent. This feedback loop can degrade recommendation quality by amplifying biases in past interactions rather than adapting to evolving user needs. One notable approach to mitigating this issue in CRSs involves formulating the recommendation task as a path-finding problem on a graph. In this approach, both users and items are modeled as nodes, while their relationships, such as past interactions, shared attributes, or contextual factors, form the edges. CRSs navigate through this graph, dynamically exploring paths based on user feedback, which can help disentangle true user preferences from spurious correlations by considering alternative pathways for exploration. However, despite these advanced capabilities, CRSs still face significant challenges related to computational resource demands, particularly in real-time applications where maintaining efficient and responsive dialogues requires substantial processing power. These challenges become even more pronounced when dealing with long, complex interactions that involve high-dimensional user representations, limiting the scalability of current CRSs in practical large-scale deployments.

In summary, advanced recommendation techniques have made significant progress in mitigating spurious correlations by employing sophisticated modeling approaches. Deep learning architectures extract hierarchical representations, capturing complex patterns beyond simple correlations. Graph-based models, particularly graph neural networks, utilize high-order connectivity to propagate structured information and improve relational un-

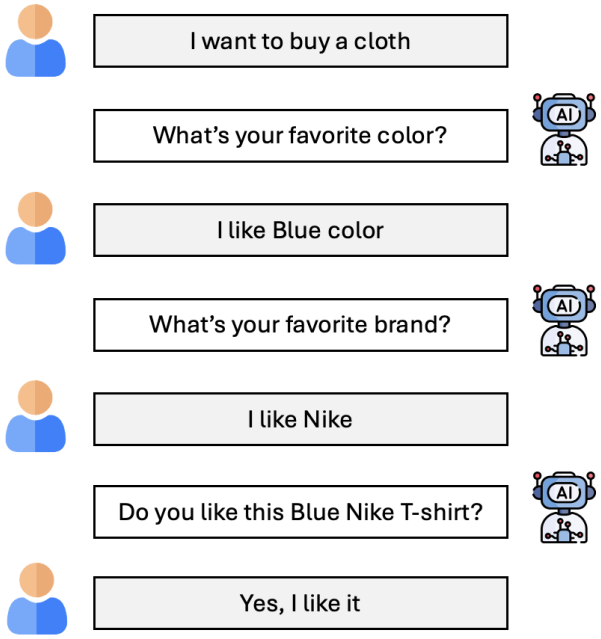


Figure 2.8: A toy example of interactive dialog in conversational recommendations.

derstanding. Sequential and conversational models incorporate temporal dynamics, enabling recommendations to adapt to evolving user preferences through interactive or time-aware learning mechanisms. However, despite these advancements, these methods still rely on correlation-based learning, making them susceptible to spurious patterns in the data. While they effectively model associations, they struggle to distinguish true causal relationships from confounded observations. This limitation has led to the exploration of causal inference approaches, which systematically address spurious correlations by explicitly modeling and controlling for confounders. By shifting from correlation-based learning to causal reasoning, recommender systems can achieve more robust, unbiased, and interpretable recommendations.

2.4 Causality-based Recommendation Techniques

To address the shortcomings of correlation-based learning, causal inference has emerged as a principled statistical framework for identifying and quantifying true cause-effect relationships in recommendation [129, 192]. Generally, causal inference leverages five complementary techniques to systematically address spurious correlations from different perspectives: (1) causal graphs visually represent dependencies and help identify confounders; (2) do-calculus provides rules for unbiased causal effect estimation from observational

data; (3) interventions simulate the impact of hypothetical actions on user preferences; (4) propensity score methods balance covariates to correct for selection and exposure bias; and (5) counterfactual reasoning explores hypothetical scenarios to infer potential outcomes. By systematically integrating causal techniques, recommender systems can better distinguish true user preferences from spurious correlations, isolating genuine causal factors while controlling for confounders. For example, Bonner and Vasile [12] proposed causal embedding models that explicitly represent confounding factors in recommendation, leading to less biased user representations. Zhu et al. [257] developed deep structural causal models to disentangle direct and indirect effects in recommendation tasks. Li et al. [99] introduced causal graph neural networks to correct for bias propagation along user-item interaction graphs. Zhang et al. [248] leveraged counterfactual inference to address selection bias and improve ranking performance in RSs. However, current causal recommenders still face challenges—such as assuming observable confounders, limited adaptability to complex scenarios, and difficulties in achieving both robustness and interpretability. These gaps motivate the development of novel causal models that can more effectively address the problem of spurious correlations in real-world recommendations, which is the main focus of this dissertation. To lay the groundwork for novel causal models, the following section analyzes key causal inference techniques, critically evaluating their methodologies, contributions, and limitations.

First is the causal graph, also known as a directed acyclic graph (DAG), which serves as a foundational framework for visualizing and analyzing causal relationships, enabling the distinction between true causal effects and spurious correlations [51, 131]. These graphs serve as comprehensive analytical blueprints by representing key variables (e.g., user demographics, item characteristics, and contextual factors) as nodes and their causal relationships as directed edges, enabling systematic identification of true causal pathways that influence user behaviors [255, 248]. As shown in Figure 2.9, in the left panel, the DAG represents a generic causal structure where treatment (T) influences outcome (Y), but is confounded by X1 and X2, both of which are also affected by an unobserved variable (U). This structure illustrates the importance of identifying confounding pathways, such as the path through X1 or X2, which can introduce biases in estimating the effect of T on Y. Controlling for these confounders is necessary for obtaining an unbiased causal effect estimate. In the right panel, the DAG illustrates the causal mechanisms underlying education and wage determination. Family income affects both education level and wage, suggesting a confounding relationship that must be accounted for when estimating the impact of education on wage. Additionally, intelligence influences both SAT scores and education, forming

an alternative pathway that could confound causal inference. By explicitly modeling these relationships, the graph highlights both back-door and front-door adjustments [233, 155]. The back-door path (through family income) can be blocked by conditioning on income, while the front-door adjustment via SAT scores allows for an alternative estimation of the education effect. These causal structures highlight the necessity of systematic identification of confounding factors and mediation effects, guiding the application of causal inference techniques such as back-door and front-door adjustments. By leveraging DAGs, researchers can design more robust causal estimation methods that accurately capture the true effects of interventions while mitigating biases arising from unobserved confounders.

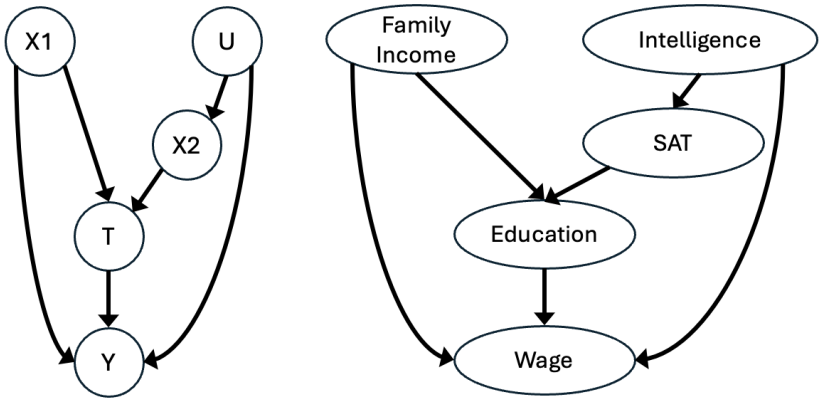


Figure 2.9: A toy example of the causal graph, showing the cause and effect relationships among variables.

Second is the causal intervention, which is a fundamental technique for estimating the effect of hypothetical actions or treatments in complex decision-making settings [239]. In the context of recommender systems, causal intervention enables the systematic evaluation of how modifications in recommendation strategies influence user engagement and satisfaction while distinguishing true causal effects from spurious correlations [191, 186, 233, 58, 226]. The causal graph in Figure2.9 illustrates an analogous scenario in education, where factors such as family income and intelligence influence both education and wage. In recommender systems, similar confounding occurs when historical exposure, platform ranking mechanisms, or external social influences affect both the recommended items and user engagement, leading to biased estimates if not properly adjusted. To formally estimate causal effects in recommender systems, two principal causal intervention methods are widely used: First is the back-door adjustment, which controls for confounding variables that introduce spurious correlations between recommendations and user engagement [251, 248]. In the education analogy, family income and intelligence influence both

education and wage, forming back-door paths that distort the estimation of education's true effect. Similarly, in recommender systems, historical exposure and platform biases can influence both recommendation exposure and engagement, creating misleading associations. Controlling for these confounders ensures that the estimated effect of recommendations reflects genuine user preferences. When back-door confounders are unobservable, front-door adjustment offers an alternative by leveraging intermediate variables along the causal pathway [220]. In education, SAT scores serve as a mediator between education and wage, providing a valid causal estimate even when intelligence remains unmeasured. Likewise, in recommender systems, dwell time or add-to-cart interactions act as mediators between recommendations and final engagement outcomes. Since these mediating factors are directly influenced by recommendations before impacting long-term engagement, they enable causal estimation without needing full confounder information. By integrating both back-door and front-door adjustments as causal interventions, recommender systems can systematically optimize decision-making, improve user experience, and ensure that recommendations reflect genuine user preferences rather than artifacts of biased data.

Third, counterfactual reasoning enables the estimation of outcomes under alternative scenarios, allowing recommender systems to answer "what-if" questions and evaluate the effects of different recommendation strategies without requiring costly A/B testing [158, 243], as shown in Figure 2.10. Unlike traditional correlation-based approaches that rely on observed interactions, counterfactual reasoning leverages structural causal models to predict user behavior under hypothetical conditions, effectively disentangling true causal effects from spurious correlations [182, 196]. For example, by estimating how users would interact with an item if it were displayed in a different ranking position, counterfactual reasoning can help mitigate exposure bias, ensuring that item recommendations are not unfairly influenced by placement effects. Beyond addressing confounding bias, counterfactual reasoning also generates counterfactual samples, effectively augmenting the original dataset by simulating alternative user-item interactions [195]. This augmentation is particularly valuable in cold-start scenarios or when user feedback is sparse, as it provides additional training signals to enhance recommendation quality. Furthermore, counterfactual reasoning supports explainability by generating contrastive explanations, which highlight the factors that would need to change for a different recommendation to be made [167]. These explanations improve transparency by helping users and researchers understand why certain recommendations were given and how they could be altered. Additionally, counterfactual reasoning facilitates off-policy evaluation by simulating user responses to unseen recommendations, making it possible to assess new strategies before deploying

them in real-world systems [160, 141]. By enabling robust evaluation and refining recommendation logic, counterfactual reasoning enhances both decision-making and recommendation reliability, ensuring that systems generate outcomes that reflect genuine user preferences rather than misleading statistical artifacts.

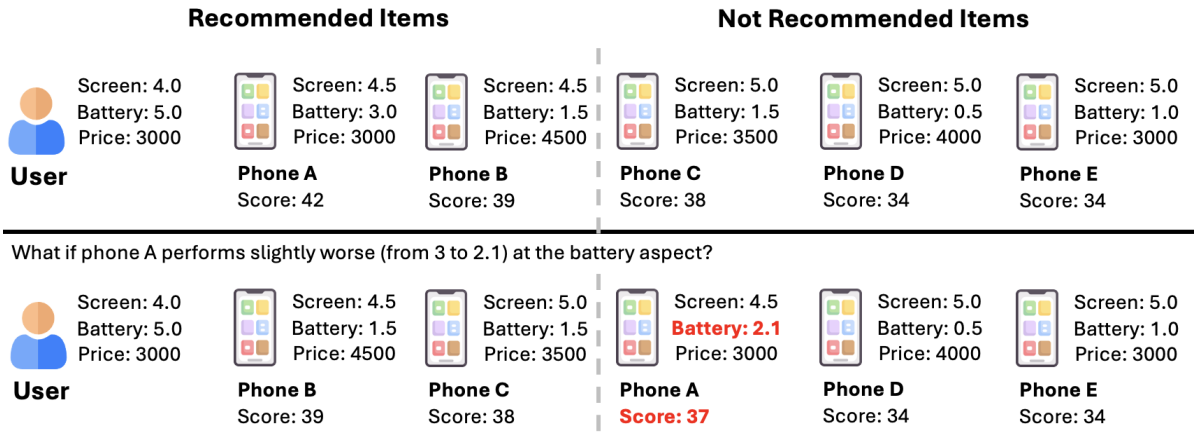


Figure 2.10: A toy example of the counterfactual reasoning in recommendations.

Fourth, do-calculus provides a mathematical framework for reasoning about interventions and counterfactuals in causal models, offering a principled approach to estimating causal effects in recommender systems[34, 164]. Unlike correlation-based techniques that rely on observational data, do-calculus defines a set of formal rules for manipulating causal graphs to quantify the effects of hypothetical interventions, such as exposing a user to a particular item[79]. This method enables recommender systems to infer how user behavior would change under controlled interventions, allowing for causal effect estimation even in the presence of confounding biases. A major advantage of do-calculus is its ability to control for confounding factors that simultaneously influence both item exposure and user preferences, ensuring that spurious correlations do not distort causal conclusions [43]. By systematically transforming probabilistic expressions involving interventions, do-calculus determines when causal effects can be estimated from observational data, reducing reliance on costly randomized experiments. Additionally, it provides a principled method for combining multiple data sources, improving the validity and robustness of causal inferences [131]. For example, in a recommendation scenario where an item appears popular due to frequent exposure rather than genuine preference, do-calculus can separate the direct causal effect of an item's relevance from confounded popularity effects, ensuring that recommendations reflect true user interests rather than visibility biases. Through these unbiased estimates, do-calculus enhances both recommendation accuracy and interpretabil-

ity, leading to more reliable strategies for optimizing user engagement while mitigating the risks of misleading statistical associations.

Fifth, propensity scores mitigate selection/exposure biases in observational data by ensuring that the distribution of covariates is balanced across treatment groups [250, 112]. In recommender systems, propensity scores quantify the probability of a user being exposed to or interacting with an item based on observed characteristics [12, 212]. Since user interactions are often influenced by algorithmic exposure rather than purely by preference, selection bias arises when users who engage with certain items are systematically different from those who do not. Without adjusting for these biases, recommendations may reinforce spurious correlations, leading to the misinterpretation of user behavior as a genuine preference rather than a consequence of prior exposure. The figure 2.11 illustrates propensity score matching (PSM) as a technique to reduce bias by creating comparable groups. Initially, users who received the treatment (e.g., took the pill) and those who did not (e.g., did not take the pill) exhibit imbalanced distributions of covariates, leading to biased comparisons. PSM matches individuals with similar propensity scores to ensure that the treated and untreated groups are comparable, reducing confounding effects. Beyond matching, propensity score methods employ additional techniques for more refined adjustments. For instance, weighting corrects for differences in exposure likelihood by assigning appropriate statistical weights to observations, thereby ensuring unbiased effect estimation across the entire sample. Similarly, stratification divides users into subgroups based on propensity scores, facilitating structured comparisons that control for systematic differences [113, 146]. These complementary approaches provide RSs with multiple tools to address selection bias from different perspectives. For example, when evaluating whether increasing product recommendations improves engagement, propensity scores help disentangle algorithmic influence from actual user interest by controlling for prior exposure effects. By integrating these methods, propensity scores enable recommender systems to obtain unbiased estimates of user preferences, improving both evaluation methodologies and optimization strategies while mitigating the risks of exposure bias.

In summary, causal inference provides a comprehensive statistical framework for systematically identifying true cause-and-effect relationships among variables in recommender systems, enabling the distinction between true user preferences and spurious correlations. Generally, causal inference leverages multiple complementary techniques to address different aspects of spurious correlations: causal graphs serve as the foundation by visually mapping relationships and identifying confounders, do-calculus establishes mathematical rules for valid causal estimation from observational data, causal intervention assesses

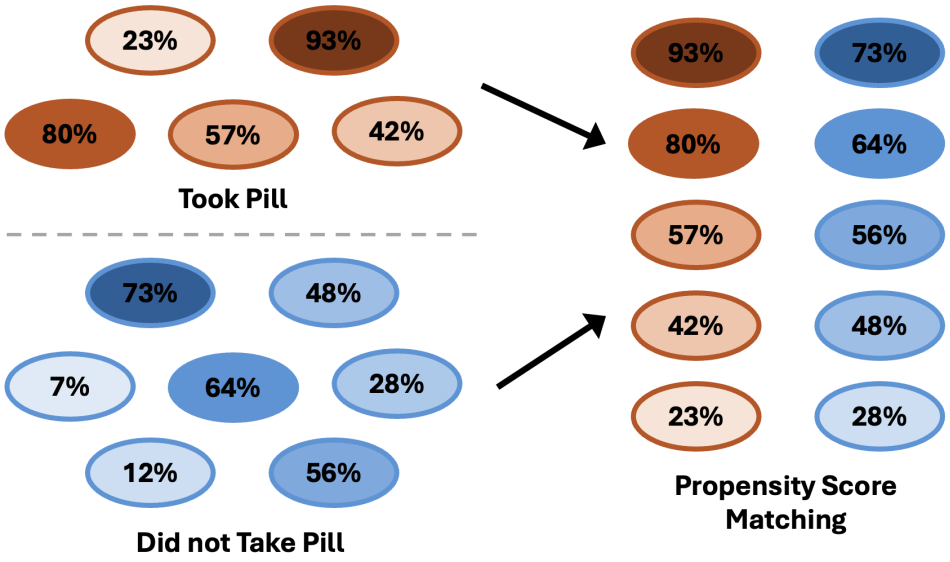


Figure 2.11: A toy example of the propensity score matching.

treatment effects through controlled experiments, propensity scoring mitigates selection and exposure biases, and counterfactual reasoning enables exploration of hypothetical scenarios. Through systematic integration of these causal techniques, recommender systems can effectively isolate true factors influencing user preferences to address spurious correlations, ultimately enabling more robust and explainable recommendations.

2.5 Summary of Model Contributions

To provide a clear overview of the thesis structure and model progression, Table 2.1 provides a comprehensive summary of the proposed models, their scenarios, tasks, datasets, metrics, and corresponding chapters. This organization illustrates the distinct challenges tackled in each stage, and the cumulative advancement toward robust and explainable causal recommendation.

Table 2.1: Summary of Proposed Models: Scenarios, Tasks, Datasets, Metrics, and Chapters

Model	**Scenario**	**Task**	**Datasets**	**Metric(s)**	**Chapter**
CCR	Conversational Recommendation	Causal-based attribute selection for unbiased conversational recommendation	Yelp, Douban-Book, MovieLens	SR@5/10/20, AT	3
LDPE	Sequential Recommendation	Dual propensity score estimation for unbiased sequential recommendation	MovieLens-1M, Amazon-Books, Taobao	Recall@5/10/20, Precision@5/10/20, NDCG@5/10/20	3
CGSR	Session-based Recommendation	Causal shortcut path blocking in session-based graph recommendation	Diginetica, Yoochoose, Gowalla	Recall@10/20, NDCG@10/20, MRR@10/20	4
GCRec	High-order Graph Recommendation	Causal adjustment for multiple confounders in graph-based recommendation	Amazon-Books, Taobao	Recall@20, Precision@20, NDCG@20, HR@20	4
CEDA	Social Network Recommendation	Causal echo chamber attenuation for social diffusion recommendation	Twitter, Facebook, Google+	Diversity, Echo Chamber Index, Precision@20, Recall@20, NDCG@20	4
SeDLR	Explainable Recommendation	Semantics-guided disentangled learning for explainable recommendation	Walmart Recruit, Yelp	Explainability Score, Recall@10/20, NDCG@10/20	5
CECR	Conversational Recommendation	Counterfactual explanation for conversational recommendation	REDIAL, TGReDial, MovieLens-1M	Success Rate, Explainability Score, Recall@1/10, Precision@1/10	5

CHAPTER 3

CAUSAL MODELS FOR UNBIASED RECOMMENDATIONS

This chapter addresses RQ1 by introducing two complementary causal models—CCR and LDPE—each designed to mitigate spurious correlations in distinct recommendation settings. CCR targets Conversational Recommender Systems, addressing attribute-level confounders that arise from dynamic user intents during multi-turn dialogues through causal stratification and matching. LDPE, on the other hand, focuses on Sequential Recommender Systems, leveraging LLMs and causal inferences to debias user behavior sequences affected by exposure bias. Together, these models provide a unified causal framework that systematically addresses different sources of confounding bias, demonstrating the versatility and necessity of causal inference in enhancing robustness across diverse recommendation scenarios. Through detailed methodologies, rigorous experimental evaluations, and comprehensive findings, this chapter establishes the foundational role of causal inference in uncovering true user preferences for unbiased recommendations.

3.1 A Causal-Based Attribute Selection Strategy for Conversational Recommender Systems

3.1.1 Overview

Conversational recommender systems (CRSs) provide personalized recommendations by strategically querying attributes matching users' preferences through interactive dialogues. Unlike traditional static recommenders, CRSs iteratively refine the understanding of user

preferences through real-time feedback, making them highly effective for personalized recommendations. However, this process suffers from confounding effects that distort the true causal drivers of user behavior. First, time confounders arise from evolving user preferences over conversation rounds. For example, a user who initially accepts sun hat recommendations may later reject them and prefer raincoats due to weather changes. Such temporal shifts in preferences create spurious correlations between items, making it difficult for CRSs to discern users' true preferences. Second, user attribute confounders emerge from unique user characteristics leading to diverse preferences among seemingly similar users. For instance, a lawyer and an athlete may share several attributes but have vastly different clothing preferences due to their occupations. These attribute-driven differences create spurious relationships between shared attributes and user preferences, challenging CRSs' ability to learn generalizable patterns. Together, these confounding effects can mask the true causal factors driving user decisions, potentially leading to biased and ineffective recommendations.

3.1.1.1 Research Objective

This study aims to address RQ1 by developing a principled causal framework, that can simultaneously mitigate both the time and user attribute confounders while accurately identifying the true causal drivers of user behaviors in CRSs. Generally, we propose to synergistically integrate stratification to ensure attribute independence from conversational rounds, with matching mechanisms to pair similar users, effectively addressing both time and user attribute confounding effects. This integration enables us to accurately estimate the unbiased Average Treatment Effect (ATE) of attributes on user preferences, identifying the most causally significant attributes for user queries. In this way, we aim to uncover the true causal factors shaping user preferences while providing more accurate and personalized recommendations.

3.1.1.2 The Proposed Method

To achieve the aforementioned objectives, we propose a novel causal framework called **C**ausal **C**onversational **R**ecommender (**CCR**) with four synergistic components:

- ***Offline Representation Learning*** creates initial user, item, and attribute representations from historical data, with offline user representation capturing long-term preferences that will be updated with online feedback.

- ***Causal-based Attribute Selection*** addresses time and user attribute confounders through stratification and matching, while identifying the causal-based attribute in each round with the highest Average Treatment Effect (ATE).
- ***User Preferences Refining*** iteratively updates user preferences based on asked causal-based attributes and recommended items, along with online feedback, producing a refined representation capturing both short-term and long-term preferences.
- ***Recommendation*** leverages the refined user representation for Top-K recommendations, ensuring a dynamically adapted and personalized CRS.

The key contributions of this research are:

- We are the first to investigate how causal factors, especially round-specific attribute selection, shape user interactions in CRSs, providing novel insights for enhancing CRSs through causal analysis.
- We develop an innovative causality-driven approach named CCR that simultaneously handles time and user attribute confounders in CRSs while accurately identifying true causal drivers of user behavior.
- We propose a novel attribute selection mechanism grounded in causality theory that prioritizes asking the most influential attributes on user preferences, enhancing conversation efficiency and personalization.
- Comprehensive evaluations across three datasets demonstrate that CCR achieves superior performance compared to existing state-of-the-art CRSs.

3.1.2 CCR

3.1.2.1 Problem Definition

Let $\mathscr{U}$, $\mathscr{I}$ and $\mathscr{A}$ be the sets of users, items and attributes, respectively. Each item $i \in \mathscr{I}$ is associated with a set of attributes $\mathscr{A}_i \subseteq \mathscr{A}$ and interacted by users $\mathscr{U}_i \subseteq \mathscr{U}$. $\mathscr{I}_u \subseteq \mathscr{I}$ denotes items interacted by the user u. $\mathscr{I}_a \subseteq \mathscr{I}$ denotes items that share the common attribute $a \in \mathscr{A}$. $\mathscr{U}_u \subseteq \mathscr{U}$ denotes users have direct connection with the user u. $\mathscr{U}_a \subseteq \mathscr{U}$ denotes users share the same attribute $a \in \mathscr{A}$. Given a user $u \in \mathscr{U}$ who initiates a dialogue with a preferred attribute $a \in \mathscr{A}$, our goal is to recommend items that match their preferences through an interactive dialogue with fewer rounds.

To achieve this, it is crucial to address two key confounders mentioned: time confounders arising from evolving preferences across conversation rounds, and user attribute confounders stemming from unique user characteristics. As shown in Figure 3.1, we design a causal graph that reveals the intricate relationships among five key variables in CRSs: time R, asked attribute A, item I, rating Y, and user attribute U. The directed causal edges represent how time R and user attributes U act as confounders by influencing both attribute selection A and rating outcomes Y, potentially creating spurious correlations that mask true preference patterns. Based on this graph, we employ stratification to ensure attribute independence from rounds and matching to pair similar users. These causal inference techniques enable accurate estimation of Average Treatment Effect (ATE) for each attribute over the causal path $A \rightarrow I \rightarrow Y$, while blocking confounding effects through back-door paths $R \rightarrow A$ and $U \rightarrow A$. The attribute with the highest ATE is selected as the causal-based attribute per round for user queries. Through iterative preference refinement based on real-time feedback, we aim to provide increasingly personalised and robust recommendations.

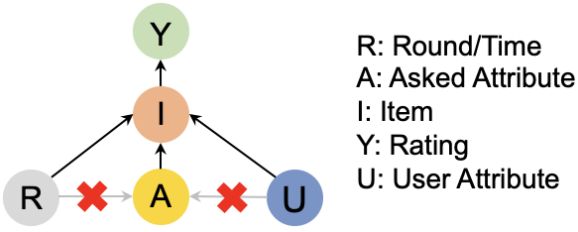


Figure 3.1: Our designed causal graph for CRSs.

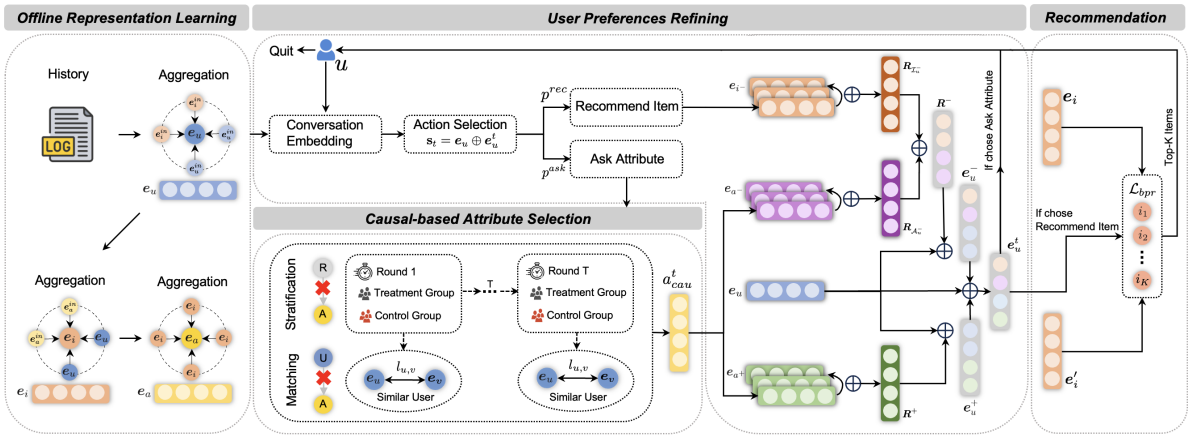


Figure 3.2: The overall framework of our proposed method CCR.

3.1.2.2 Methodology

Figure 3.2 presents the framework of our proposed CCR, which comprises four components. The first component ***Offline Representation Learning*** learns offline user/item/attribute representations from historical data, forming a solid starting point for conversational interactions. Specifically, we first employ a retrieval function [137] that converts categorical attributes into numerical encodings. For instance, given an item with attributes "Color": ["Red", "Blue", "Green"] and "Size": ["Small", "Medium", "Large"], we encode "Blue" as [0,1,0] and "Medium" as [0,1,0], then concatenate them to form the initial item representation $\boldsymbol{e}_i^{in}$ for blue medium item. The same encoding process is used for generating initial user representation $\boldsymbol{e}_u^{in}$ and initial attribute representation $\boldsymbol{e}_a^{in}$.

Social behavior theory suggests that user preferences are reflected in both consumption patterns and community influences [150, 80]. To quantify these dual influences, we construct a preference modeling framework using two complementary interaction spaces: a social network matrix $\boldsymbol{U}$ encoding interpersonal connections, where $U_{u,v}$ represents the presence of user-user relationships, and a behavioral matrix $\boldsymbol{O}$ capturing consumption patterns, where $O_{u,i}$ indicates user-item interaction rating. Next, we synthesize these spaces through a weighted approach to create offline user representation $\boldsymbol{e}_u$ as

$$\boldsymbol{e}_u = \alpha \cdot \frac{\sum_i O_{u,i} \cdot \boldsymbol{e}_i^{in}}{|\mathscr{I}_u|} + (1-\alpha) \cdot \frac{\sum_v U_{u,v} \cdot \boldsymbol{e}_v^{in}}{|\mathscr{U}_u|} \tag{3.1}$$

where α is a parameter to control the importance between consumption patterns and social influences. $\boldsymbol{e}_v^{in}$ is the initial user representation of user v, follows the same procedure as generating $\boldsymbol{e}_u^{in}$. $\boldsymbol{e}_u$ is the offline user representation, capturing rich semantics from historical interactions and social connections, denoting user long-term preferences.

Moreover, research indicates that item characteristics serve as preference indicators since users engaging with identical items often exhibit preference similarities [186]. Thus, we integrate both user engagement signals and item attributes to learn the offline item representation. Specifically, we introduce an item-attribute matrix $\boldsymbol{A}$, where each element $A_{i,a}$ is a binary indicator that means the presence of attribute a in item i. Then, we employ a weighted approach to create offline item representation $\boldsymbol{e}_i$ as

$$\boldsymbol{e}_i = \beta \cdot \frac{\sum_u O_{u,i} \cdot \boldsymbol{e}_u}{|\mathscr{U}_i|} + (1-\beta) \cdot \frac{\sum_a A_{i,a} \cdot \boldsymbol{e}_a^{in}}{|\mathscr{A}_i|} \tag{3.2}$$

where β balances the contribution of user engagement signals and attribute information, with $\boldsymbol{e}_u$ derived from Eq (3.1). This fusion creates semantically rich item embeddings that reflect the preferences of users who have engaged with the item and the intrinsic characteristics of the item itself.

On the other hand, we learn the offline attribute representation $\boldsymbol{e}_a$ by considering the collective influence of an attribute across various items and users. The core idea is that items sharing the same attribute a can effectively encapsulate the semantics and potential characteristics associated, while users who have interacted with the same attribute a can share similar preferences. Specifically, we employ a weighted approach to generate the offline attribute representation $\boldsymbol{e}_a$ as

$$\boldsymbol{e}_a = \psi \cdot \frac{\sum_u U_{u,a} \cdot \boldsymbol{e}_u}{|\mathscr{U}_a|} + \frac{\sum_i A_{i,a} \cdot \boldsymbol{e}_i}{|\mathscr{I}_a|} \tag{3.3}$$

where ψ is a parameter that moderates the fusion of user signals and item characteristics. $\boldsymbol{e}_a$ is the offline attribute representation for attribute a, capturing rich semantics from user-attribute and item-attribute relationships.

The second component ***Causal-based Attribute Selection*** builds upon learned offline representations to identify causally significant attributes per round while mitigating confounding effects from time and user attributes. Following user initialization, we employ Deep Q-learning [69] to adaptively choose between immediate recommendations or attribute exploration. To ensure unbiased attribute selection, we address two key confounding factors, where time confounders stem from shifting user preferences and user attribute confounders stem from each user's unique attributes. Generally, we employ stratification and matching to ensure attributes are independent of conversational round and user attributes, respectively. This can block confounding effects from time and user attributes by applying the back-door adjustment of causal inference, allowing us to accurately estimate the unbiased ATE of each attribute to assess its potential impact on user preferences. The attribute exhibiting the highest ATE is selected as the causal-based attribute per round, as substantial effects on outcomes typically indicate causal drivers of user behavior [43].

Formally, we first define the set $\mathscr{X} = x_q = (u_q, t_q, v_q, b_q)$, where each tuple x_q represents a feedback instance. Here, t_q denotes the round, v_q represents user attributes, and b_q indicates attribute exposure ($b_q = 1$ if user u_q was asked about attribute a before round t_q, 0 otherwise). Then, we partition $\mathscr{X}$ into T temporal subgroups $\mathscr{X}_1, ..., \mathscr{X}_t, ..., \mathscr{X}_T$ based on round number, which isolates time-specific data and mitigates temporal confounding within each subgroup. Within $\mathscr{X}_t$, we classify users into treatment ($b_q = 1$) and control ($b_q = 0$) groups based on attribute exposure. This enables estimation of the Average Treatment Effect (ATE) for each subgroup:

$$\mathrm{ATE}_{t,q} = \bar{L}_t^{\mathrm{t}} - \bar{L}_t^{\mathrm{c}} = \frac{\sum_{x_q \in \mathscr{X}_t} b_q v_q}{\sum_{x_q \in \mathscr{X}_t} b_q} - \frac{\sum_{x_q \in \mathscr{X}_t} (1 - b_q) v_q}{\sum_{x_q \in \mathscr{X}_t} (1 - b_q)} \tag{3.4}$$

where $\text{ATE}_{t,q}$ represents the average treatment effect for user q in round t. $\bar{L}_t^t$ denotes the average outcome value for the treatment group in round t. $\bar{L}_t^c$ is the average outcome value for the control group in round t. v_q indicates the observed outcome for user q and $\mathscr{X}_t$ represents the set of all user feedback instances in round t. This stratification approach effectively blocks the confounding path through time while preserving the ability to measure true causal effects of attributes on user preferences.

After computing subgroup ATEs, we aggregate them into an overall time-aware ATE:

$$\text{ATE}_{time} = \sum_{t=1}^{T} \frac{|\mathscr{X}_t|}{|\mathscr{X}|} \text{ATE}_{t,q} \tag{3.5}$$

where $\frac{|\mathscr{X}_t|}{|\mathscr{X}|}$ weights each subgroup by its relative size. To efficiently compute these subgroups, we design a user sampling approximation strategy that categorizes users based on their exposure timing relative to each subgroup t. We partition users into three groups: $\mathscr{U}_1$ contains users receiving attribute recommendations before the subgroup ($P(A = 1 \mid R = t_q, \mathscr{U}_1 \subseteq \mathscr{U}) = 1$); and $\mathscr{U}_2$ contains users exposed after the subgroup ($P(A = 1 \mid R = t_q, \mathscr{U}_2 \subseteq \mathscr{U}) = 0$); and $\mathscr{U}_3$ contains users who received attribute recommendations at the time of subgroup ($P(A = 1 \mid R = t_q, \mathscr{U}_3 \subseteq \mathscr{U})$). Note that we mainly focus on first two user groups while ignoring the last user group, as its infrequent nature may skew the performances and add complexities. Mathematically, we have

$$X_t' = \{x_q \mid u \in \mathscr{U}_1 \cup \mathscr{U}_2\} \tag{3.6}$$

For samples in X_t', we can show independence between time and attribute exposure:

$$\begin{aligned} & P(A = 1 \mid R = t) \\ & \overset{(a)}{=} \sum_{u \in \mathscr{U}_1 \cup \mathscr{U}_2} P(A = 1 \mid U = u, R = t) P(U = u \mid R = t) \\ & \overset{(b)}{=} \sum_{u \in \mathscr{U}_1 \cup \mathscr{U}_2} P(A = 1 \mid U = u) P(U = u \mid R = t) \\ & \overset{(c)}{=} \sum_{u \in \mathscr{U}_1 \cup \mathscr{U}_2} P(A = 1 \mid U = u) P(U = u) \overset{(d)}{=} P(A = 1) \end{aligned} \tag{3.7}$$

where $P(A = 1 \mid R = t)$ denotes the probability of attribute exposure at round t, (a) follows from the Law of Total Probability, (b) reflects conditional independence between attribute exposure and round given user, (c) assumes independence between round and user assignment, and (d) shows the equivalence to the overall exposure probability $P(A = 1)$, indicating time and attribute exposure are independent. Thus, by replacing $\mathscr{X}_t$ with X_t' in the ATE calculation, we obtain unbiased estimates of attribute effects on user preferences during the conversations.

While addressing time confounders through stratification, samples within each subgroup may still be affected by user attribute confounders. To address this, we employ a matching technique that pairs treated and control users based on similar attributes, ensuring independence between user attributes U and asked attribute A by blocking the backdoor path $U \rightarrow A$. For each sample q, we define outcomes $b_q(t_q = 1)$ and $b_q(t_q = 0)$ for treatment and control conditions respectively. Mathematically, we have

$$\begin{aligned} \hat{b}_q\left(t_q = 1\right) &= \begin{cases} b_q, & \text{if } t_q = 1 \\ b_{q'}, & \text{if } t_q = 0 \end{cases} \\ \hat{b}_q\left(t_q = 0\right) &= \begin{cases} b_{q'}, & \text{if } t_q = 1 \\ b_q, & \text{if } t_q = 0 \end{cases} \end{aligned} \tag{3.8}$$

where $\hat{b}_q(t_q = 1)$ and $\hat{b}_q(t_q = 0)$ denote the estimated value of $b_q\left(t_q = 1\right)$ and $b_q\left(t_q = 0\right)$. q' is the index that identifies the matched user by minimizing the distance metric $l_{u,v}$ as

$$l_{u,v} = \frac{1}{2}\left(1 - \frac{\boldsymbol{e}_u^t \boldsymbol{e}_v^t}{\left\|\boldsymbol{e}_u^t\right\| \left\|\boldsymbol{e}_v^t\right\|}\right) \tag{3.9}$$

where $\boldsymbol{e}_u^t$ and $\boldsymbol{e}_v^t$ are user representations for the user u and user v at round t, and $l_{u,v}$ measures the distance between them. Using these matched pairs, we compute the unbiased ATE as:

$$\text{ATE}_{user} = \sum_{x_q \in X_t'} \left(\hat{b}_q\left(t_q = 1\right) - \hat{b}_q\left(t_q = 0\right)\right) \tag{3.10}$$

where ATE_{user} is the unbiased ATE free from the effects of user attribute confounders. This matching approach effectively mitigates user attribute confounding by ensuring balanced attribute distributions between treatment and control groups.

Having addressed both time and user attribute confounders, we now select the most causally significant attribute per round by evaluating their potential impact on user preferences. Specifically, we compute the ATE for each attribute a at round t using ATE_{user} as:

$$\text{ATE}_{a,t} = \sum_{x_q \in \mathscr{X}_{(a,t)}} \frac{\text{ATE}_{user}}{|\mathscr{X}_{(a,t)}|} \tag{3.11}$$

where $\mathscr{X}_{(a,t)}$ is the subgroup of users who were asked about attribute a at round t, derived from $\mathscr{X}$. $\text{ATE}_{a,t}$ is the ATE for attribute a at round t. The causal-based attribute for each round is then selected as:

$$a_{cau}^t = \arg\max_a \text{ATE}_{a,t} \tag{3.12}$$

where a_{cau}^t denotes the attribute with the highest causal impact on user preferences at round t. This approach ensures attribute selection aligns with true causal drivers of user behavior.

Algorithm 1 Learning Algorithm for CCR

Input: Set the maximum round number T. Set rewards r_t^s, r_t^i, r_t^l, r_t^a, r_t^t as 1, -0.5, 0.1, -0.5, -1. Set q_t^{ask} and q_t^{rec} as 0. P is a matrix containing attributes contained in the user's interacted items.

Initialize $\mathscr{A}_u^+$, $\mathscr{A}_u^-$, and $\mathscr{I}_u^-$
Initialize $\{\boldsymbol{e}_u\}$, $\{\boldsymbol{e}_i\}$, $\{\boldsymbol{e}_a\}$ based on Eq (3.1), Eq (3.2), Eq (3.3)
Initialize $\boldsymbol{e}_u^t = \boldsymbol{e}_u$, $p^{ask} = p^{rec} = P[u]$
for $t = 1, 2, 3, \ldots, T$ **do**
    select an action p
    if $p = p^{ask}$ **then**
        $a_{cau}^t = Causal\ Attribute\ Selector(\boldsymbol{e}_u^t, \boldsymbol{e}_v^t, \mathscr{X})$
        if user u likes a_{cau}^t **then**
            $\mathscr{A}_u^+ \leftarrow \mathscr{A}_u^+ \cup \{a_{cau}^t\}$
        else
            $\mathscr{A}_u^- \leftarrow \mathscr{A}_u^- \cup \{a_{cau}^t\}$
        end if
    else
        Recommend Top-K items to user u
        if user u accepts item i **then**
            Return the accepted item i
        end if
    end if
    Generate $\boldsymbol{R}_{\mathscr{A}_u^+}$, $\boldsymbol{R}_{\mathscr{A}_u^-}$, $\boldsymbol{R}_{\mathscr{I}_u^-}$ using $\mathscr{A}_u^+$, $\mathscr{A}_u^-$, $\boldsymbol{R}_{\mathscr{I}_u^-}$
    Generate $\boldsymbol{e}_u^+$ using $\boldsymbol{R}_{\mathscr{A}_u^+}$
    Generate $\boldsymbol{e}_u^-$ using $\boldsymbol{R}_{\mathscr{A}_u^-}$, $\boldsymbol{R}_{\mathscr{I}_u^-}$
    Refine user preference: $\boldsymbol{e}_u^t \leftarrow \boldsymbol{e}_u + \boldsymbol{e}_u^+ - \boldsymbol{e}_u^-$
    Generate state vector $\mathbf{s}_t$
    Update q_t^{ask}, q_t^{rec} to y_t^{ask}, y_t^{rec}
    Compute $\mathscr{L}_t$ using q_t^{ask}, q_t^{rec}, y_t^{ask}, y_t^{rec}
    Update network parameter: $\delta_{t+1} \leftarrow \delta_t - \eta \cdot \frac{\partial L_t}{\partial \delta_t}$
end for

Output: Accepted Top-K items or conversation ends at round T

As conversations progress, our third component *User Preferences Refining* iteratively refines user preferences by incorporating causal-based attributes a_{cau}^t, recommended items, and online feedback. User feedback is categorized into positive feedback (accepted at-

Algorithm 2 Causal-based Attribute Selector

Input: Dataset $\mathscr{X} = \{x_q\} = \{(u_q, t_q, v_q, b_q)\}$. Refined user representation $\boldsymbol{e}_u^t$ and $\boldsymbol{e}_v^t$.

Initialize $max_{ATE} = 0$
Split $\mathscr{X}$ into T subgroups based on round number t
for each attribute a **do**
  for $t = 1, 2, 3 \ldots T$ **do**
    Compute $X_t' = \{x_q \mid u \in \mathscr{U}_1 \cup \mathscr{U}_2\}$ by Eq (3.6)
    Generate $\mathscr{X}_{(a,t)}$ by selecting users who were asked for a at round t from $\mathscr{X}$
    for each sample x_q in the treatment group of X_t' **do**
      $l_{u,v} = \frac{1}{2}\left(1 - \frac{\boldsymbol{e}_u^t \boldsymbol{e}_v^t}{\|\boldsymbol{e}_u^t\| \|\boldsymbol{e}_v^t\|}\right)$
      Select the matched user $x_{q'}$ by minimizing $l_{u,v}$
      Compute $\hat{b}_q\left(t_q = 1\right)$ and $\hat{b}_q\left(t_q = 0\right)$ by Eq (4.6).
    end for
    $\text{ATE}_{user} = \sum_{x_q \in X_t'} \left(\hat{b}_q\left(t_q = 1\right) - \hat{b}_q\left(t_q = 0\right)\right)$
  end for
  $\text{ATE}_{a,t} = \sum_{x_q \in \mathscr{X}_{(a,t)}} \frac{\text{ATE}_{user}}{|\mathscr{X}_{(a,t)}|}$
  if $\text{ATE}_{a,t} > max_{ATE}$ **then**
    $max_{ATE} = \text{ATE}_{a,t}$
    $a_{cau}^t = a$
  end if
end for

Output: Causal-based attribute a_{cau}^t per round

tribute set $\mathscr{A}_u^+$) and negative feedback (rejected attribute set $\mathscr{A}_u^-$ and rejected item set $\mathscr{I}_u^-$), represented as $\boldsymbol{e}_{a^+}$, $\boldsymbol{e}_{a^-}$, and $\boldsymbol{e}_i^-$ respectively. Specifically, the refinement process consists of three sequential operations: (1) Input Layer to collect and encode user feedback; (2) Feedback Aggregation Layer to summarize and aggregate collected feedback; and (3) Update Layer to refine the user representation based on the aggregated feedback. Online feedback is round-specific, denoted by subscript t. For the Input Layer, we encode feedback into matrices:

$$\boldsymbol{R}_{\mathscr{A}_u^+} = \begin{bmatrix} \boldsymbol{e}_{a^+} \\ \vdots \\ \boldsymbol{e}_{a^+} \end{bmatrix}, \boldsymbol{R}_{\mathscr{A}_u^-} = \begin{bmatrix} \boldsymbol{e}_{a^-} \\ \vdots \\ \boldsymbol{e}_{a^-} \end{bmatrix}, \boldsymbol{R}_{\mathscr{I}_u^-} = \begin{bmatrix} \boldsymbol{e}_{i^-} \\ \vdots \\ \boldsymbol{e}_{i^-} \end{bmatrix} \tag{3.13}$$

where $\boldsymbol{R}_{\mathscr{A}_u^+}$, $\boldsymbol{R}_{\mathscr{A}_u^-}$ and $\boldsymbol{R}_{\mathscr{A}_u^+}$ denote the matrix representations of accepted attributes, rejected attributes, and rejected items respectively. Then, we aggregate positive feedback us-

ing attention mechanisms:

$$\boldsymbol{R}^{+} = \text{softmax}\left(\frac{\boldsymbol{R}_{\mathscr{A}_u^+}\boldsymbol{W}^j\left(\boldsymbol{R}_{\mathscr{A}_u^+}\right)^T\boldsymbol{W}^k}{\sqrt{d}}\right)\boldsymbol{R}_{\mathscr{A}_u^+}\boldsymbol{W}^f \tag{3.14}$$

where $\boldsymbol{R}^{+}$ is the aggregated representation of all positive feedback, encompassing contextual information from neighboring positive feedback. $\boldsymbol{W}^j$, $\boldsymbol{W}^k$, and $\boldsymbol{W}^f$ are projection matrices [106]. softmax is the softmax activation applied to the attention score. On the other hand, given a matrix $\boldsymbol{R}_{ne}$ including rejected items and rejected attributes as $\boldsymbol{R}_{ne} = \boldsymbol{R}_{\mathscr{A}_u^-} \oplus \boldsymbol{R}_{\mathscr{I}_u^-}$. Similarly, we aggregate the representation of negative feedback as

$$\boldsymbol{R}^{-} = \text{softmax}\left(\frac{\boldsymbol{R}_{ne}\boldsymbol{W}^j\left(\boldsymbol{R}_{ne}\right)^T\boldsymbol{W}^k}{\sqrt{d}}\right)\boldsymbol{R}_{ne}\boldsymbol{W}^f \tag{3.15}$$

where $\boldsymbol{R}^{-}$ is the aggregated representation of all negative feedback, capturing context from rejected items and attributes.

Finally, we update the user representation based on the aggregated representation of online feedback as

$$\begin{aligned}\boldsymbol{e}_u^{+} &= \boldsymbol{R}^{+} \oplus \boldsymbol{e}_u \\ \boldsymbol{e}_u^{-} &= \boldsymbol{R}^{-} \oplus \boldsymbol{e}_u \\ \boldsymbol{e}_u^{t} &= \boldsymbol{e}_u + \boldsymbol{e}_u^{+} - \boldsymbol{e}_u^{-}\end{aligned} \tag{3.16}$$

where $\boldsymbol{e}_u^{+}$ and $\boldsymbol{e}_u^{-}$ denote the updated user representation that incorporates all positive feedback and negative feedback, respectively. $\boldsymbol{e}_u^{t}$ represents the refined user preference at round t, capturing long-term preferences from historical data and short-term preferences from the current conversation.

With the refined user preference representation $\boldsymbol{e}_u^{t}$, our last component ***Recommendations*** aims to make Top-K recommendations to the user by optimizing a pairwise Bayesian Personalized Ranking (BPR) loss:

$$\mathscr{L}_{bpr} = \sum_{(u,i)\in\mathscr{D}} -\ln\sigma\left((\boldsymbol{e}_u^{t})^T\boldsymbol{e}_i - (\boldsymbol{e}_u^{t})^T\boldsymbol{e}_i'\right) + \lambda_\theta|\theta|^2 \tag{3.17}$$

where $\mathscr{D}$ represents the user-item training pairs, θ denotes model parameters with regularization coefficient λ_θ, and $(\boldsymbol{e}_u^{t})^T\boldsymbol{e}_i$ indicates the predicted preference score between user u and item i. When the system chooses to make direct item recommendations, as determined by the action selection process, we recommend the Top-K items with the highest predicted scores to the user. However, if the system decides to inquire about additional attributes, we proceed with the *Causal-based Attribute Selection* component and continue the conversational loop.

For cost-effective evaluations, we employ a user simulator [157, 92, 217, 246] that emulates real user interactions by processing system queries and generating appropriate responses. The simulator takes historical preferences as input and produces feedback on attribute preferences and item recommendations. This enables efficient training and evaluation of the model's ability to adapt recommendations to evolving user preferences while avoiding costly real-user studies. The recommendation quality is assessed by comparing system outputs against the simulator's ground truth preferences.

3.1.3 Experiments

3.1.3.1 Datasets

We evaluate our model on three benchmark datasets: *Yelp*[1], *Douban-Book*[2] and *MovieLens*[3]. The statistics of these datasets are summarized in Table 3.1. All three datasets include users' attributes, social relationships, and user-item interaction data with the ratings. To ensure data quality, we filter users with fewer than 10 interactions and remove infrequent attributes. The remaining data is then split into training/testing/validation sets with ratios of 70%/20%/10%, respectively.

Table 3.1: CCR: Statistical details of three datasets.

Statistics	**Yelp**	**Douban-Book**	**MovieLens**
# User	27,675	13,024	943
# Item	70,311	22,347	1,682
# Interaction	1,368,606	792,062	100,000
# Density	0.07%	0.27%	6.30%

3.1.3.2 Baselines and Evaluation

For implementation, we set the embedding dimension to 128 with regularization parameter $1e-5$, using Adam optimizer with learning rate tuned in [0.0001, 0.1]. Baseline models are tuned using the same hyperparameter range as ours for fair comparison. For recommendation performance evaluation, we employ two widely used metrics Success Rate (SR) and Average Turns (AT) [93, 217, 94]. SR@t measures the percentage of successful recommendations achieved in t conversation rounds, and Average Turns (AT) indicates the average number of rounds needed to reach a successful recommendation outcome.

[1]https://www.yelp.com/dataset/
[2]https://github.com/librahu/HIN-Datasets-for-Recommendation-and-Network-Embedding
[3]https://github.com/librahu/HIN-Datasets-for-Recommendation-and-Network-Embedding

We compare CCR with the following state-of-the-art CRSs to verify its effectiveness:

EAR [93]: integrates estimation, action and reflection stages through reinforcement learning to estimate user preferences in CRS.

UNICORN [41]: adopts a unified approach for multi-stage decision optimization to enhance the CRS.

FPAN [217]: designs gating modules for negative and positive feedback processing, and then adapts user preferences based on online feedback in CRS.

SCPR [94]: converts the decision-making in CRS into a path-finding issue via Deep Q-learning.

CRM [157]: utilizes a belief tracker and trains an RL-based policy network to guide user-item interactions in CRSs.

CECR [232]: generates counterfactual samples of negative terms to enhance the robustness of CRSs.

Max Entropy (MaE): selects attributes with the highest entropy to ask the user each round.

3.1.3.3 Result Analysis

Performance Comparison: Table 3.2 demonstrates CCR's performance compared to state-of-the-art baselines across three datasets. Obviously, CCR consistently achieves superior results, with notable improvements in SR@20 of 4.66%, 2.06%, and 1.63% on *Yelp*, *Douban-Book*, and *MovieLens*, respectively. Moreover, the performance advantage increases with conversation rounds, indicating effective preference refinement through feedback integration. Notably, CCR requires fewer conversation turns to reach successful recommendations, as evidenced by the lowest AT values among all models. These results validate the effectiveness of our causal-based attribute selection strategy and debiasing approach in accurately capturing user preferences in CRSs.

Ablation Study: To evaluate component contributions, we conduct ablation experiments on *MovieLens* and *Yelp* in terms of SR@20 and AT. *A* denotes removing the stratification, which can address the time confounders. *B* denotes removing the matching, which can address the user attribute confounders. *C* denotes removing the causal-based attribute selection. *D* denotes removing the user preference refinement. *E* denotes the complete CCR model. As shown in Figure 3.3, removing stratification and matching lead to a notable decline in both the SR@20 and AT, which highlights the importance of addressing time and user attribute confounders in CRSs. Also, ignoring the user preference refinement leads to a marked performance decrease, emphasizing its considerable importance. The causal-

Table 3.2: CCR: Recommendation performance comparison: the best results are bolded, while the best baselines are underlined.

Dataset	**Yelp**				**Douban-Book**				**MovieLens**			
ModelMetrics	SR@5	SR@10	SR@20	AT	SR@5	SR@10	SR@20	AT	SR@5	SR@10	SR@20	AT
EAR	0.793	0.818	0.859	4.882	0.812	0.833	0.866	4.821	0.844	0.871	0.943	4.313
FPAN	0.811	0.851	0.911	4.662	0.833	0.867	0.918	4.515	0.851	0.888	0.935	4.114
UNICORN	0.805	0.852	0.903	5.012	0.836	0.869	0.914	4.891	0.843	0.865	0.937	4.172
SCPR	0.805	0.868	.922	4.782	0.842	0.861	0.929	4.231	0.846	0.875	0.959	4.124
CRM	0.657	0.773	0.812	5.014	0.744	0.804	0.889	4.971	0.724	0.805	0.852	4.673
CECR	.852	.901	0.920	.319	.866	.922	.973	.229	.891	.949	.982	.775
MaE	0.553	0.578	0.613	7.742	0.581	0.589	0.602	7.221	0.612	0.645	0.689	7.013
Our CCR	**0.863**	**0.929**	**0.965**	**4.188**	**0.872**	**0.948**	**0.993**	**4.201**	**0.901**	**0.962**	**0.998**	**3.526**
Improv. %	1.29%	3.11%	4.66%	3.03%	0.69%	2.82%	2.06%	0.66%	1.12%	1.37%	1.63%	6.60%

based attribute selection emerged as the most critical component, evidenced by the most significant performance drop when removed. These results confirm the effectiveness of our causal-based attribute selection in exploring causal drivers of user behavior, which can significantly enhance the robustness of CRSs.

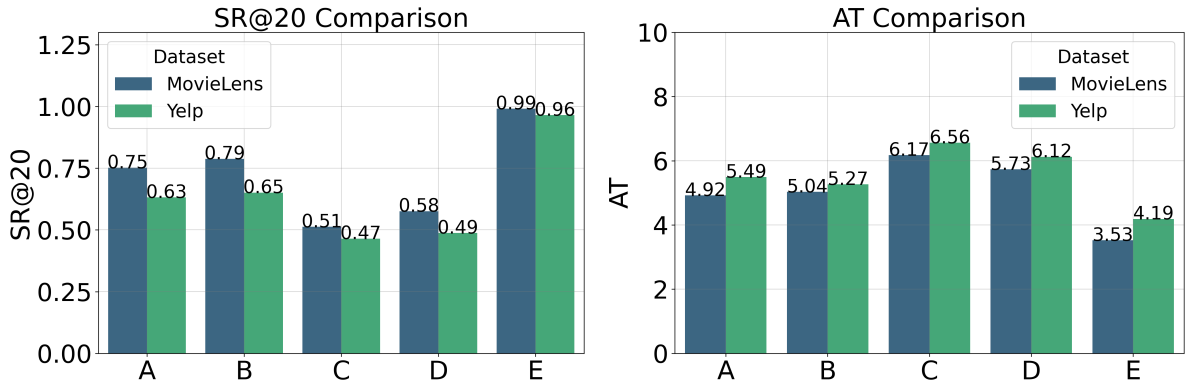


Figure 3.3: CCR: Ablation study.

Effective of Causal-based Attribute: To evaluate our causal-based attribute selection strategy, we compare CCR's SR@20 performance against baselines across multiple conversation rounds on *MovieLens* and *Yelp*. As shown in Figure 3.4, we find that CCR consistently achieves superior performance throughout all rounds. This sustained advantage stems from our method's ability to identify attributes with true causal impact on recommendations, rather than randomly choose one attribute. Moreover, the performance gap widens in later rounds, indicating that causal-based selection enables more accurate preference modeling and leads to increasingly personalized recommendations as conversations progress. These results highlight the importance of exploring causal factors, particularly in the context of attribute selection, for developing a more robust CRS.

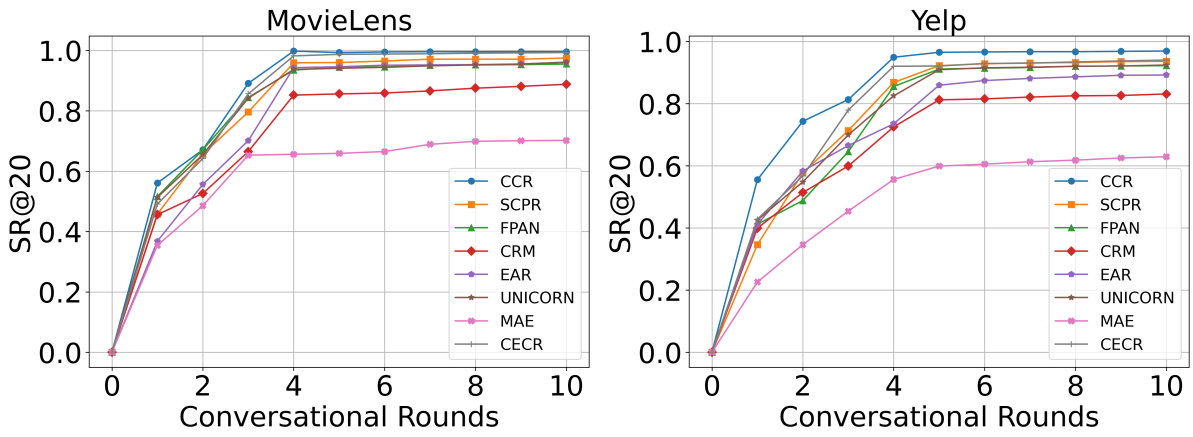


Figure 3.4: CCR: The effectiveness of our causal-based attributes.

Personalization Comparison: To evaluate CCR's personalization capability, we compare the asking likelihood of Top-40 frequent attributes across different models. The asking likelihood of an attribute refers to the percentage of users predicted by the system who like that attribute. As shown in Figure 3.5, UNICORN shows a high likelihood of 0.78 for Top-10 attributes, which means UNICORN predicts 78% of users like the Top-10 frequent attributes. In contrast, our CCR shows a lower average asking likelihood of 0.51 for Top-10 frequent attributes, suggesting that attribute selection is more personalized than commonly popular attributes. Remarkably, the CCR consistently had the lowest query likelihood of the Top-40 attributes across all baselines, highlighting the CCR's ability to personalize better than the other baselines.

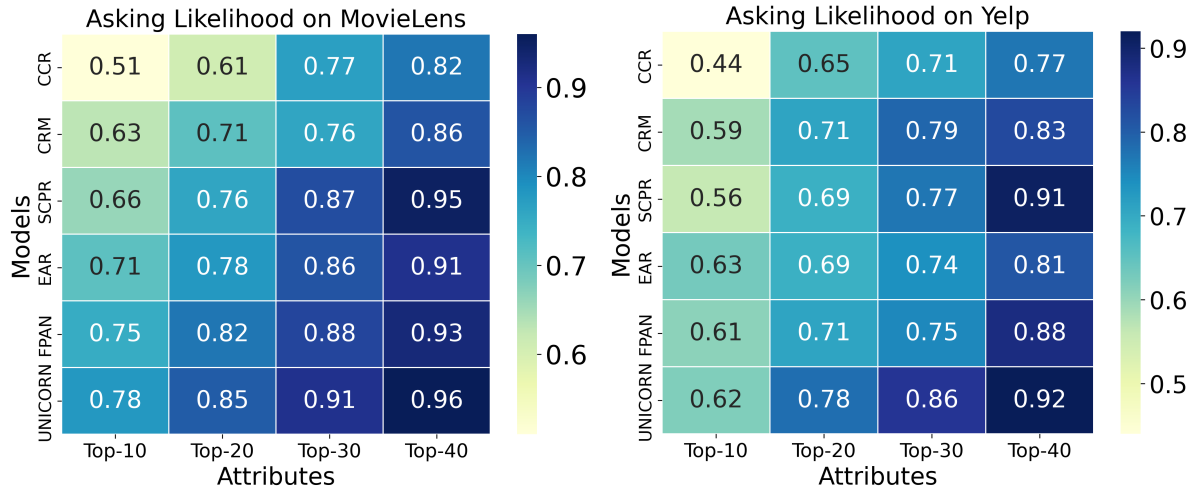


Figure 3.5: CCR: Comparison of personalization capabilities.

Time Complexity Comparison: To evaluate our CCR's computational efficiency, we compare recommendation generation times with other baselines. As shown in Figure 3.6, CCR achieves the fastest response time of 0.71 seconds through its efficient three-layer feedback aggregation mechanism, surpassing all baselines: UNICORN (1.44s), EAR (1.12s), CRM (1.05s), FPAN (0.88s), SCPR (0.77s), CECR (0.74s), and MaE (1.51s). The performance differences primarily stem from varying algorithmic complexities, with MaE being the slowest due to its exhaustive attribute exploration approach. These results demonstrate CCR's exceptional ability to maintain computational efficiency while delivering high-quality recommendations.

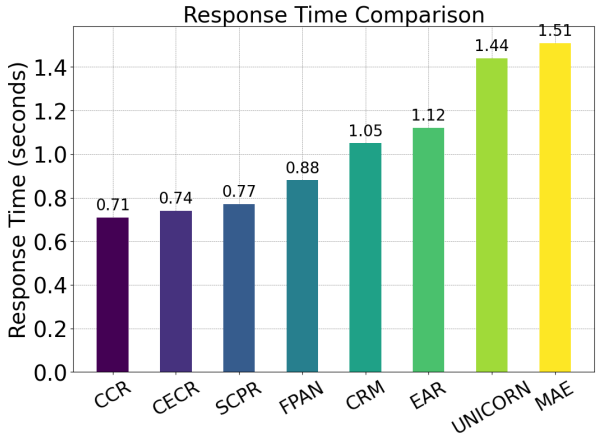


Figure 3.6: CCR: Time complexity comparison.

3.1.3.4 Summary

In this work, we propose CCR, a novel causal framework that addresses RQ1 regarding bias mitigation in conversational recommender systems. By synergistically integrating stratification and matching techniques of causal inference, CCR effectively mitigates both time and user attribute confounders while accurately identifying causally significant attributes through ATE estimation. Extensive experiments demonstrate CCR's superior performance over state-of-the-art baselines in terms of recommendation accuracy, personalization, and computational efficiency. Future work will explore item-level causal factors to provide deeper insights into user preference modeling in conversational recommendations.

3.2 LLMs Meet Causal Inference: Semantic-Rich Dual Propensity Score for Sequential Recommendation

3.2.1 Overview

Sequential Recommender Systems (SRSs) have emerged as a pivotal tool for analyzing user-item interaction sequences to suggest relevant items to users [70, 132]. However, a fundamental challenge in SRSs is the exposure bias, which occurs when users are under- or over-exposed to certain items [193], It can create a self-reinforcing feedback loop where popular items gain more visibility while less exposed items struggle, regardless of their true relevance [11]. Current approaches primarily focus on analyzing sequential dependencies among items to address exposure biases, but they often overlook user-side exposure bias arising from varying user activity levels and neglect rich semantic information embedded in item content [178, 177]. This can significantly limit the SRSs's ability to understand less active users' preferences and accurately capture user interests in less exposed items, leading to suboptimal recommendations.

3.2.1.1 Research Objective

This study aims to address RQ1 by developing an innovative causal-guided framework that fully addresses exposure biases caused by spurious correlation in sequential recommendations. Generally, we propose to synergistically integrate LLMs' semantic modeling capabilities with causal inference's debiasing mechanisms through dual propensity scoring. This integration enables us to simultaneously capture rich semantic patterns from textual data while mitigating exposure bias from both item popularity and user activity perspectives, thereby providing more accurate and unbiased sequential recommendations.

3.2.1.2 The Proposed Method

To achieve the aforementioned objectives, we propose a causal-based approach called **L**LM-enhanced **D**ual **P**ropensity Score **E**stimation (**LDPE**) with three synergistic components:

- ***Unbiased Representation Learning*** employs LLMs to generate semantically rich user/item embeddings from textual data, then integrates collaborative information to enhance generalizability and reduce exposure biases inherent in LLM training data for unbiased LLM-based user/item embeddings.

- ***Dual Propensity Score Estimation*** uses unbiased LLM-based embeddings to estimate propensity scores from both item and user sides, which includes interaction timestamps as exposure patterns can change over time, thus fully addressing exposure biases.
- ***Transformer*** processes item and user sequences while integrating temporal factors and dual propensity scores to accurately model users' true preferences.

The key contributions of this research are summarized as follows:

- We highlight the limitations of existing SRSs in fully mitigating exposure bias, specifically their insufficient understanding of interaction semantics and focus on only item-side exposure bias while neglecting user-side.
- We propose a novel approach called LDPE, which leverages LLMs and causal inference to fully mitigate exposure bias stemming from both item popularity and varying user activity levels in SRSs.
- We integrate collaborative information from collaborative filtering into LLMs to balance overrepresented trends, enhancing generalizability and reducing exposure biases for unbiased LLM-based embeddings.
- We conduct extensive experiments to demonstrate the superiority of LDPE over state-of-the-art baselines.

3.2.2 LDPE

3.2.2.1 Problem Definition

Let $\mathscr{U}$ and $\mathscr{I}$ be the sets of users and items, respectively. Given a user $u \in \mathscr{U}$ and his/her historical interactions $\mathscr{D} = (u, i, t)$, where $i \in \mathscr{I}$ and each tuple (u, i, t) denotes user u interacted with item i at time t. Our goal is to predict the next sequential item i_{t+1} that the user is most likely to interact with at time $t+1$.

To achieve accurate predictions, it's crucial to fully mitigate exposure biases in the sequential data from both the item and user sides, while integrating rich interaction semantic. Specifically, we employ LLMs to generate semantic-rich LLM-based item/user embeddings $\boldsymbol{e}_i^{LLM}$ / $\boldsymbol{e}_u^{LLM}$ from textual data, then align them with CF-based item/user embeddings to obtain unbiased embeddings $\boldsymbol{e}_i$ / $\boldsymbol{e}_u$. Then, we introduce our causal graph as Figure 3.7 showing causal relationships between user U, item I, exposure E, time T, and rating Y.

Based on this graph, we construct for each interaction (u,i,t) an item sequence $\boldsymbol{h}_u^{<t} = [i_1,\ldots,i_K]$ capturing items user u interacted with before t, and a user sequence $\boldsymbol{h}_i^{<t} = [u_1,\ldots,u_K]$ containing users who interacted with item i before t. Using unbiased LLM-based embeddings $\boldsymbol{e}_i/\boldsymbol{e}_u$ of items/users in these sequences, we estimate dual propensity scores $P\left(u,\boldsymbol{h}_i^{<t}\right)$ and $P\left(i,\boldsymbol{h}_u^{<t}\right)$, accounting for item-side and user-side exposure bias, respectively. These scores enable us to adjust interaction importance by weighting less exposed items/users higher while considering temporal effects. In such way, we can fully mitigate exposure biases from both the item and user sides while incorporating rich semantic information for robust sequential recommendations.

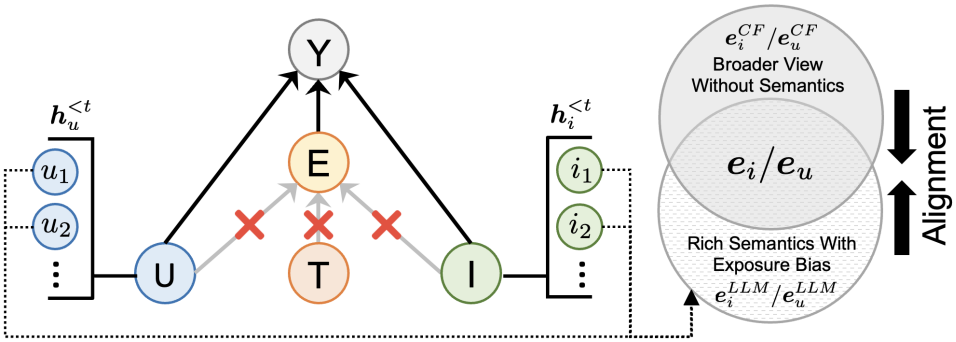


Figure 3.7: Our designed causal graph for SRSs.

3.2.2.2 Methodology

Figure 3.8 shows the framework of our proposed LDPE, which has three components. The first component ***Unbiased Representation Learning*** constructs semantically rich yet unbiased item/user embeddings for estimating propensity scores. Although LLMs excel at extracting semantic information from texts to generate embeddings [98], their embeddings inherit exposure biases in the training data that overrepresents popular items [19, 212]. To address this issue of LLMs, we integrate collaborative filtering signals to balance these representation-level exposure biases and enhance the generalizability of LLM-based embeddings. As shown in Figure 3.9, we first generate descriptive texts using predefined templates. For items, we replace placeholders with item names (e.g., "<The Avengers>"), while for users, we incorporate all available information of users (e.g., name, age, gender, interaction history). These texts are processed by an LLM encoder to obtain LLM-based item/user embeddings by averaging the hidden states from the last layer as

$$\boldsymbol{e}_i^{LLM} = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{LLM}(i)_n, \boldsymbol{e}_u^{LLM} = \frac{1}{M}\sum_{m=1}^{M}\boldsymbol{LLM}(u)_m \tag{3.18}$$

where $\boldsymbol{e}_i^{LLM} \in \mathbb{R}^d$ is the LLM-based item representation for item i, capturing rich semantics from the textual data of the item i. $\boldsymbol{e}_u^{LLM}$ is the LLM-based user embeddings of user u, capturing rich semantics from all available information of user u. N represents the total number of tokens in an item's description, with each token's hidden state from an LLM denoted as $\boldsymbol{LLM}(i)_n$. M is the total number of tokens in a user's description, with each token's hidden state from an LLM represented as $\boldsymbol{LLM}(u)_m$.

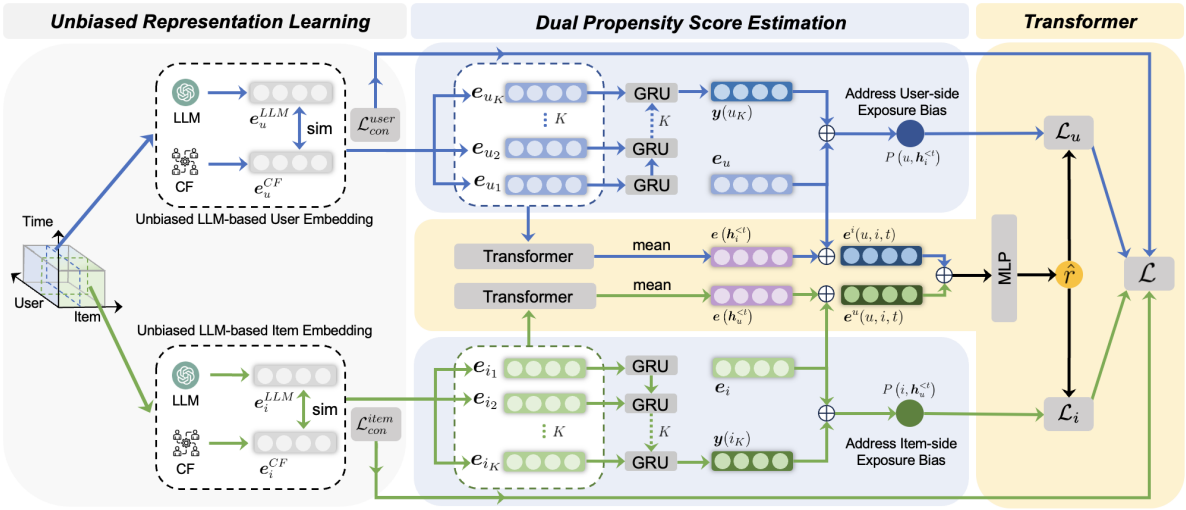


Figure 3.8: The overall framework of our proposed method LDPE.

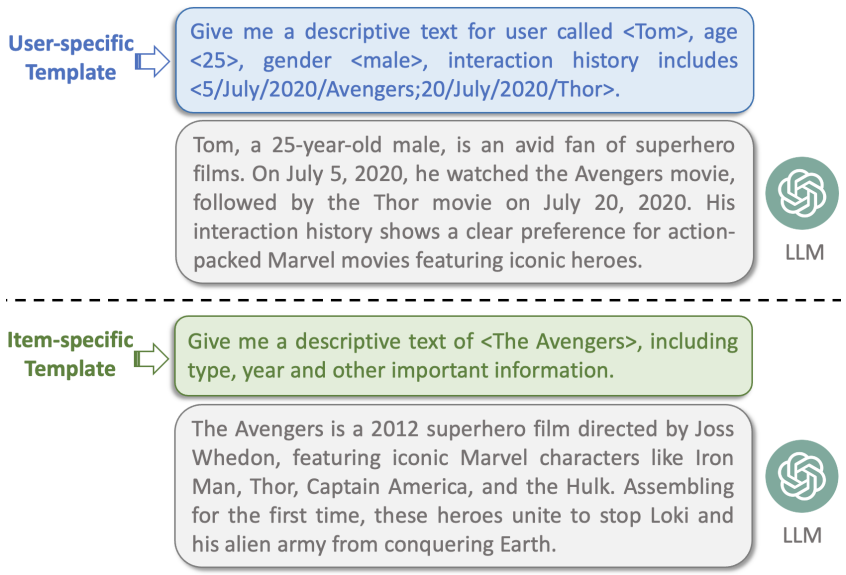


Figure 3.9: Examples of generating descriptive texts for the target item/user using predefined prompt templates.

To mitigate the inherent exposure biases in these semantic-rich LLM-based embeddings, we learn Collaborative Filtering (CF)-based representations [198, 135] that capture generalized interaction patterns using a collaborative filtering model $\mathscr{R}$:

$$\boldsymbol{e}_i^{CF} = \mathscr{R}(\boldsymbol{e}_i^{int}) \quad , \quad \boldsymbol{e}_u^{CF} = \mathscr{R}(\boldsymbol{e}_u^{int}) \tag{3.19}$$

where CF-based item embedding $\boldsymbol{e}_i^{CF}$ and user embedding $\boldsymbol{e}_u^{CF}$ capture generalized interaction patterns across multiple user-item interactions. $\boldsymbol{e}_u^{int}$ and $\boldsymbol{e}_i^{int}$ are initial user and item embedding, derived using a retrieval function [137] to include all associated attributes, which are then numerically encoded for compatibility.

Following that, we align LLM-based and CF-based embeddings through contrastive learning [224] to preserve semantic richness while mitigating inherent exposure biases in LLMs' training text. Generally, we align the CF-based and LLM-based embeddings of the same user/item as positive pairs, while treating the embeddings of different user/item as negative pairs. Mathematically, we have

$$\begin{aligned} \mathscr{L}_{con}^{item} &= -\frac{1}{|\mathscr{P}|} \sum_{(i,u)\in\mathscr{P}} \log \frac{\exp(\text{sim}(\sigma_{\downarrow}(\boldsymbol{e}_i^{LLM}), \boldsymbol{e}_i^{CF}))}{\sum_{i\in\mathscr{N}_i} \exp(\text{sim}(\sigma_{\downarrow}(\boldsymbol{e}_i^{LLM}), \boldsymbol{e}_i^{CF}))} \\ \mathscr{L}_{con}^{user} &= -\frac{1}{|\mathscr{P}|} \sum_{(i,u)\in\mathscr{P}} \log \frac{\exp(\text{sim}(\sigma_{\downarrow}(\boldsymbol{e}_u^{LLM}), \boldsymbol{e}_u^{CF}))}{\sum_{u\in\mathscr{N}_u} \exp(\text{sim}(\sigma\downarrow(\boldsymbol{e}_u^{LLM}), \boldsymbol{e}_u^{CF}))} \end{aligned} \tag{3.20}$$

where $\mathscr{L}_{con}^{item}$ and $\mathscr{L}_{con}^{user}$ denote the contrastive losses for items and users. Sets $\mathscr{N}_u$ and $\mathscr{N}_i$ contain non-interacting users and items used as negative samples for user u and item i, respectively. $\mathscr{P}$ represents positive interaction pairs. The contrastive losses aim to maximize the similarity between the corresponding LLM-based and CF-based embeddings for pairs in $\mathscr{P}$, while minimizing similarity with the embeddings in $\mathscr{N}_u$ and $\mathscr{N}_i$. sim$(\cdot)$ is the similarity function to compute cosine similarity among embeddings, and $\sigma\downarrow$ is a multilayer perceptron aligning LLM-based embeddings to the CF-based feature space. By minimizing these two contrastive losses, we compute the final unbiased LLM-based item/user embedding as

$$\boldsymbol{e}_i = \boldsymbol{e}_i^{LLM} - \alpha\nabla\boldsymbol{e}_i^{LLM}\mathscr{L}_{con}^{item} \quad , \quad \boldsymbol{e}_u = \boldsymbol{e}_u^{LLM} - \alpha\nabla\boldsymbol{e}_u^{LLM}\mathscr{L}_{con}^{user} \tag{3.21}$$

where α is the parameter balance the influence of LLM-based and CF-based embeddings. $\nabla\boldsymbol{e}_i^{LLM}$ and $\nabla\boldsymbol{e}_u^{LLM}$ are the gradients of the contrastive loss with respect to the $\boldsymbol{e}_i^{LLM}$ and $\boldsymbol{e}_u^{LLM}$. $\boldsymbol{e}_i$ and $\boldsymbol{e}_u$ are the generated unbiased LLM-based embeddings for item i and user u, respectively, capturing rich semantics while mitigating the inherent exposure biases in LLMs' training texts.

The second component ***Dual Propensity Score Estimation*** leverages the estimated unbiased LLM-based embeddings to compute dual propensity scores from both item and user sides. These propensity scores estimate the likelihood of observing a user-item interaction in a given situation, which can help address exposure biases from item popularity and varying user activity levels by reweighting each interaction's importance [218, 211]. Moreover, as exposure patterns may change over time and introduce temporal confounding effects, thus we integrate interaction timestamps into the propensity score calculations to address temporal confounding effects as well.

Specifically, we employ two Gated Recurrent Unit (GRU) networks [32] to process the item and user sequences separately. Since GRUs have been proven for efficiently capturing sequential dependencies, making them ideal for modeling user-item sequential interactions [234]. For the item-side propensity score, we extract the item interaction sequence $\boldsymbol{h}_u^{<t} = [i_1, \ldots, i_k, \ldots, i_K]$, where i_k is the k-th item that user u interacted with before time t. K is the total item number. The GRU processes this sequence with unbiased LLM-based embeddings $\boldsymbol{e}_{i_k}$ (e.g., $\boldsymbol{e}_i$) from Eq (3.21) to output the item sub-sequence representation:

$$\boldsymbol{y}(i_k), \boldsymbol{z}_k = \mathrm{GRU}\left(\boldsymbol{e}_{i_k}, \boldsymbol{z}_{k-1}\right) \tag{3.22}$$

where $\boldsymbol{z}_k$ and $\boldsymbol{z}_{k-1}$ denote the GRU hidden states at layers k and $k-1$ respectively. GRU is the GRU cell to processes the sequence. The output item sub-sequence representation $\boldsymbol{y}(i_k)$ encodes the sequential dependencies for items $[i_1, \ldots, i_k]$. After processing the complete sequence, we compute the item-side propensity score $P\left(i, \boldsymbol{h}_u^{<t}\right)$ as

$$P\left(i, \boldsymbol{h}_u^{<t}\right) = \max\left(\frac{\exp\left(\boldsymbol{e}_i^\top \boldsymbol{y}(i_K)\right)}{\sum_{i' \in \mathscr{I}} \exp\left(\boldsymbol{e}_{i'}^\top \boldsymbol{y}(i_K)\right)}, Q\right) \tag{3.23}$$

where $P\left(i, \boldsymbol{h}_u^{<t}\right)$ represents the likelihood of user u interacting with item i at time t, given their historical sequence $\boldsymbol{h}_u^{<t}$. The threshold Q ensures numerical stability by preventing extreme propensity values.

For the user-side propensity score, we also extract the user interaction sequence $\boldsymbol{h}_i^{<t} = [u_1, \ldots, u_k, \ldots, u_K]$, which captures the sequence of users who interacted with item i before time t, where K denotes the number of users. Likewise, the GRU processes $\boldsymbol{h}_i^{<t}$ sequentially using unbiased LLM-based user embeddings $\boldsymbol{e}_{u_k}$ (e.g., $\boldsymbol{e}_u$) obtained from Eq (3.21)) to generate the user sub-sequence representation:

$$\boldsymbol{y}(u_k), \boldsymbol{z}_k = \mathrm{GRU}\left(\boldsymbol{e}_{u_k}, \boldsymbol{z}_{k-1}\right) \tag{3.24}$$

where $\boldsymbol{z}_k$ and $\boldsymbol{z}_{k-1}$ denote the GRU hidden states at layers k and $k-1$ respectively. The output user sub-sequence representation $\boldsymbol{y}(u_k)$ encodes the sequential dependencies for

users $[u_1, \ldots, u_k]$ who previously interacted with item i. Upon processing the complete sequence, we compute the user-side propensity score $P\left(u, \boldsymbol{h}_i^{<t}\right)$ as

$$P\left(u, \boldsymbol{h}_i^{<t}\right) = \max\left(\frac{\exp\left(\boldsymbol{e}_u^{\top} \boldsymbol{y}\left(u_K\right)\right)}{\sum_{u' \in \mathscr{U}} \exp\left(\boldsymbol{e}_{u'}^{\top} \boldsymbol{y}\left(u_K\right)\right)}, Q\right) \tag{3.25}$$

where $P\left(u, \boldsymbol{h}_i^{<t}\right)$ represents the likelihood of user u interacting with item i at time t, conditioned on the item's historical user interaction sequence $\boldsymbol{h}_i^{<t}$.

The third component ***Transformer*** leverages transformer architecture as our model backbone to integrate the estimated dual propensity scores for accurate preference modeling. Given transformers' proven capability in capturing complex temporal dependencies [207, 37], they serve as an ideal choice for sequential recommendation tasks. Through parallel processing of item and user sequences, the transformer generates comprehensive user and item embeddings that encode both temporal dynamics and interaction patterns. These embeddings are then fusion through a multi-layer perceptron (MLP) to predict user preferences, effectively mitigating exposure bias and temporal confounding effects for unbiased sequential recommendations.

Specifically, we first use two transformers to process the item sequence $\boldsymbol{h}_u^{<t}$ and user sequence $\boldsymbol{h}_i^{<t}$ separately. For a target interaction tuple (u, i, t), we compute the overall item representation as the concatenation of unbiased LLM-based item embedding and user historical sequence embedding:

$$\boldsymbol{e}^u(u, i, t) = \left[\boldsymbol{e}\left(\boldsymbol{h}_u^{<t}\right) \oplus \boldsymbol{e}_i\right] \tag{3.26}$$

where $\boldsymbol{e}^u(u, i, t)$ captures both item semantics and sequential dependencies. $\oplus$ is the concatenation. $\boldsymbol{e}\left(\boldsymbol{h}_u^{<t}\right)$ is the vector that encodes the sequence $\boldsymbol{h}_u^{<t}$ by averaging the transformer's output:

$$\boldsymbol{e}\left(\boldsymbol{h}_u^{<t}\right) = \text{Mean}\left(\text{Transformer}\left(\left[\boldsymbol{e}_{i_1}, \ldots, \boldsymbol{e}_{i_K}\right]\right)\right) \tag{3.27}$$

where Mean is the mean pooling operation. Transformer is a transformer architecture. Likewise, the overall user representation is obtained by concatenating the unbiased LLM-based user embedding and item historical sequence embedding:

$$\boldsymbol{e}^i(u, i, t) = \left[\boldsymbol{e}\left(\boldsymbol{h}_i^{<t}\right) \oplus \boldsymbol{e}_u\right] \tag{3.28}$$

where $\boldsymbol{e}^i(u, i, t)$ is the overall user representation of the tuple (u, i, t), capturing both the semantics of the user u and sequential dependencies from the item i's past interactions. $\boldsymbol{e}\left(\boldsymbol{h}_i^{<t}\right)$ is the vector that encodes the sequence $\boldsymbol{h}_i^{<t}$ by averaging the transformer's output:

$$\boldsymbol{e}\left(\boldsymbol{h}_i^{<t}\right) = \text{Mean}\left(\text{Transformer}\left(\left[\boldsymbol{e}_{u_1}, \ldots, \boldsymbol{e}_{u_K}\right]\right)\right) \tag{3.29}$$

Next, we predict user preferences by feeding the concatenated item and user representations into an MLP:

$$\hat{r} = \sigma\left(\text{MLP}\left(\boldsymbol{e}^u(u,i,t) \oplus \boldsymbol{e}^i(u,i,t)\right)\right) \tag{3.30}$$

where MLP is a multi-layer perception outputting the preference score $\hat{r}$ for the target tuple (u,i,t). σ is the sigmoid function mapping the preference score $\hat{r}$ to a probability between 0 and 1.

To train our model effectively, we design an unbiased learning objective incorporating propensity scores and contrastive alignment. First, we define the user-side $\mathscr{L}_u$ and item-side $\mathscr{L}_i$ propensity losses:

$$\mathscr{L}_u = \sum_{(u,i,t)\in\mathscr{D}} \frac{\delta(c,\hat{r})}{P\left(i,\boldsymbol{h}_u^{<t}\right)} \quad , \quad \mathscr{L}_i = \sum_{(u,i,t)\in\mathscr{D}} \frac{\delta(c,\hat{r})}{P\left(u,\boldsymbol{h}_i^{<t}\right)} \tag{3.31}$$

where δ is the cross-entropy loss between ground truth c and prediction $\hat{r}$. $\mathscr{D}$ is the training set. By incorporating the propensity scores $P\left(i,\boldsymbol{h}_u^{<t}\right)$ and $P\left(u,\boldsymbol{h}_i^{<t}\right)$, we reweight the importance of each sample to mitigate exposure biases, assigning higher weights to under-exposed samples and lower weights to over-exposed ones. Then, we combine the $\mathscr{L}_u$ and $\mathscr{L}_i$ with the contrastive alignment of unbiased LLM-based embeddings to obtain the final learning objective:

$$\mathscr{L} = \beta\mathscr{L}_{con}^{item} + \epsilon\mathscr{L}_{con}^{user} + \gamma\mathscr{L}_u + (1-\epsilon-\beta-\gamma)\mathscr{L}_i \tag{3.32}$$

where β, ϵ, and γ are parameters to balance the item contrastive loss, user contrastive loss, and user side loss, respectively. By optimizing $\mathscr{L}$, we jointly learn unbiased item/user representations while mitigating exposure bias from both item and user sides.

3.2.3 Experiments

3.2.3.1 Datasets

We conduct experiments on three datasets: MovieLens-1M[4], Amazon-Books[5], and Amazon-Beauty[6]. As shown in Table 3.3, *MovieLens-1M* has 1 million movie ratings from 6040 users on 3706 movies, *Amazon-Books* has 22 million ratings from 8 million users on 2 million books, and *Amazon-Beauty* has 198502 reviews from 22363 users on 12101 beauty products. For each dataset, we filter out users and items with less than 10 interactions to ensure data quality, then split the data into train/validation/test as 70%/10%/20%, respectively.

[4]https://grouplens.org/datasets/movielens/1m/
[5]http://jmcauley.ucsd.edu/data/amazon/
[6]http://jmcauley.ucsd.edu/data/amazon/

Table 3.3: LDPE: Statistical details of three datasets.

Statistics	**MovieLens-1M**	**Amazon-Book**	**Amazon-Beauty**
#User	6,040	8,000,000	22,363
#Item	3,706	2,000,000	12,101
#Interaction	1,000,000	22,000,000	198,502
#Density	4.47%	0.00014%	0.0733%

3.2.3.2 Baselines and Evaluation

For implementation, we set the embedding dimension to 128, using Adam optimizer with batch size 256. The learning rate is tuned in [0.0001, 0.001, 0.01, 0.1], while parameters α, β, ϵ, and γ are searched in [0, 1] with step size 0.1. The threshold Q is tuned in [0.01, 0.1, 1]. For a fair comparison, we tune the baselines' parameters within the same ranges as ours and employ T5 [31] as our LLM encoder, consistent with the baselines. For evaluation metrics, we adopt three widely used metrics [59, 96, 233]: Pr@K for recommendation quality, Re@K for completeness, and NDCG@K for ranking quality considering item positions.

To evaluate our proposed LDPE, we selected the below state-of-the-art methods:

- **MOJITO** [168]: leverages Gaussian mixtures in attention mechanisms within transformers to enhance sequential recommendations.
- **DRec** [212]: implements dual propensity scoring to reduce exposure biases by considering both items and users in SRSs.
- **ITPS** [37]: uses inverse temporal propensity scores with bidirectional transformers to tackle exposure bias in SRSs.
- **DRO** [225]: applies distributionally robust optimization to system exposure data to mitigate exposure bias.
- **G4Rec** [70]: uses RNNs to detect sequential patterns and facilitate recommendations via session-parallel mini-batch training.
- **NARM** [96]: integrates RNNs with item-level attention to highlight users' primary interests, enhancing sequential recommendations.
- **ReLLa** [107]: employs a retrieval-enhanced method using LLMs to parse extensive user behavior sequences for better recommendations.
- **C4Rec** [256]: merges collaborative filtering with LLMs, adopting a soft+hard prompting strategy to refine recommendation processes.

Table 3.4: LDPE: Recommendation performance comparison: the best results are bolded, while the best baselines are underlined.

Dataset	Metric	MOJITO	DRec	ITPS	G4Rec	NARM	ReLLa	C4Rec	RLMRec	DRO	KHGT	LDPE	Improv %
MovieLens -1M	Pr@5	48.83	47.01	46.42	45.83	48.85	50.25	51.23	49.36	45.92	50.18	**56.62**	10.52%
	Pr@10	51.22	51.53	48.99	47.83	50.15	52.36	53.39	53.68	48.61	54.46	**59.63**	11.09%
	Pr@20	56.51	55.78	51.33	52.36	53.66	56.03	55.78	57.77	51.25	56.81	**62.02**	7.36%
	NDCG@5	55.63	56.66	51.36	48.69	50.68	61.02	58.95	55.78	49.68	57.99	**64.58**	5.83%
	NDCG@10	59.81	58.91	54.47	51.14	56.91	65.51	61.25	58.91	52.36	62.39	**68.36**	4.35%
	NDCG@20	64.57	62.83	59.47	53.24	58.85	68.68	65.52	60.14	55.78	64.86	**70.01**	1.94%
	Re@5	50.12	55.12	48.82	42.63	45.61	59.85	50.18	48.87	43.23	47.63	**63.06**	5.36%
	Re@10	53.52	57.82	52.25	46.86	47.18	63.33	55.58	52.74	45.98	49.69	**66.87**	5.59%
	Re@20	59.89	60.01	56.37	52.13	49.99	65.65	60.04	55.99	48.71	52.39	**71.05**	8.23%
Amazon -Book	Pr@5	35.55	38.88	36.87	40.14	52.22	53.63	55.98	50.05	41.14	44.78	**59.91**	7.02%
	Pr@10	36.17	40.45	39.81	42.69	54.37	56.88	57.76	53.36	44.86	48.17	**62.39**	8.01%
	Pr@20	39.46	44.23	41.32	45.68	58.86	60.05	61.25	58.59	47.98	52.36	**66.66**	8.83%
	NDCG@5	51.59	48.99	41.55	42.13	55.55	58.89	56.18	54.41	52.02	55.66	**61.36**	4.19%
	NDCG@10	59.94	52.23	44.44	43.39	57.69	61.25	58.92	56.39	55.74	58.91	**64.33**	5.03%
	NDCG@20	62.33	58.67	48.61	46.25	60.06	65.33	61.99	59.88	58.49	61.17	**67.02**	2.59%
	Re@5	40.44	38.91	33.65	36.63	47.74	58.88	59.85	57.72	48.87	48.89	**63.88**	6.73%
	Re@10	45.68	45.61	37.54	38.91	49.81	61.78	61.05	59.06	50.07	51.25	**66.74**	8.04%
	Re@20	50.08	48.87	39.99	41.05	52.05	63.57	62.99	60.85	53.67	54.32	**68.81**	8.24%
Amazon -Beauty	Pr@5	40.41	41.58	38.14	41.25	55.38	58.89	58.81	55.56	50.05	47.79	**61.69**	4.75%
	Pr@10	42.68	44.68	41.12	46.36	58.69	62.25	60.09	58.29	53.46	51.36	**63.58**	2.14%
	Pr@20	46.32	47.52	45.68	48.83	60.36	64.44	62.33	60.08	56.65	54.69	**67.77**	5.17%
	NDCG@5	50.12	48.99	44.61	45.56	58.81	60.33	62.26	59.81	60.05	56.65	**67.51**	8.43%
	NDCG@10	55.78	51.23	45.85	48.91	61.25	62.87	64.66	63.07	62.75	59.99	**69.99**	8.24%
	NDCG@20	60.12	55.67	51.05	52.36	63.33	66.67	67.89	65.26	65.84	63.34	**71.25**	4.95%
	Re@5	45.39	45.51	39.41	41.14	50.15	53.31	54.98	52.28	56.89	57.81	**60.05**	3.87%
	Re@10	47.81	48.96	42.57	45.36	52.36	56.87	57.08	55.64	58.99	60.36	**64.78**	7.32%
	Re@20	51.98	52.01	45.66	47.77	55.55	59.05	60.01	57.81	60.06	62.36	**66.37**	6.43%

- **RLMRec** [135]: uses LLMs for detailed profiling of users/items, aligning this with collaborative data to bolster representation learning.
- **KHGT** [207]: uses LLMs to analyze multiplex user-item interactions and knowledge-aware item relations for multi-behavior recommendation.

3.2.3.3 Result Analysis

Performance Comparison: To evaluate the recommendation performance of LDPE, we compare it with state-of-the-art baselines across three datasets using Pr@K, NDCG@K, and Re@K. As shown in Table 3.4, we found that LDPE consistently outperforms all baselines, with significant improvements on *MovieLens-1M* (e.g., 11.09% on Pr@10, 4.35% on NDCG@10, 5.59% on Re@10) and similar gains on *Amazon-Book* and *Amazon-Beauty*. These superior results can be attributed to LDPE's novel integration of LLMs and causal inference, which captures rich interaction semantics while mitigating both item-side and user-side exposure biases for robust sequential recommendations. Moreover, the consistent improvements across three different datasets show LDPE's superior adaptability in diverse domains.

Ablation Study: To analyze component contributions, we conduct ablation studies by removing or replacing each component. URL denotes replacing the unbiased representa-

tion learning component with traditional matrix factorization. *CF* and *LLM* denote only use CF-based embedding and LLM-based embedding, respectively. *IPS* and *UPS* denote removing the item-side and user-side propensity score estimation, respectively. *ALL* is the complete LDPE model. Figure 3.10 presents that unbiased representation learning is the most crucial one, with its removal causing significant drops in NDCG@20: 40.8% on *Amazon-Beauty* and 45.3% on *Amazon-Book*. Within this component, LLM-based embedding proves more impactful than CF-based embedding, highlighting the importance of semantic understanding. Additionally, removing propensity scores demonstrates their importance, with *IPS* removal showing a 14.56% drop compared to *UPS*'s 11.26% drop on *Amazon-Beauty*'s NDCG@20, confirming the necessity of addressing exposure bias from both perspectives.

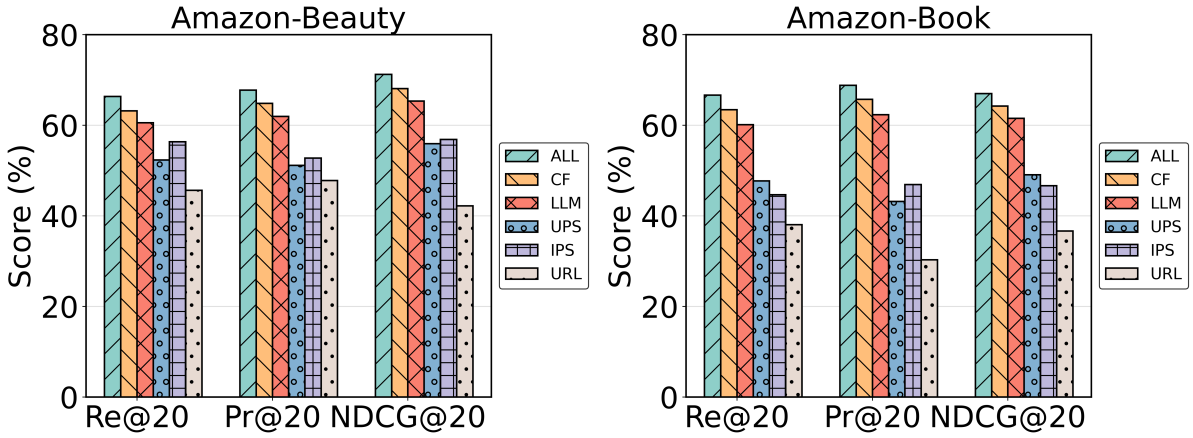


Figure 3.10: LDPE: Ablation study.

LLM Choice Comparison: To identify the most efficient LLM for LDPE, we evaluate four state-of-the-art LLMs on LDPE's performance, including Llama2 [166], Alpaca [162], ChatGPT [204], and T5 [31]. These LLMs have been pre-trained on extensive textual data, showing exceptional capabilities in capturing rich semantics and generating robust embeddings. Figure 3.11 show that T5 consistently outperforms other LLMs across all metrics, achieving the highest NDCG@20 of 70% on *MovieLens-1M* and 67% on *Amazon-Book*, surpassing the second-best performers around 4%. T5's superior performance can be attributed to its text-to-text framework, which enables the unified handling of complex text-understanding tasks. Based on these results, we select T5 as LDPE's LLM encoder for its strong capability in semantic information extraction.

Parameter Sensitivity Analysis: To evaluate LDPE's sensitivity to various parameter configurations, we conduct a sensitivity analysis on four key parameters: β, ϵ, γ, and α. α manages the trade-off between LLM-based and CF-based embeddings, while β, ϵ, and

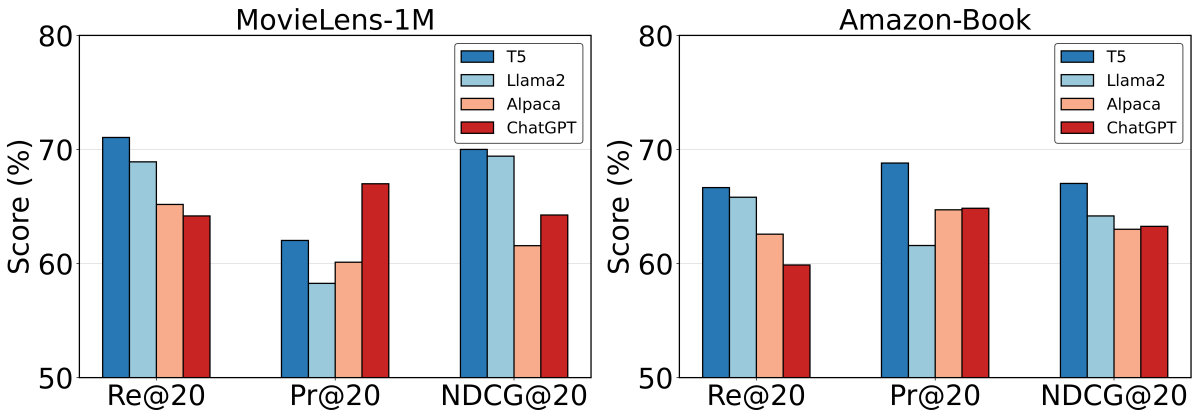


Figure 3.11: LDPE: Backbone LLM encoder selection comparison.

γ balance the item contrastive loss, user contrastive loss, and user side loss, respectively. Each parameter was tuned within the range of [0, 1] with a step size of 0.1. *MovieLens-1M* and *Amazon-Book* are selected to examine the parameter sensitivity as they differ in density and domain, which can ensure the scalability of LDPE. Figure 3.12 presents the LDPE's performance on these datasets, where optimal values are $\alpha = 0.4$, $\beta = 0.6$, $\epsilon = 0.4$, and $\gamma = 0.2$. It means a slight emphasis on CF-based embedding, and a higher emphasis on item contrastive loss compared to user contrastive loss and user side loss, enables LDPE's best performance. These findings show LDPE's robustness across different parameter settings and reveal the relative significance of each component in the final objective.

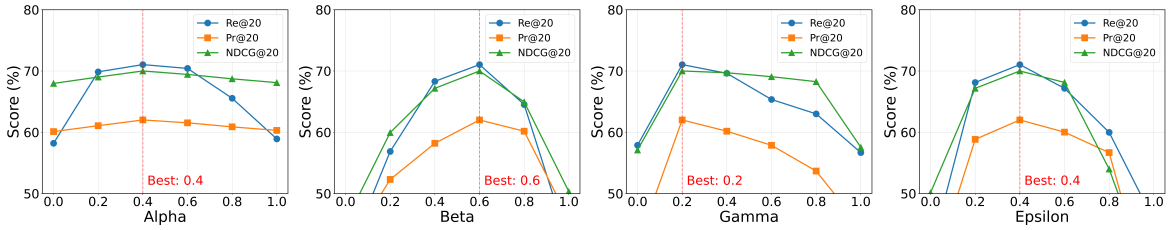


Figure 3.12: LDPE: Parameter analysis over *MovieLens-1M*.

Time Complexity Analysis: Table 3.5 presents the average inference response time per query for all baseline models across three benchmark datasets: MovieLens-1M, Amazon-Books, and Amazon-Beauty. Generally, G4Rec and NARM achieve the lowest response times due to their streamlined recurrent architectures, while methods such as MOJITO, DRec, and ITPS display moderate inference speeds. More complex models like ReLLa, C4Rec, RLMRec, and DRO incur slightly higher computational costs, reflecting their use of advanced semantic or robust optimization modules. KHGT exhibits the highest response times

on all datasets, attributable to its knowledge-aware reasoning. Our proposed LDPE model demonstrates competitive efficiency, with average response times of 84 ms, 91 ms, and 98 ms on MovieLens-1M, Amazon-Books, and Amazon-Beauty, respectively. The results indicate that, despite the increased complexity introduced by causal dual propensity estimation, LDPE maintains practical inference efficiency across diverse recommendation scenarios, striking a favorable balance between computational overhead and improved performance.

Table 3.5: Average inference response time (ms) of all baseline models on three benchmark datasets.

Model	**MovieLens-1M**	**Amazon-Books**	**Amazon-Beauty**
MOJITO	62	70	77
DRec	76	83	90
ITPS	79	86	92
G4Rec	55	62	68
NARM	59	65	71
ReLLa	88	94	102
C4Rec	92	99	109
RLMRec	89	97	105
DRO	83	89	96
KHGT	99	106	116
LDPE	84	91	98

3.2.3.4 Summary

In this work, we propose LDPE, a novel framework that integrates LLMs and causal inference to address RQ1 regarding bias mitigation in recommender systems. By aligning LLM-based semantic embeddings with collaborative information, LDPE generates unbiased user/item embeddings and estimates time-aware dual propensity scores to fully mitigate both item-side and user-side exposure biases. Extensive experiments show LDPE's superior performance over state-of-the-art baselines, while future work will focus on enhancing recommendation explainability through the LLMs and other causal techniques.

CHAPTER 4

CAUSAL MODELS FOR COMPLEX RECOMMENDATIONS

Building upon the foundational debiasing strategies introduced in Chapter 3, this chapter extends causal modeling to more complex recommendation scenarios where confounding biases manifest in intricate graph and sequential structures. While Chapter 3 focuses on isolating attribute-level (CCR) and exposure-level (LDPE) confounders in relatively structured settings, the models in this chapter—CGSR, GCRec, and CEDA—further generalize these ideas to address higher-order confounding effects in session graphs, multi-relational user-item networks, and social diffusion patterns, respectively. Specifically, CGSR inherits the idea of blocking spurious paths (like in CCR) but applies it to session graphs by disrupting shortcut propagation through causal edge dropout. GCRec builds on LDPE's use of causal graphs and propensity scoring but extends to multi-confounder settings using backdoor adjustment strategies. Finally, CEDA leverages causal intervention principles to restructure information diffusion paths in social networks, inspired by the exposure adjustment logic of LDPE but adapted for user-user influence graphs. Together, the three causal models in this chapter demonstrate how causal debiasing principles can be adapted and expanded to fit the complexity of real-world recommendation environments.

4.1 Causality-Guided Graph Learning for Session-based Recommendation

4.1.1 Overview

Session-based recommendation systems (SBRs) aim to capture evolving user preferences by taking into account the sequential order of interactions within sessions [70, 153]. Generally, traditional SBRs generate complex session graphs based on sequential interactions and then explore user behavior patterns from the graph for next-item recommendations. However, these methods mainly rely on attention or pooling mechanisms that are prone to exploiting shortcut paths in session graphs. Shortcut paths were originally a concept in computer vision that occurs when a user quickly navigates between items without meaningful intermediate interactions. This can lead to suboptimal recommendations by causing the model to miss important contextual information and causal relationships that can improve the accuracy of recommendations.

4.1.1.1 Research Objective

This study aims to address RQ2 regarding how causal inference can enhance recommender system robustness against spurious correlations in complex scenarios. Generally, we focus on developing a principled causal framework that can effectively identify and block shortcut paths in complex session graphs to distinguish between spurious corrections and true causal relationships in SRSs. Thereby, generating more robust and accurate recommendations by focusing on true causal patterns while providing interpretable explanations for the recommended items.

4.1.1.2 The Proposed Method

To achieve the aforementioned objectives, we propose a novel approach called **C**ausality-guided **G**raph Learning for **S**ession-based **R**ecommendation (**CGSR**) with two synergistic components:

- ***Distillation*** employs back-door adjustment of causality to block shortcut paths on the session graph, producing a distilled session graph that captures only genuine causal relations among items.

- ***Aggregation*** performs multi-layer aggregation over different edge types on the distilled session graph to estimate session preferences, incorporating rich semantics from various interaction patterns to learn users' true intentions.

The key contributions of this research are summarized as follows:

- We propose a novel causal-based method CGSR that effectively blocks shortcut paths and exploits the causal features for capturing the user's true intentions in the session-based recommendation.
- We design an innovative causality-guided graph learning framework that can generate interpretable graphs with pathways to highlight the items that have a significant causal impact on user preferences.
- We develop a principled approach to perform high-order aggregation over the distilled session graph, enabling more accurate modeling of complex user preferences while improving the explainability.

4.1.2 CGSR

4.1.2.1 Problem Definition

Let $\boldsymbol{S} = \boldsymbol{s}_1, \boldsymbol{s}_2, \cdots, \boldsymbol{s}_s$ be the set of anonymous sessions over the item universe $I = i_1, i_2, \cdots, i_{|I|}$. To capture transitions between items, we construct a session graph $\mathscr{G} = \mathscr{I}, \mathscr{E}$ using all available sessions in the training data, where $\mathscr{I}$ is the set of items and $\mathscr{E} = \varepsilon i j$ is the set of directed edges (see Figure 4.1). Specifically, an edge $\varepsilon i j$ is created if item i_j directly follows item i_i in any session. However, it is important to note that this transformation from session sequences to session graphs is generally lossy. That is, the session graph does not preserve the complete sequential order of items within each session; consequently, it is often impossible to reconstruct the original session sequences from the graph alone. Multiple distinct sequences can be mapped to the same graph structure, resulting in information loss regarding the order of user interactions. To address this limitation, recent work [27] has proposed lossless transformation strategies that augment the session graph with additional information, such as edge attributes or order encodings. These enhancements allow the original session sequence to be uniquely reconstructed from its graph representation, thereby minimizing information loss during the transformation process. In this work, we focus on accurately predicting users' next interactions and mitigating shortcut paths in the session graph, which can distort the estimation of users' true preferences.

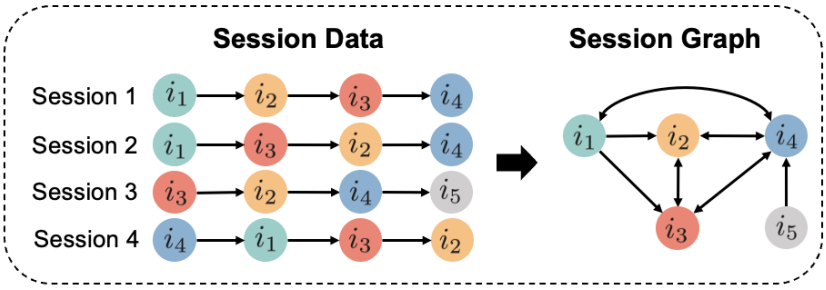


Figure 4.1: An example of a session graph derived from the session data.

To achieve accurate predictions, it's crucial to analyze the session graph from a causal perspective and construct a structural causal model to mitigate the influence of shortcut paths [188, 187, 186]. As shown in Figure 4.2, we design a causal graph showing relationships among five key variables: graph representation A, shortcut feature B, causal feature C, graph data D, and prediction Y. A back-door path lies between C and Y (i.e., $C \leftarrow D \rightarrow B \rightarrow A \rightarrow Y$), where B acts as a confounder creating spurious correlations. To block this back-door path, we employ the back-door adjustment of causal inference:

$$P(Y \mid do(C)) = P(Y \mid C) = \sum_{B \in \mathcal{T}} P(Y \mid C, B) P(B) \tag{4.1}$$

where $\mathcal{T}$ denotes the confounder set, $P(Y \mid C, B)$ represents the conditional probability given C and B, and $P(B)$ is the confounder's prior probability. Through this back-door adjustment, we can effectively block shortcut paths and preserve true causal relationships in the session graph, enabling more accurate recommendations while providing fine-grained explanations for interactions.

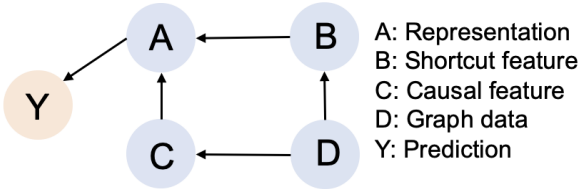


Figure 4.2: Our designed causal graph for SRSs.

4.1.2.2 Methodology

Figure 4.3 shows the overall framework of our proposed CGSR, which operates on a pre-trained initial session graph to address shortcut paths through ***Distillation*** and ***Aggregation*** component. To represent the structure of the session graph, we begin by constructing

node feature matrix $\boldsymbol{X} \in \mathbb{R}^{|\mathscr{I}| \times \mathrm{d}}$ of $\mathscr{G}$, where each row corresponds to a node in the graph and each column corresponds to a feature. Then, we build the adjacency matrix $\mathbf{A} \in \mathbb{R}^{|\mathscr{I}| \times |\mathscr{I}|}$, where $\mathbf{A}[i, j] = 1$ if edge ε_{ij} exists, otherwise is 0. To capture temporal dependencies in the session graph, we employ GRUs to learn item representations by processing sequential input $\{\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, ..., \boldsymbol{x}_{it}\}$, where each is the attribute vector of the item i_i at time t. The initial hidden state of the GRUs at time step $t = 0$ is defined as $\boldsymbol{h}_{i0} = \mathbf{0} \in \mathbb{R}^{\mathrm{d}}$, which is an all-zero vector in d dimension. Then, the GRU updates hidden states as follows:

$$
\begin{aligned}
\boldsymbol{z}_t &= \sigma\left(\mathbf{W}_z \boldsymbol{x}_{it} + \mathbf{U}_z \boldsymbol{h}_{i(t-1)} + \boldsymbol{b}_z\right) \\
\boldsymbol{r}_t &= \sigma\left(\mathbf{W}_r \boldsymbol{x}_{it} + \mathbf{U}_r \boldsymbol{h}_{i(t-1)} + \boldsymbol{b}_r\right) \\
\boldsymbol{h}_{it} &= (1 - \boldsymbol{z}_t) \odot \boldsymbol{h}_{i(t-1)} + \boldsymbol{z}_t \odot \tilde{\boldsymbol{h}}_{it} \\
\tilde{\boldsymbol{h}}_{it} &= \tanh\left(\mathbf{W} \boldsymbol{x}_{it} + \mathbf{U}\left(\boldsymbol{r}_t \odot \boldsymbol{h}_{i(t-1)}\right) + \boldsymbol{b}\right)
\end{aligned}
\tag{4.2}
$$

where $\boldsymbol{h}_{it} \in \mathbb{R}^{\mathrm{d}}$ is the hidden state at time step t and represents the item i_i at that time step. σ is the sigmoid function. $\odot$ is element-wise multiplication. $\mathbf{W}$ and $\mathbf{U}$ are weight matrices for transferring the previous hidden state into the candidate hidden state $\tilde{\boldsymbol{h}}_t$. The weight matrices $\mathbf{W}_r$, $\mathbf{U}_r$, $\mathbf{W}_z$, and $\mathbf{U}_z$ are additional learnable parameters.

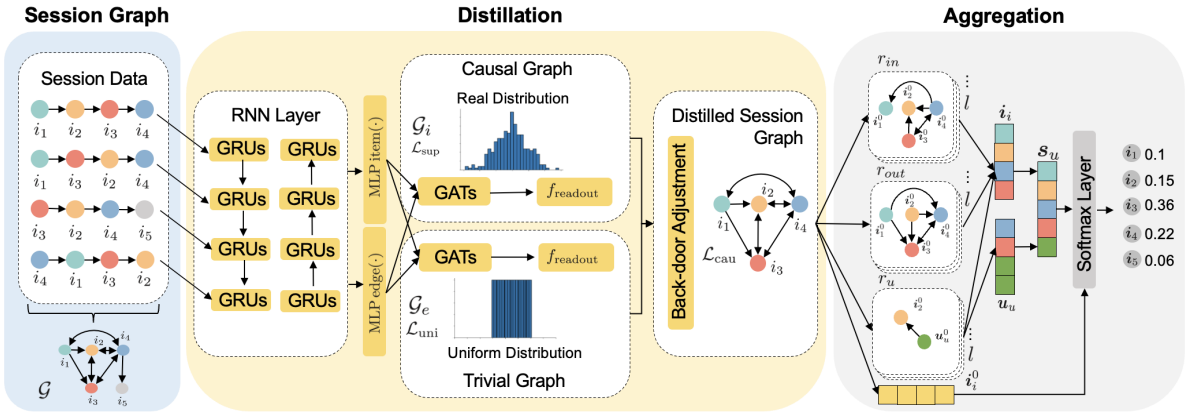


Figure 4.3: The overall framework of our proposed method CGSR.

Using the obtained item representations $\boldsymbol{h}_{it}$, we decompose the input session graph into causal and trivial proposals, i.e., causal graph and trivial graph [155]. Specifically, we employ Multi-Layer Perceptrons (MLPs) to compute attention scores from both item and edge perspectives:

$$
\begin{aligned}
&\alpha_{c_i} = softmax\left(\mathrm{MLP}\left(\boldsymbol{h}_{it}\right)\right), \alpha_{t_i} = softmax\left(\mathrm{MLP}\left(\boldsymbol{h}_{it}\right)\right) \\
&\beta_{c_{ij}} = softmax\left(\mathrm{MLP}\left(\boldsymbol{h}_{it} \| \boldsymbol{h}_{jt}\right)\right), \beta_{t_{ij}} = softmax\left(\mathrm{MLP}\left(\boldsymbol{h}_{it} \| \boldsymbol{h}_{jt}\right)\right)
\end{aligned}
\tag{4.3}
$$

where $\boldsymbol{h}_{jt}$ is the hidden state of item i_j at time t, which is the same as $\boldsymbol{h}_{it}$ for item i_i. $\|$ denotes the concatenation and MLP denotes the MLPs. $\alpha_{c_i}/\alpha_{t_i}$ and $\beta_{c_{ij}}/\beta_{t_{ij}}$ are normalized attention scores for items and edges. These scores are used to construct soft masks [238] $\mathbf{M}_a \in \mathbb{R}^{|\mathscr{I}|\times|\mathscr{I}|}$ and $\mathbf{M}_x \in \mathbb{R}^{|\mathscr{I}|\times 1}$. Each element ranged between 0 and 1 denotes the attention score relevant to the task of interest, defining the importance of each element to the prediction. Next, we define their complementary masks as $\overline{\mathbf{M}}_a = \mathbf{1} - \mathbf{M}_a$ and $\overline{\mathbf{M}}_x = \mathbf{1} - \mathbf{M}_x$, where $\mathbf{1}$ is the all-one matrix. Lastly, we divide the $\mathscr{G}$ into the causal graph (e.g., captures the causal relations for users' preferences) and trivial graph (e.g., includes the shortcut patterns) as $\mathscr{G}_i = \{\mathbf{A} \odot \mathbf{M}_a, \boldsymbol{X} \odot \mathbf{M}_x\}$ and $\mathscr{G}_e = \left\{\mathbf{A} \odot \overline{\mathbf{M}}_a, \boldsymbol{X} \odot \overline{\mathbf{M}}_x\right\}$, respectively.

Having the causal and trivial graphs, we then adopt two Graph Attention Networks [153] to learn their representations and make predictions:

$$\begin{aligned} \mathbf{E}_{\mathscr{G}_i} &= f_{\text{readout}}\left(\text{Att}\left(\mathbf{W}_i\mathbf{X}_i, \mathbf{A}_i\right)\right) \\ \mathbf{E}_{\mathscr{G}_e} &= f_{\text{readout}}\left(\text{Att}\left(\mathbf{W}_e\mathbf{X}_e, \mathbf{A}_e\right)\right) \end{aligned} \tag{4.4}$$

where Att$(\cdot)$ performs multi-head attention over transformed node features $\mathbf{W}_i\mathbf{X}_i/\mathbf{W}_e\mathbf{X}_e$ and adjacency matrices $\mathbf{A}_i/\mathbf{A}_e$ f_{readout} is readout function [155] aggregates node-level features into graph-level representations. $\mathbf{W}_i\mathbf{X}_i$ and $\mathbf{W}_e\mathbf{X}_e$ are linear transformations of the input feature matrices for $\mathscr{G}_i$ and $\mathscr{G}_e$, which are computed as the multiplication between the $\mathbf{W}_i$ and $\mathbf{X}_i$. The same computation process is applied to $\mathbf{W}_e\mathbf{X}_e$. Following that, the cross-entropy loss functions for $\mathscr{G}_i$ and $\mathscr{G}_e$ are defined as

$$\begin{aligned} \mathscr{L}_{\text{sup}} &= -\frac{1}{|\mathscr{D}|}\sum_{\mathscr{G}\in\mathscr{D}} \mathbf{y}_{\mathscr{G}}^{\top}\log\left(\mathbf{E}_{\mathscr{G}_i}\right) \\ \mathscr{L}_{\text{uni}} &= \frac{1}{|\mathscr{D}|}\sum_{\mathscr{G}\in\mathscr{D}} \text{KL}\left(\mathbf{y}_{\text{uni}}, \mathbf{E}_{\mathscr{G}_e}\right) \end{aligned} \tag{4.5}$$

where $\mathscr{L}_{\text{sup}}$ minimizes cross-entropy between predicted distribution $\mathbf{E}_{\mathscr{G}_i}$ and ground-truth $\mathbf{y}_{\mathscr{G}}$ over training data $\mathscr{D}$. $\mathscr{L}_{\text{uni}}$ uses KL-divergence to push predictions in trivial graph $\mathbf{E}_{\mathscr{G}_e}$ toward uniform distribution $\mathbf{y}_{\text{uni}}$, helping capture shortcut patterns [155].

Now, we begin to implement the back-door adjustment over the generated two graphs, which can eliminate the effects of shortcut features on the shortcut path in the session graph. Specifically, we pair the target causal graph with each stratification of the trivial graph to create the intervened graphs, which means the back-door adjustment is implemented on the representation level. Mathematically, we have

$$\begin{aligned} \mathbf{z}_{\mathscr{G}'} &= \Phi\left(\mathbf{E}_{\mathscr{G}_i} \odot \mathbf{E}_{\mathscr{G}_e}\right) \\ \mathscr{L}_{\text{cau}} &= -\frac{1}{|\mathscr{D}|\cdot|\hat{\mathscr{T}}|}\sum_{\mathscr{G}\in\mathscr{D}}\sum_{t'\in\hat{\mathscr{T}}} \mathbf{y}_{\mathscr{G}}^{\top}\log\left(\mathbf{z}_{\mathscr{G}'}\right) \end{aligned} \tag{4.6}$$

where $\mathbf{z}_{\mathcal{G}'}$ is the representation of the intervened graph obtained through element-wise multiplication $\odot$ and mapping function Φ. $\hat{\mathcal{T}}$ is the estimated stratification set of the trivial graph capturing the shortcut features from training data. The resulting $\mathcal{G}'$ can be inferred as the distilled session graph. Finally, the overall distillation objective $\mathcal{L}_{dis}$ is defined as:

$$\mathcal{L}_{dis} = \mathcal{L}_{\text{sup}} + \mathcal{L}_{\text{uni}} + \mathcal{L}_{\text{cau}} \tag{4.7}$$

By optimizing the $\mathcal{L}_{dis}$, we generate a distilled session graph that preserves only true causal relationships by blocking shortcut paths.

Next, the second component ***Aggregation*** aims to predict user preferences by performing high-order information propagation over the distilled session graph $\mathcal{G}'$. Specifically, for each item i_i and user u_u, we define their neighbor sets $\mathcal{S}_i$ and $\mathcal{S}_u$ based on edge types r_{out} and r_u respectively. This multi-edge type enables capturing complex user-item relationships and rich semantic patterns that can enhance recommendations. Given that sequential item interactions often exhibit attribute similarities reflecting user preferences [95], we compute normalized similarity scores between adjacent items i_i and i_j as

$$w_{i,j} = \boldsymbol{x}_i^\top \cdot \boldsymbol{x}_j, \quad \hat{w}_{i,j} = \frac{w_{i,j}}{\sum_{k \in \mathcal{S}_i} w_{i,k}} \tag{4.8}$$

where the $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are attribute vectors for these two items. $\cdot$ is the dot product. $w_{i,j}$ represents the similarity score among items i_i and i_j. $\hat{w}_{i,j}$ represents the normalized similarity score for edge type r_{out}. Similar normalization is applied for edge types r_{in} and r_u using their respective neighbor sets, forming the foundation for the adjacency matrix of the distilled session graph $\mathcal{G}'$. Following that, we generate high-quality item representations for session preference based on the estimated normalized similarity score. Specifically, we employ GNNs to aggregate messages from different edge types. Taking the outgoing edges r_{out} as an example, we accumulate the messages as

$$\begin{aligned} \boldsymbol{i}_{\mathcal{S}_i}^{(l)} &= \frac{1}{|\mathcal{S}_i|} \sum_{k \in \mathcal{S}_i} \boldsymbol{i}_i^{(l-1)} \\ \boldsymbol{i}_{i,r_{out}}^{(l)} &= RELU\left(\hat{w}_{i,j}^{(l)} \left[\boldsymbol{i}_{\mathcal{S}_i}^{(l)} \| \boldsymbol{i}_i^{(l-1)}\right]\right) \end{aligned} \tag{4.9}$$

where $\boldsymbol{i}_i^{(l-1)}$ is the item representation at layer $l-1$, initialized with $\boldsymbol{h}_{2t}$ from the GRU to capture temporal dependencies. $RELU$ denotes the relu activation function. The normalized similarity score $\hat{w}_{i,j}^{(l)}$ weights the importance of neighbor information when generating the aggregated representation $\boldsymbol{i}_{i,r_{out}}^{(l)}$ with edge type r_{out}.

Similarly, we perform aggregation operations for edge types r_u and r_{in} to capture diverse relationship patterns. For user-connected edges $\boldsymbol{i}_{i,r_u}^{(l)}$, we compute $\boldsymbol{i}_{i,ru}^{(l)}$ by aggregating information from user neighbors. For incoming edges, we compute $\boldsymbol{i}_{i,rin}^{(l)}$ from input

item neighbors. Both processes use normalized weights $\hat{w}_{u,i}^{(l)}$ and $\hat{w}_{i,j}^{(l)}$ respectively. The final layer-wise representation combines information across all edge types:

$$\boldsymbol{i}_i^{(l)} = \text{mean}\left(\boldsymbol{i}_{i,r_u}^{(l)}, \boldsymbol{i}_{i,r_{in}}^{(l)}, \boldsymbol{i}_{i,r_{out}}^{(l)}\right) \tag{4.10}$$

where mean$(\cdot)$ represents a mean operation that takes the average of all messages. This multi-type aggregation generates comprehensive item representations $\boldsymbol{i}_i^{(l)}$ in l layer that captures rich semantic patterns from high-order neighborhood interactions.

Likewise, once anonymous users interact with items in sessions, we generate high-quality user representations through their high-order connections as

$$\begin{aligned} \boldsymbol{u}_{\mathscr{S}_u}^{(l)} &= \frac{1}{|\mathscr{S}_u|} \sum_{i \in \mathscr{S}_u} \boldsymbol{i}_i^{(l-1)} \\ \boldsymbol{u}_{u,r_u}^{(l)} &= RELU\left(\hat{w}_{u,i}^{(l)} \left[\boldsymbol{u}_{\mathscr{S}_u}^{(l)} \| \boldsymbol{i}_i^{(l-1)}\right]\right) \end{aligned} \tag{4.11}$$

where $\hat{w}_{u,i}^{(l)}$ is the normalized similarity score between user u_u and item i_i in the l layer. The resulting $\boldsymbol{u}_{u,r_u}^{(l)}$ denotes the aggregated user representation in the l layer with edge type r_u. Since users are only connected with the interacted items in sessions via the edge type r_u, thus we define $\boldsymbol{u}_u^{(l)}$ as the aggregated user representation as

$$\boldsymbol{u}_u^{(l)} = \text{mean}\left(\boldsymbol{u}_{u,r_u}^{(l)}\right) \tag{4.12}$$

where $\boldsymbol{u}_u^{(l)}$ aggregates information from a user's interacted items through weighted averaging, capturing user preferences through the lens of item-user relationships. This aggregation allows us to capture the user's preferences based on specific edge types, representing the fine-grained level relationship between the user and items.

Next, we aggregate the item and user representations at each layer to construct the final global-level item and user representation as

$$\boldsymbol{u}_u = \sum_{l=0}^{L} \alpha_{(l)} \boldsymbol{u}_u^{(l)}, \quad \boldsymbol{i}_i = \sum_{l=0}^{L} \alpha_{(l)} \boldsymbol{i}_i^{(l)} \tag{4.13}$$

where $\alpha(l)$ are learnable importance weights for each layer through backpropagation. $\boldsymbol{u}_u$ and $\boldsymbol{i}_i$ are the final global-level user and item representation, respectively. To account for the asymmetric contributions of users and items to session preferences, we employ an adaptive gating mechanism g:

$$\begin{aligned} g &= \sigma\left(\theta\left(\boldsymbol{i}_i \| \boldsymbol{u}_u\right)\right) \\ \boldsymbol{s}_u &= g \cdot \boldsymbol{i}_i + \left(1 - g\right) \cdot \boldsymbol{u}_u \end{aligned} \tag{4.14}$$

where g is a scalar value between 0 and 1, computed using a trainable parameter θ and concatenated representations of $\boldsymbol{i}_i$ and $\boldsymbol{u}_u$. $\boldsymbol{s}_u$ is the final session preference embedding generated by dynamically weighting the user and item representations through gate g, capturing high-order interaction patterns from the distilled session graph. If g is close to 1, more weight is given to the $\boldsymbol{i}_i$, otherwise, more weight is given to the $\boldsymbol{u}_u$. So far, we have the session preference representation $\boldsymbol{s}_u$ of the user containing rich semantics from high-order neighbors over the distilled session graph.

Finally, our last component ***Recommendation*** leverages the generated session preference $\boldsymbol{s}_u$ to make Top-K recommendations. Specifically, we compute the recommendation probability $\hat{\mathbf{y}}$ of candidate items and its objective can be formulated as

$$\begin{aligned} \hat{y}_i &= softmax\left(\boldsymbol{s}_u{}^\top \cdot \boldsymbol{i}_i^0\right) \\ \mathscr{L} &= -\sum_{i=1}^{|\mathscr{I}|} y_i \log\left(\hat{y}_i\right) + \left(1-y_i\right)\log\left(1-\hat{y}_i\right) + \lambda \mathscr{L}_{dis} \end{aligned} \tag{4.15}$$

where y_i is the binary interaction label and $\boldsymbol{i}_i^0$ is the item's initial representation capturing its intrinsic features independent of session context. The model computes similarity scores between $\boldsymbol{s}_u$ and $\boldsymbol{i}_i^0$, normalizes them through $softmax$, and optimizes the objective $\mathscr{L}$ with distillation regularization controlled by λ. By optimizing the $\mathscr{L}$, we can block the effect of shortcut paths in session graphs for more robust SRSs.

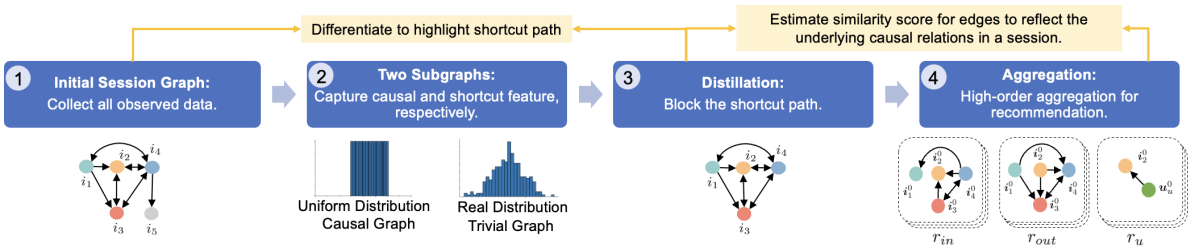


Figure 4.4: CGSR: The workflow of our proposed CGSR.

4.1.3 Experiments

4.1.3.1 Datasets

We evaluate our proposed CGSR on two widely used real-world datasets: *Yoochoose*[1] and *Diginetica*[2], as shown in Table 4.1. Following prior work [96, 110, 203], since *Yoochoose* is too

[1] http://2015.recsyschallenge.com/challege.html
[2] http://cikm2016.cs.iupui.edu/cikm-cup

large for training, we use 1/64 and 1/4 subsets of *Yoochoose*, denoted as *Yoochoose 1/64* and *Yoochoose 1/4* containing click events in session form. *Diginetica* is a dataset from CIKM Cup 2016 that comprises sessions extracted from E-commerce search logs. All datasets are split into training, testing, and validation sets with a ratio of 70%/20%/10%, respectively.

Table 4.1: CGSR: Statistical details of the three datasets.

Statistics	**Yoochoose 1/64**	**Yoochoose 1/4**	**Diginetica**
# clicks	557,248	8,326,407	982,961
# training sessions	369,859	5,917,746	719,470
# test sessions	55,898	55,898	60,858
# items	16,766	29,618	43,097
Average session length	6.16	5.71	5.12

4.1.3.2 Baselines and Evaluation

Following prior work [127, 27, 57], we use three widely used metrics for evaluating the Top-K recommendation: Precision (Pr@K) measuring the proportion of relevant items, Hit Ratio (HR@K) capturing the proportion of interacted items, and Mean Reciprocal Rank (MRR@K) indicating the ranking position of the first relevant item in Top-K recommendations. For implementation, we set embedding dimension d=100 and initialize all parameters using a Gaussian distribution with a mean of 0 and a standard deviation of 0.1. To optimize parameters, we use the Adam optimizer with learning rate 0.001 (decayed by 0.1 every 5 epochs) and batch size 100. To ensure data quality, sessions with fewer than 5 interactions are filtered out. We compare CGSR with the below state-of-the-art SBRs to verify its effectiveness:

- **Item-KNN** [144] recommends items similar to previously interacted items based on the cosine similarity among their embeddings.
- **STAMP** [110] uses a memory attention mechanism to learn general preferences and the last item as the recent preferences to generate recommendations.
- **SR-GNN** [203] employs GNNs to model complex dependencies among items over the session graph to generate recommendations.
- **GCE-GNN** [194] employs GNNs to learn item transitions in the session graph for more robust SBRs.
- **A-PGNN** [241] employs GCNs to learn item transitions in the session graph for personalized SBRs.

Table 4.2: CGSR: Recommendation performance comparison: the best results are in bold, while the best baselines are underlined.

Dataset	Yoochoose 1/64						Yoochoose 1/4						Diginetica					
ModelMetrics	Pr@5	Pr@10	HR@5	HR@10	MRR@5	MRR@10	Pr@5	Pr@10	HR@5	HR@10	MRR@5	MRR@10	Pr@5	Pr@10	HR@5	HR@10	MRR@5	MRR@10
ItemKNN	39.84	42.67	30.15	33.28	15.51	17.36	41.83	44.51	34.52	36.66	7.71	10.25	28.83	30.25	20.01	21.55	8.12	10.99
STAMP	60.14	64.45	41.16	44.82	27.72	28.88	65.18	68.04	44.71	48.01	26.36	29.45	40.1	44.01	31.25	34.51	11.11	14.32
GRU4Rec	48.17	51.98	35.61	37.77	18.77	20.14	47.78	50.17	36.61	38.96	22.22	25.51	23.34	27.77	14.57	16.76	6.86	10.57
NARM	58.11	62.33	42.11	45.51	21.14	25.41	59.81	64.58	40.14	42.25	24.48	27.88	38.67	45.36	29.91	31.26	12.21	15.22
SG-GNN	61.28	65.78	43.68	47.7	26.63	29.43	61.15	67.81	42.66	45.11	29.91	36.55	42.61	47.55	33.75	35.91	12.63	16.51
GCE-GNN	67.77	70.5	45.55	47.13	26.89	34.14	67.51	74.41	46.71	48.81	31.11	34.99	47.88	53.32	41.56	42.52	20.25	28.62
H-RNN	65.65	68.71	42.39	46.14	25.55	30.11	62.52	75.14	45.52	47.77	30.04	31.99	48.38	52.15	24.44	26.55	17.67	24.44
A-PGNN	67.98	70.14	46.61	47.89	27.71	33.35	71.27	75.12	50.12	52.69	33.56	35.76	52.15	56.66	41.11	45.18	21.76	28.71
Ours	**74.25**	**74.56**	**50.12**	**52.22**	**30.16**	**35.77**	**77.47**	**80.1**	**54.12**	**56.78**	**38.11**	**40.14**	**55.55**	**59.78**	**44.63**	**49.81**	**25.11**	**35.15**
Improv.	9.22	5.76	7.53	9.04	8.80	4.77	8.70	6.60	7.98	7.76	13.56	9.82	6.52	5.51	7.39	10.25	15.40	22.43

- **H-RNN** [133] uses RNNs to model cross-session user preferences for more robust session-based recommendations.

- **GRU4Rec** [70] uses RNNs to model the sequential information and makes recommendations with session-parallel mini-batch training processes.

- **NARM** [96] uses RNNs to model the sequential information and adopts item-level attention to learn users' primary preferences for recommendations.

4.1.3.3 Result Analysis

Performance Comparison: As shown in Table 4.2, CGSR demonstrates superior performance across all datasets and metrics compared to state-of-the-art baselines. Notably, CGSR achieves substantial improvements in MRR@5, surpassing the best baselines by 8.80%, 13.56%, and 15.40% on *Yoochoose 1/64*, *Yoochoose 1/4*, and *Diginetica*, respectively. These outcomes suggest that our CGSR can effectively mitigate the effects of shortcut paths in the session graph, leading to more accurate predictions of user preferences. The performance gain is particularly pronounced on *Yoochoose 1/4*, which can be attributed to its higher data density enabling better learning of causal patterns. In summary, these results show that CGSR can effectively block shortcut paths, resulting in significant performance improvements over baselines, especially on denser datasets.

Ablation Study: As shown in Figure 4.5, we conducted an ablation study to examine the contribution of model components. Dis(w/o) denotes the removal of the distillation, directly using the initial session graph. Agg(w/o) denotes the removal of the aggregation, directly using the first layer's user/item representations to make session representations. We find that removing the distillation function leads to a dramatic 50% drop in HR@5 across both *Yoochoose* datasets, while removing the aggregation function results in a 32.45% decrease in Pr@5 on *Yoochoose 1/4*. These substantial performance degradations demonstrate that both components are essential for the model to effectively capture and utilize session

patterns. On the other hand, we analyze edge type importance by sequentially blocking input edges, output edges, and user neighbor connections. In(w/o) denotes the removal of the effects from the input edge associated with item neighbors. Out(w/o) denotes the removal of the effects from the output edge associated with item neighbors. U(w/o) denotes the removal of the effects from its associated user neighbors. We find that input edges have the greatest impact as they provide critical information from previous interactions. Output edges provide secondary information by indicating subsequent items, while user connections have the least impact because they do not directly model sequential patterns. These results validate the importance of considering edge types in session graph learning.

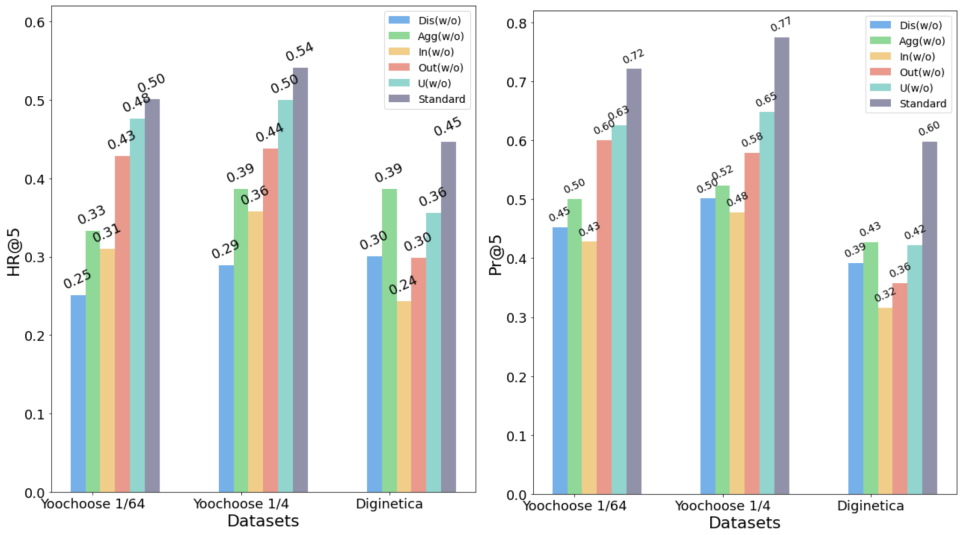


Figure 4.5: CGSR: Ablation study.

Interpretability: To verify the interpretability of our proposed CGSR, we conduct a case study on the *Diginetica*, as shown in Figure 4.6. The solid and dotted lines represent edges from the causal and trivial graphs respectively, derived from soft masks applied to the initial session graph. Edge weights indicate normalized similarity scores between sequential items, with higher scores suggesting a stronger likelihood of subsequent interactions. This representation enables clear visualization of how CGSR distinguishes between causal and spurious (e.g., shortcut) item relationships. In our case study, user 25 first interacts with item 1189, which has connections through three edge types: r_{in} (orange), r_{out} (black), and r_r (blue). While the edge between items 65407 and 57539 shows high similarity leading to a recommendation of item 7492, our causal analysis reveals item 49272 has stronger overall session-wide relationships. By identifying and blocking the 65407-57539 connection as a shortcut path, CGSR recommends item 49272 based on its robust causal relation-

ships across the session graph. Then, we can explain item 49272 as the most similar item to item 1189 due to strong causal relations of item 49272 within the entire session graph. This demonstrates that our CGSR improves model interpretability by considering causality while blocking shortcut paths, leading to more accurate and transparent SBRs.

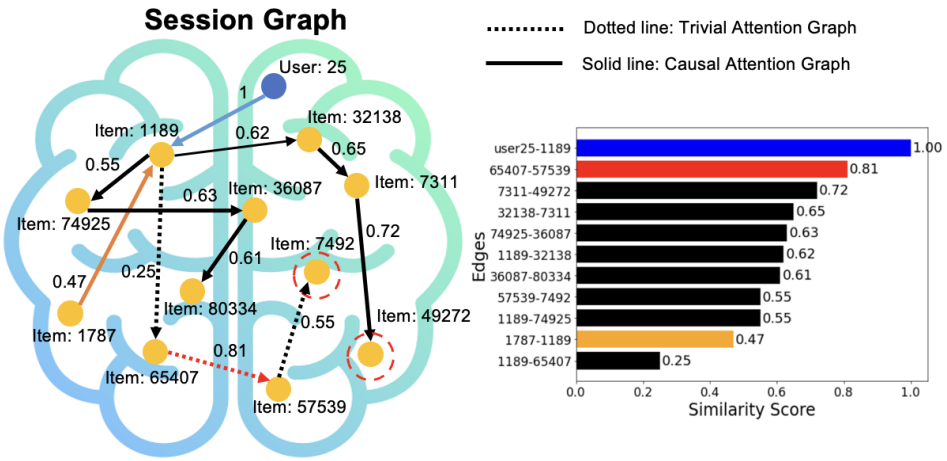


Figure 4.6: CGSR: Interpretability analysis on *Diginetica*.

Parameter Analysis: Figure 4.7 shows experimental results that investigate the influence of two important parameters: the number of aggregation layers (l) and the selection of Top-K items. The parameter l determines the level of aggregation, where a larger value considers more connections. We observe that the improvement rate diminishes after the 3rd layer, as lower layers capture simple features while higher layers model complex ones, potentially leading to overfitting. Moreover, increasing Top-K leads to higher accuracy as it can increase the candidate pool, increasing the likelihood of finding a better match. Based on these findings, we set $l = 3$ as optimal while Top-K can be adjusted according to specific requirements, ensuring robust recommendations while preventing overfitting.

4.1.3.4 Summary

In this work, we propose CGSR, a novel causal framework that addresses RQ2 regarding enhancing recommender system robustness against spurious correlations in complex recommendation scenarios. By synergistically integrating causal intervention and high-order aggregation techniques, CGSR effectively blocks shortcut paths in session graphs while capturing true causal relationships through distillation and preference modeling. Extensive experiments demonstrate CGSR's superior performance over state-of-the-art baselines in terms of recommendation accuracy and interpretability. Future work will explore inte-

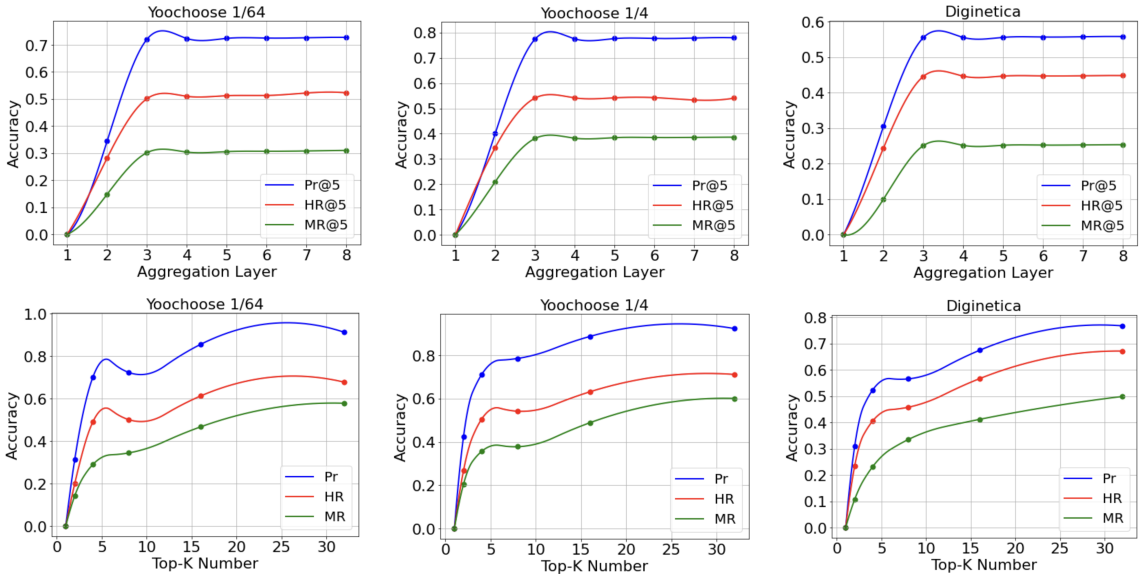


Figure 4.7: CGSR: Parameter analysis on three datasets.

grating causal inference with transfer learning to further enhance model robustness across diverse application domains.

4.2 Deconfounded Recommendation via Causal Intervention

4.2.1 Overview

Most traditional recommender systems aim to predict users' preferences by analyzing historical interaction data between users and items. However, real-world interaction data exists in complex structures that simultaneously contain multiple confounding biases caused by spurious correlations [20], making it challenging to capture users' true preferences. As shown in Figure 4.8, in social networks, users' purchase behaviors are often influenced by their friends' activities rather than true interests; while items in popular item groups always receive higher recommendation probabilities regardless of their actual relevance. Though existing approaches have attempted to address confounding bias, they usually only handle individual types of confounders, making them inadequate for multiple confounders in complex recommendation scenarios (RQ2). The inability to handle concurrent confounding effects severely impacts recommendation robustness, necessitating new approaches that can effectively model and mitigate multiple confounders simultaneously.

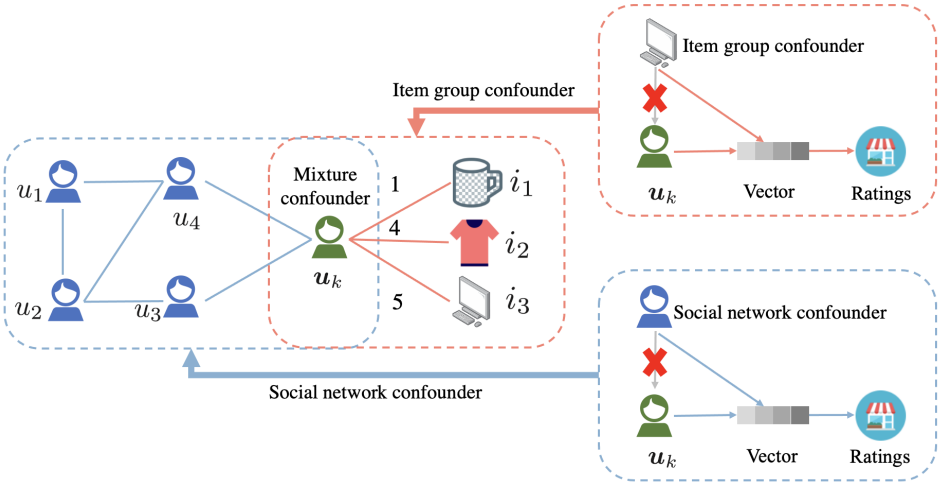


Figure 4.8: A causal view of two confounders concurrently arising from the item group and social network in the real-world.

4.2.1.1 Research Objective

This study aims to address RQ2 by developing a novel causal framework, that can enhance recommendation robustness when handling multiple concurrent confounders in complex scenarios. Specifically, we propose to integrate Graph Neural Networks (GNNs) with causal intervention techniques through back-door adjustment, enabling us to address confounding effects from both social networks and item groups in complex structures. This integration allows us to model the true causal relationships driving user preferences while eliminating spurious correlations in complex structures. Thereby, improving recommendation robustness against multiple confounding effects in real-world scenarios.

4.2.1.2 The Proposed Method

To achieve the aforementioned objectives, we propose a novel causal framework called **G**raph **C**ausal for **Rec**ommendation (**GCRec**) with three synergistic components:

- ***Embedding Initialization*** constructs initial user/item embeddings by incorporating rich attribute information, providing a strong foundation for modeling complex relational patterns.

- ***High-order Aggregation*** employs GNNs to refine the initial embeddings into comprehensive user/item representations by aggregating information across social networks and interaction histories, capturing sophisticated relational patterns.

- ***Confounder Debiasing*** implements causal intervention on the aggregated high-order GNN-based user/item representations to address multiple confounding effects simultaneously, with an innovative approach that adjusts intervention strength based on preference distribution divergence.

- ***Top-K Recommendation*** uses the generated high-order GNN-based representation and adjustable back-door adjustment to compute accurate prediction scores for each user-item pair, producing robust recommendations that reflect user true preferences.

The key contributions of this research are summarized as follows:

- We pioneer a causal framework that explicitly models dual confounding effects from social networks and item groups, enabling a deeper understanding of bias generation mechanisms to improve recommendations.

- We leverage Graph Neural Networks (GNNs) to learn high-order user representations that serve as deconfounders, addressing mixed confounders to more accurately capture user preferences for recommendations

- We introduce an adaptive causal intervention strategy that dynamically balances positive and negative confounding impacts through principled back-door adjustment.

4.2.2 GCRec

4.2.2.1 Problem Definition

Let $\mathscr{U}$, $\mathscr{I}$ and $\mathscr{G}$ be the sets of users, items and item groups respectively. For each user $u \in \mathscr{U}$, we observe their historical item interactions $\mathbf{O}_{u,i}$ with items $i \in \mathscr{I}$ and social connections $\mathbf{S}_{u,v}$ with other users $v \in \mathscr{U}$. Each item i belongs to one or more groups $g \in \mathscr{G}$ based on its attributes.

Our goal is to accurately predict users' true preferences while addressing two key confounders caused by spurious corrections in interaction data: (1) social network confounder, where users' behaviors are influenced by their friends rather than true interests; and (2) item group confounder, where items from popular groups receive higher recommendation probabilities regardless of relevance. As shown in Figure 4.9, we design a causal graph revealing the intricate relationships between user U, item I, confounder C, inherent preference M, and prediction Y. In conventional recommender systems, the prediction relies

on estimating conditional probability $P(Y|U,I)$, which can be affected by spurious correlations due to confounders:

$$\begin{aligned} P(Y|U=u,I=i) &= \sum_c \sum_m P(c|u)P(M(c,u)|c,u)P(Y|u,i,M(c,u)) \\ &= \sum_c P(c|u)P(Y|u,i,M(c,u)) \end{aligned} \tag{4.16}$$

where c represents the user's historical distribution over item groups and social connections, and $M(c,u)$ denotes the inherent preference representation under distribution c. This formulation shows how confounders c can impact both user representation u and final prediction Y through $M(c,u)$, creating spurious correlations. To address this, we employ back-door adjustment to block the confounding effects through causal paths $C \rightarrow U$ and $(C,U) \rightarrow M$:

$$\begin{aligned} P(Y|do(U=u),I=i) &= \sum_c P(c|do(U=u))P(Y|do(U=u),i,M(c,do(U=u))) \\ &= \sum_c P(c)P(Y|u,i,M(c,u)) \end{aligned} \tag{4.17}$$

where $do(U=u)$ represents the intervention operation that forces user representation to u regardless of confounders c. This intervention effectively removes the direct influence of confounders on user representation by replacing $P(c|u)$ with prior probability $P(c)$, thereby breaking spurious correlations while preserving genuine causal relationships between users and items. The resulting $P(Y|do(U=u),I=i)$ better reflects users' true preferences by excluding confounding effects from both social networks and item groups.

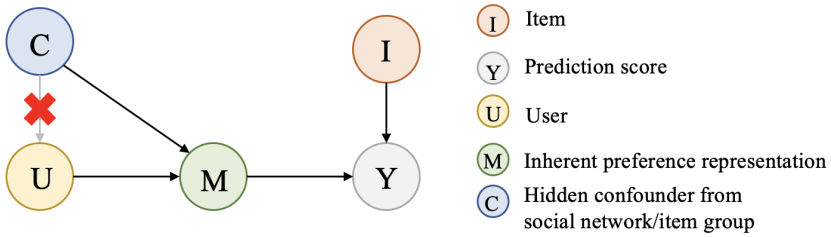


Figure 4.9: Our designed causal graph for debiasing two confounders through the back-door adjustment.

4.2.2.2 Methodology

Figure 4.10 presents our GCRec framework, which systematically addresses multiple confounders through four integrated components. The first component ***Embedding Initialization*** aims to build strong foundation representations by effectively encoding rich attribute

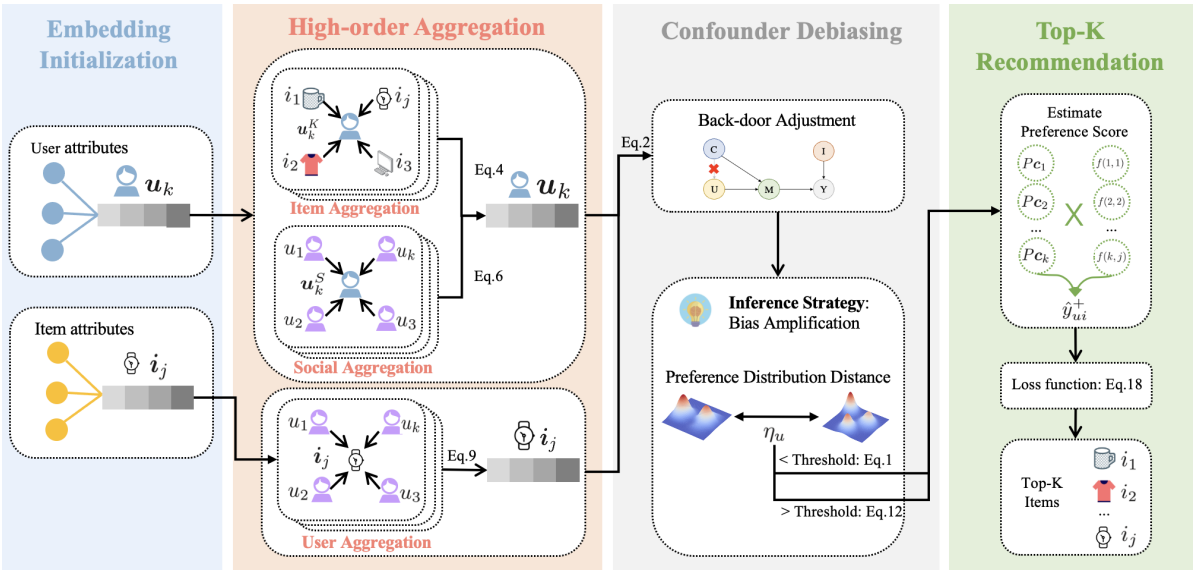


Figure 4.10: The overall framework of our proposed method GCRec.

information. Specifically, we transform user and item attributes into numerical embeddings through a database retrieval function to generate initial representations $\boldsymbol{u}_k \in \mathbb{R}^q$ for user k and $\boldsymbol{i}_j \in \mathbb{R}^a$ for item j [137, 35]. To simplify the notation, we ignore the attribute index q and a for k-th user and j-th item. Beyond these basic embeddings, we capture interaction feedback through a dense vector $\mathbf{e}_y$ that models rating scores (e.g., encoding a 5-star rating as $[0,0,0,0,5]$) [48]. These carefully constructed initial embeddings serve as the building blocks for subsequent GNN-based high-order aggregation, enabling us to learn comprehensive deconfounded representations while maintaining model generalizability.

Next, the second component ***High-order Aggregation*** aims to refine the initial embeddings into comprehensive user/item embeddings by fusing information across both social networks and historical interactions. Specifically, we construct rating-aware representation $\mathbf{x}_{kj}$ that denotes the interaction between the u_k and i_j with rating y:

$$\mathbf{x}_{kj} = MLP\left(\left[\boldsymbol{i}_j \oplus \mathbf{e}_y\right]\right) \tag{4.18}$$

where $\oplus$ is the concatenation operation. $\mathbf{x}_{kj}$ is the output of MLP that integrates the interactive information with the rating. Based on these rating-aware vectors, we generate item-space user representations $\boldsymbol{u}_k^K$ using an attention-weighted aggregation:

$$\begin{aligned} \boldsymbol{u}_k^K &= \sigma\left(\mathbf{W} \cdot \text{Aggre}_{items}\left(\mathbf{x}_{kj}, \forall j \in S_k\right) + \mathbf{b}\right) \\ &= \sigma(\mathbf{W} \cdot \sum_{j \in S_k} \alpha_{kj} \mathbf{x}_{kj} + \mathbf{b}) \end{aligned} \tag{4.19}$$

where S_k is the user's interacted item set and σ is a ReLU activation. $Aggre_{items}$ is an item aggregation function that uses a linear approximation of a localized spectral convolution [88] to compute the element-wise mean of vectors. Although it assumes all neighbours contribute equally to the user representation, the effects of interactions on users can differ greatly. Thus, we allow interactions to contribute differently by assigning a weight to each interaction:

$$\begin{aligned} \alpha^*_{kj} &= \mathbf{W}_2^T \cdot \sigma\left(\mathbf{W}_1 \cdot \left[\mathbf{x}_{kj} \oplus \boldsymbol{u}_k\right] + \mathbf{b}_1\right) + b_2 \\ \alpha_{kj} &= \frac{\exp\left(\alpha^*_{kj}\right)}{\sum_{j \in S_k} \exp\left(\alpha^*_{kj}\right)} \end{aligned} \tag{4.20}$$

where α^*_{kj} is the attention weight of the interaction, and α_{kj} is the attention of $Aggre_{items}$ computed using a two-layer neural network. Moreover, based on social correlation theory [150], where users' preferences typically align with their direct social connections. Thus, we aggregate high-order social relationships into user representations as deconfounders:

$$\begin{aligned} \boldsymbol{u}^S_k &= \sigma\left(\mathbf{W} \cdot \text{Aggre}_{neighbors}\left(\boldsymbol{u}^K_o, \forall o \in N_k\right) + \mathbf{b}\right) \\ &= \sigma(\mathbf{W} \cdot \sum_{o \in N_k} \beta_{ko} \boldsymbol{u}^K_o + \mathbf{b}) \end{aligned} \tag{4.21}$$

where $\boldsymbol{u}^S_k$ aggregates neighbor vectors through element-wise mean over social node set N_k. $Aggre_{neighbors}$ is an aggregation function on the user's neighbors by taking the element-wise mean of embeddings in $\left\{\boldsymbol{u}^K_o, \forall o \in N_k\right\}$. Since users share preferences more with strong social ties [150], we introduce social attention βko using a two-layer neural network:

$$\begin{aligned} \beta^*_{ko} &= \mathbf{W}_2^T \cdot \sigma\left(\mathbf{W}_1 \cdot \left[\boldsymbol{u}^K_o \oplus \boldsymbol{u}_k\right] + \mathbf{b}_1\right) + b_2 \\ \beta_{ko} &= \frac{\exp\left(\beta^*_{ko}\right)}{\sum_{o \in N_k} \exp\left(\beta^*_{ko}\right)} \end{aligned} \tag{4.22}$$

where the β_{ko} denotes the final social attention weight of the strengths that normalized from attentive scores β^*_{ko} with Softmax. The final user latent factor combines item-space factors $\boldsymbol{u}^K_k$ from item set S_k and social-space factors $\boldsymbol{u}^S_k$ through an ℓ-layer MLP:

$$\begin{aligned} \boldsymbol{u}^{(0)}_k &= MLP(\left[\boldsymbol{u}^K_k \oplus \boldsymbol{u}^S_k\right]) \\ \boldsymbol{u}^{(1)}_k &= \sigma\left(\mathbf{W}_2 \cdot \boldsymbol{u}^{(0)}_k + \mathbf{b}_2\right) \\ &\cdots \\ \boldsymbol{u}_k &= \sigma\left(\mathbf{W}_\ell \cdot \boldsymbol{u}^{(\ell-1)}_k + \mathbf{b}_\ell\right) \end{aligned} \tag{4.23}$$

where $\boldsymbol{u}_k$ is the GNN-based high-order user representation. For items connected through user interactions and ratings, we learn item latent factors through user aggregation:

$$\begin{aligned} \boldsymbol{i}_j &= \sigma\left(\mathbf{W} \cdot \text{Aggre}_{users}\left(\mathbf{f}_{jk}, \forall k \in B(j)\right) + \mathbf{b}\right) \\ &= \sigma(\mathbf{W} \cdot \sum_{k \in B(j)} \mu_{jk} \mathbf{f}_{jk} + \mathbf{b}) \end{aligned} \tag{4.24}$$

where $B(j)$ denotes a set of users who interacted with $\boldsymbol{i}_j$. $\mathbf{f}_{jk} = \text{MLP}([\boldsymbol{u}_k \oplus \mathbf{e}_y])$ represents rating-aware interactions for users in set $B(j)$. Given heterogeneous interaction influences benefit item representation learning [150], we develop attention mechanism μ_{jk} to capture these effects:

$$\begin{aligned} \mu^*_{jk} &= \mathbf{W}_2^T \cdot \sigma\left(\mathbf{W}_1 \cdot \left[\mathbf{f}_{jk} \oplus \boldsymbol{i}_j\right] + \mathbf{b}_1\right) + b_2 \\ \mu_{jk} &= \frac{\exp\left(\mu^*_{jk}\right)}{\sum_{k \in B_{(j)}} \exp\left(\mu^*_{jk}\right)} \end{aligned} \tag{4.25}$$

where μ_{jk} represents normalized attention weights indicating the importance of each user's contribution to the item representation $\boldsymbol{i}_j$.

Then, the third component ***Confounder Debiasing*** aims to address both social network and item group confounders through back-door adjustment. For social network confounders, we directly apply user representation $\boldsymbol{u}_k$ encoded by GNNs to back-door adjustment. For item group confounders, we compute an approximation due to the unlimited sample space of historical distributions. Specifically, we construct a discrete set $\tilde{\mathscr{C}}$ containing users' historical distributions $\boldsymbol{c}_u$ over item groups as $\boldsymbol{c}_u = \left[p_{\boldsymbol{u}_k}(g_1), \ldots, p_{\boldsymbol{u}_k}(g_n)\right] \in \tilde{\mathscr{C}}$, where $p_{\boldsymbol{u}_k}(g_n)$ is the interaction frequency of k-th user on group g_n. Mathematically, we have

$$p_{\boldsymbol{u}_k}(g_n) = \sum_{\boldsymbol{i}_j \in I} p\left(g_n \mid \boldsymbol{i}_j\right) p(\boldsymbol{i}_j \mid \boldsymbol{u}_k) = \frac{\sum_{\boldsymbol{i}_j \in \mathscr{H}_{\boldsymbol{u}_k}} q_{g_n}^{\boldsymbol{i}_j}}{\left|\mathscr{H}_{\boldsymbol{u}_k}\right|} \tag{4.26}$$

where $q_{g_n}^{\boldsymbol{i}_j}$ is the likelihood of j-th item belongs to the item group g_n. Since we sample C based on the user-item interactions,so the probability $P(\boldsymbol{c}_u)$ for the user can be computed by $\frac{\left|\mathscr{H}_{\boldsymbol{u}_k}\right|}{\sum_{v \in \mathscr{U}} \mathscr{H}_v}$. Each $\boldsymbol{c}_u$ denotes the distribution of user u and we employ a Factorization Machine $f(\cdot)$ [137] to compute the conditional probability $P(Y \mid \boldsymbol{u}_k, \boldsymbol{i}_j, M(\boldsymbol{c}_u, \boldsymbol{u}_k))$. Based on Jensen Gap and machine learning theory [180], we can approximate the back-door adjust-

ment as:

$$
\begin{aligned}
P\left(Y \mid do\left(U=\boldsymbol{u}_k\right), I=\boldsymbol{i}_j\right) &\approx \sum_{\boldsymbol{c}\in\tilde{\mathscr{C}}} P\left(\boldsymbol{c}_u\right) P\left(Y \mid \boldsymbol{u}_k, \boldsymbol{i}_j, M\left(\boldsymbol{c}_u, \boldsymbol{u}_k\right)\right) \\
&\approx \sum_{\boldsymbol{c}\in\tilde{\mathscr{C}}} P(\boldsymbol{c}_u) f(\boldsymbol{u}_k, \boldsymbol{i}_j, M(\boldsymbol{c}_u, \boldsymbol{u}_k)) \\
&\approx f(\boldsymbol{u}_k, \boldsymbol{i}_j, M(\sum_{\boldsymbol{c}\in\tilde{\mathscr{C}}} P\left(\boldsymbol{c}_u\right) \boldsymbol{c}_u, \boldsymbol{u}_k))
\end{aligned} \tag{4.27}
$$

where $M(\boldsymbol{c}_u, \boldsymbol{u}_k)$ is inherent preference representation capturing user preferences under historical distribution $\boldsymbol{c}_u$. The preference function $f(\cdot)$ takes user embedding $\boldsymbol{u}_k$, item embedding $\boldsymbol{i}_j$, and historical distribution representation $M(\overline{\boldsymbol{c}}, \boldsymbol{u}_k)$ as inputs, where $\overline{\boldsymbol{c}} = \sum_{\boldsymbol{c}\in\tilde{\mathscr{C}}} P\left(\boldsymbol{c}_u\right) \boldsymbol{c}_u = \left[p\left(g_1\right), \ldots, p\left(g_n\right)\right]$. Given that high-order connectivity provides richer semantics [186], we employ a Factorization Machine to aggregate from high-order connectivity as

$$
M(\overline{\boldsymbol{c}}, \boldsymbol{u}_k) = \sum_{n=1}^{N} \sum_{k=1}^{K} \left(p\left(g_n\right) \boldsymbol{v}_n\right) \odot \left(u_k[1] \boldsymbol{u}_k\right) \tag{4.28}
$$

where $\odot$ denotes element-wise product, $u_k[1]$ is the first attribute value, and $\boldsymbol{v}_n$ represents randomly initialized group embeddings. This captures high-order user preferences under historical distributions across N item groups and K users. To implement the intervention on the user, we estimate preference scores through: $P(Y \mid \boldsymbol{u}_k, \boldsymbol{i}_j, M(\overline{\boldsymbol{c}}_u, \boldsymbol{u}_k)) \propto f\left(\boldsymbol{u}_k, \boldsymbol{i}_j, M(\overline{\boldsymbol{c}}, \boldsymbol{u}_k)\right)$. Mathematically, we have

$$
\begin{aligned}
& f\left(\boldsymbol{u}_k, \boldsymbol{i}_j, M(\overline{\boldsymbol{c}}, \boldsymbol{u}_k)\right) \\
&= \boldsymbol{u}_k \cdot \mathrm{ReLU}(\mathbf{E}_\ell \cdot \mathrm{ReLU}(\mathbf{E}_{\ell-1} \cdot \mathrm{ReLU}(\cdots(\mathbf{E}_1 \cdot \mathrm{ReLU}[\begin{matrix} \boldsymbol{i}_j \\ M(\overline{\boldsymbol{c}}, \boldsymbol{u}_k) \end{matrix}])\ldots)))^\top
\end{aligned} \tag{4.29}
$$

where $\mathbf{E}_\ell$ represents trainable weights in the ℓ-th MLP layer. ReLU is the rectified linear unit activation function [4]. For a triple $\left(\boldsymbol{u}_k, \boldsymbol{i}_j, M(\overline{\boldsymbol{c}}, \boldsymbol{u}_k)\right)$, we compute preference scores by encoding item and historical preference representations through MLP layers followed by dot product with user embeddings. In this way, we can effectively address both social and item group confounders based on the back-door adjustment.

To effectively balance the positive and negative effects of confounding, we introduce an adaptive inference strategy. Unlike prior approaches that only consider negative effects, we recognize that bias amplification can benefit users with stable preferences while harming those with evolving interests. Technically, we employ the symmetric Kullback-Leibler(KL) divergence to measure preference drift by splitting historical interactions into two temporal parts [170, 181]: $\boldsymbol{c}_u^1 = \left[p_{u_k^1}\left(g_1\right), \ldots, p_{u_k^1}\left(g_N\right)\right]$ and $\boldsymbol{c}_u^2 = \left[p_{u_k^2}\left(g_1\right), \ldots, p_{u_k^2}\left(g_N\right)\right]$, where $\boldsymbol{c}_u^1$ and $\boldsymbol{c}_u^2$ represent historical distributions over item groups for earlier and later periods. The

divergence η_u indicates preference stability, with larger values suggesting more dynamic preferences:

$$\begin{aligned}\eta_u &= KL\left(\boldsymbol{c}_u^1 \mid \boldsymbol{c}_u^2\right) + KL\left(\boldsymbol{c}_u^2 \mid \boldsymbol{c}_u^1\right) \\ &= \sum_{k=1}^{K} p_{u_k^1}(g_n) \log \frac{p_{u_k^1}(g_n)}{p_{u_k^2}(g_n)} + \sum_{k=1}^{K} p_{u_k^2}(g_n) \log \frac{p_{u_k^2}(g_n)}{p_{u_k^1}(g_n)}\end{aligned} \tag{4.30}$$

where η_u denotes the divergence between the user's preference distributions. We then adaptively combine conventional and causal predictions:

$$\hat{y}_{ui}^{+} = \left(1 - \hat{\eta}_u\right) * \hat{y}_{u,i}^{RS} + \hat{\eta}_u * f\left(\boldsymbol{u}_k, \boldsymbol{i}_j, M(\overline{\boldsymbol{c}}, \boldsymbol{u}_k)\right) \tag{4.31}$$

where $\hat{y}_{ui}^{+}$ denotes the inference prediction score for a user-item pair. $\hat{y}_{u,i}^{RS}$ and $f\left(\boldsymbol{u}_k, \boldsymbol{i}_j, M(\overline{\boldsymbol{c}}, \boldsymbol{u}_k)\right)$ are the prediction scores computed by the conventional recommendation and our GCRec. $\hat{\eta}_u$ is normalized as

$$\hat{\eta}_u = \left(\frac{\eta_u - \eta_{\min}}{\eta_{\max} - \eta_{\min}}\right)^{\Gamma} \tag{4.32}$$

where $\eta_{\min}$ and $\eta_{\max}$ are the minimum and maximum values of η_u for all users. $\Gamma \in [0, +\infty)$ is a hyperparameter to control the weights of $\hat{y}_{u,i}^{RS}$ and $f\left(\boldsymbol{u}_k, \boldsymbol{i}_j, M(\overline{\boldsymbol{c}}, \boldsymbol{u}_k)\right)$ by manual intervention. $\hat{\eta}_u$ will be larger if Γ is close to 0, thus users with high $\hat{\eta}_u$ will rely more on $f\left(\boldsymbol{u}_k, \boldsymbol{i}_j, M(\overline{\boldsymbol{c}}, \boldsymbol{u}_k)\right)$. In simple terms, a smaller η_u means the users' preferences are stable, so we increase Γ to amplify their preferences using bias amplification caused by confounder (i.e., positive effects). In contrast, if the users' preferences change frequently, we could reduce Γ to mitigate the bias amplification for users' preferences (i.e., negative effects).

Finally, our last ***Top-K Recommendation*** component evaluates performance using widely-adopted Top-K metrics. For each user-item pair, we compute the preference score $\hat{y}_{ui}^{+}$ and randomly sample negative items for comparison. The learning objective is defined as:

$$\mathscr{L} = -\sum_{u \in \mathscr{U}} \sum_{u_i \in \mathscr{F}_u^{+}} \sum_{u_j \in \mathscr{F}_u^{-}} \ln\left(\sigma\left(\hat{y}_{ui}^{+} - \hat{y}_{uj}^{-}\right)\right) + \lambda_{\Theta} \|\Theta\|_2^2 \tag{4.33}$$

where $\mathscr{F}_u^{+}$ and $\mathscr{F}_u^{-}$ denote observed and negative item sets respectively, Θ represents model parameters, and λ_{Θ} controls l_2 regularization. After controlling confounding effects through the inference strategy, we generate final Top-K recommendations based on predicted preference scores.

4.2.3 Experiments

4.2.3.1 Datasets

We evaluate our proposed GCRec on two benchmark datasets: `MovieLens`[3], and `Douban-Movie`[4]. As shown in Table 4.3, for each dataset, we record the number of nodes including users, items and their attributes, relationships between entities, density, and average degrees. We binarize user feedback by treating ratings $y \geq 4$ as positive interactions while randomly sampling uninteracted items as negative feedback.

Table 4.3: GCRec: Statistical details of the two datasets.

Dataset (Density)	Node/#	Relation A-B/#	Ave.Degree of A-B
MovieLens (6.30%)	User(U)/943 Age(A)/8 Occupation(O)/21 Movie(M)/1682 Genre(G)/18	U-M/100000 U-U/47150 U-A/943 U-O/943 M-M/82798 M-G/2861	#U-M: 106.04-59.45 #U-U: 50.00-50.00 #U-A: 1-117.88 #U-O: 1-44.90 #M-M: 49.23-49.23 #M-G: 1.70-158.94
Douban-Movie (0.63%)	User(U)/13367 Movie(M)/12677 Group(G)/2753 Actor(A)/6311 Director(D)/2449 Type(T)/38	U-M/1068278 U-G/570047 U-U/4085 M-A/33587 M-D/11276 M-T/27668	#U-M: 79.92-84.27 #U-G: 45.65-207.06 #U-U: 0.31-0.31 #M-A: 2.65-5.32 #M-D: 0.89-4.60 #M-T: 2.18-728.11

4.2.3.2 Baselines and Evaluation

For implementation, we set the embedding dimension to 128, MLP layer number ℓ is 2, Γ is 0.01, and learning rate is 0.01, with regularization parameter $1e-5$, using Adam optimizer with learning rate tuned in $[0.0001, 0.1]$. Baseline models are tuned using the same hyperparameter range as ours for fair comparison. For recommendation performance evaluation, we employ three widely used metrics [232, 186, 99] Recall@K, NDCG@K, and Precision@K where $K \in \{1, 10, 20, 40\}$. All datasets are split into training/testing/validation sets with ratios of 80%/10%/10%.

We compare GCRec to the following state-of-the-art baselines:

- **BPR** [138] is a personalized ranking model that uses matrix factorization with implicit feedback to predict user preferences.

[3]https://github.com/librahu/HIN-Datasets-for-Recommendation-and-Network-Embedding
[4]https://github.com/librahu/HIN-Datasets-for-Recommendation-and-Network-Embedding

- **DMF** [221] employs deep neural networks to filter the user-item interaction matrix for preference prediction.
- **FairCo** [120] incorporates user attributes to control selection bias in item exposure mechanisms.
- **FSF** [55] enforces fairness by removing non-sensitive user attributes to address historical distribution bias.
- **IPS** [142] employs propensity clipping techniques to debias recommender systems through causal learning.
- **DENC** [100] is a causal learning method that addresses missing-not-at-random problems through social network debiasing.
- **PDA** [248] leverages the confounding effects from popularity bias through the back-door adjustment of causality when predicting the recommendation results.
- **DecRS** [181] incorporates user and item attributes into representations and then uses the back-door adjustment to debias the item group confounder over user historical distribution.
- **DICE** [253] learns disentangled representations through causal inference to separate user interests from conformity bias.

4.2.3.3 Result Analysis

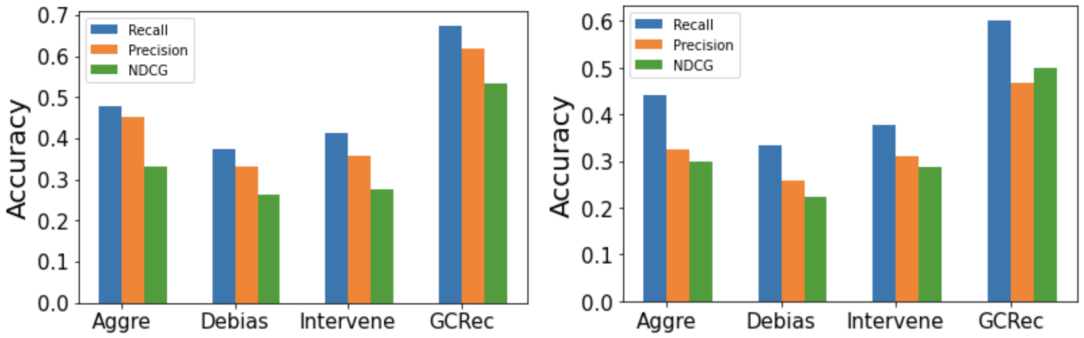
Performance Comparison: Table 4.4 demonstrates GCRec's performance against baseline methods on Top-K recommendations across two datasets. Note that *Db-Mov* is short for *Douban-Movie*. It's worth noting that causal methods obtain the highest accuracy among all baselines, verifying the causal inference is a promising direction to further enhance the recommendation systems Specifically, our results show that GCRec consistently outperforms all baselines across key metrics. For example, compared to the strongest baseline, GCRec achieves significant improvements on *MovieLens* (Recall@10: +20.28%, NDCG@10: +12.41%, Precision@10: +2.55%) and *Douban-Movie* (Recall@10: +7.04%, NDCG@10: +17.3%, Precision@10: +11.34%). The superior performance can be attributed to two key factors: (1) the effectiveness of our back-door adjustment in debiasing multiple confounders, and (2) the robust high-order connectivity modeling through user/item aggregation. Notably,

GCRec maintains strong performance even on the sparser *Douban-Movie* dataset, demonstrating its capability in handling sparse data structures through effective high-order connectivity integration.

Table 4.4: GCRec: Recommendation performance comparisons: the best results are marked as bold, strongest baselines are marked as bold with underline.

Datasets	Metrics	FairCo	FSF	IPS	DENC	BPR	DMF	PDA	DecRS	DICE	GCRec	Improv.
MovieLens	Recall@1	0.122	0.134	**0.388**	0.384	0.222	0.212	0.379	0.355	0.374	**0.409**	5.41%
	Recall@10	0.398	0.401	0.442	0.561	0.331	0.247	**0.498**	0.481	0.456	**0.599**	20.28%
	Recall@20	0.489	0.517	0.552	**0.622**	0.482	0.411	0.577	0.581	0.547	**0.675**	8.52%
	Recall@40	0.667	0.712	0.691	**0.715**	0.597	0.532	0.707	0.686	0.669	**0.758**	6.01%
	Precision@1	0.141	0.113	0.201	0.221	0.241	0.188	**0.278**	0.277	0.251	**0.288**	3.60%
	Precision@10	0.442	0.423	0.452	**0.471**	0.388	0.241	0.442	0.411	0.347	**0.483**	2.55%
	Precision@20	0.521	0.589	0.576	**0.602**	0.501	0.399	0.598	0.577	0.535	**0.618**	2.66%
	Precision@40	0.689	**0.718**	0.702	0.711	0.611	0.598	0.699	0.652	0.633	**0.741**	3.20%
	NDCG@1	0.201	0.227	0.215	0.237	0.199	0.196	**0.268**	0.225	0.201	**0.292**	8.96%
	NDCG@10	0.356	0.361	0.377	**0.395**	0.302	0.286	0.366	0.342	0.299	**0.444**	12.41%
	NDCG@20	0.412	0.421	0.402	0.501	0.355	0.491	**0.511**	0.488	0.456	**0.534**	4.50%
	NDCG@40	0.578	0.622	0.588	**0.647**	0.496	0.611	0.623	0.611	0.558	**0.673**	4.02%
Db-Mov	Recall@1	0.101	0.111	0.235	0.234	0.201	0.202	**0.288**	0.256	0.231	**0.302**	4.86%
	Recall@10	0.266	0.232	0.389	**0.412**	0.287	0.233	0.333	0.378	0.375	**0.441**	7.04%
	Recall@20	0.375	0.411	0.445	**0.563**	0.396	0.387	0.476	0.532	0.498	**0.602**	6.93%
	Recall@40	0.512	0.521	0.578	**0.652**	0.515	0.489	0.589	0.622	0.587	**0.689**	5.67%
	Precision@1	0.099	0.101	0.172	0.182	0.201	0.144	**0.206**	0.175	0.163	**0.225**	9.22%
	Precision@10	0.189	0.215	0.287	0.285	0.321	0.212	**0.335**	0.311	0.286	**0.373**	11.34%
	Precision@20	0.312	0.376	0.356	0.401	0.399	0.302	**0.439**	0.428	0.399	**0.467**	6.38%
	Precision@40	0.479	0.489	0.477	**0.522**	0.487	0.442	0.521	0.511	0.485	**0.578**	10.73%
	NDCG@1	0.111	0.123	0.147	0.155	0.146	0.155	**0.212**	0.199	0.174	**0.221**	4.25%
	NDCG@10	0.215	**0.322**	0.299	0.276	0.276	0.251	0.306	0.298	0.251	**0.378**	17.3%
	NDCG@20	0.336	0.389	0.378	**0.452**	0.301	0.376	0.398	0.401	0.357	**0.499**	10.40%
	NDCG@40	0.442	0.485	0.498	**0.546**	0.425	0.455	0.512	0.525	0.481	**0.576**	5.49%

Ablation Study: To thoroughly evaluate each component's contribution to GCRec's performance, we conduct comprehensive ablation experiments by systematically removing key components. Our model consists of three functional components: (1) high-order user/item representation generation through embedding initialization and aggregation, (2) confounder debiasing via $M(\overline{\boldsymbol{c}}, \boldsymbol{u}_k)$, and (3) user intervention through $f\left(\boldsymbol{u}_k, \boldsymbol{i}_j, M(\overline{\boldsymbol{c}}, \boldsymbol{u}_k)\right)$. As shown in Figure 4.11, we evaluate four variants: $Aggre$ (removing high-order aggregation), $Debias$ (removing confounder debiasing), $Intervene$ (removing user intervention), and complete *GCRec*. The results demonstrate that confounder debiasing has the most substantial impact, with its removal causing approximately 30% performance degradation across both datasets. The removal of aggregation ($Aggre$) and intervention ($Intervene$) components also significantly impacts performance, each leading to roughly 20% accuracy reduction. These findings validate our design choices in addressing confounding effects through multiple complementary strategies: GNN-based high-order representations, back-door adjustment, and adaptive inference. The comprehensive integration of these components en-



(a) The influence of three functional components at K@20 on *MovieLens*.

(b) The influence of three functional components at K@20 on *Douban-Movie*.

Figure 4.11: GCRec: Ablation study.

ables GCRec to achieve superior performance across all metrics on both benchmark datasets.

User Group Analysis: To evaluate GCRec's effectiveness across different user behaviors, we analyze performance across user groups stratified by preference stability measure η_u. Table 4.5 presents Recall@20 and Precision@20 metrics across user groups defined by η_u thresholds $\{0, 0.5, 1, 2, 3, 4\}$, where higher η_u indicates greater susceptibility to confounding effects from item groups and social networks. Our results show that GCRec's performance advantage increases with η_u. Taking *Douban-Movie* as an example, improvements over DICE [253] on Recall@20 consistently increase across user groups: $\eta_u > 0$ (26.47%), $\eta_u > 0.5$ (27.86%), $\eta_u > 1$ (29.04%), $\eta_u > 2$ (29.84%), $\eta_u > 3$ (30.24%), and $\eta_u > 4$ (30.33%). This pattern demonstrates the effectiveness of our adaptive inference strategy in adjusting back-door adjustment according to user preference stability. The increasing performance gap for users with higher η_u validates our approach's capability to handle confounding effects through two key mechanisms: (1) effectively reducing spurious correlations through adaptive back-door adjustment, and (2) accurately capturing diverse preferences for users with fluctuating interests. These results confirm that our inference strategy successfully controls bias amplification while maintaining robust recommendations.

Inference Strategy Analysis: To evaluate the effectiveness of our inference strategy, we further conduct more detailed ablation experiments regarding our inference component. Table 4.6 presents performance comparisons where GCRec(w/o) represents the model using only basic prediction $\hat{y}^{+}_{ui}$ without the inference strategy. The results demonstrate significant performance gains from the inference strategy, with improvements in Recall@20 of 14.47% on *MovieLens* and 6.17% on *Douban-Movie*. Notably, even without the inference strategy, GCRec(w/o) outperforms state-of-the-art baselines including DICE [253], DecRS [181], and DENC [100], validating the effectiveness of our high-order aggregation

Table 4.5: GCRec: Performance comparison across different user groups.

	MovieLens						**Douban-Movie**					
Metrics	Precision@20			Recall@20			Precision@20			Recall@20		
Threshold	DICE	GCRec	Improv.	DICE	GCRec	Improv.	DICE	GCRec	Improv.	DICE	GCRec	Improv.
0	0.598	0.618	**3.34%**	0.577	0.675	**16.98%**	0.439	0.467	**6.38%**	0.476	0.602	**26.47%**
0.5	0.602	0.629	**4.49%**	0.581	0.689	**18.59%**	0.446	0.481	**7.85%**	0.406	0.515	**26.85%**
1	0.605	0.638	**5.45%**	0.592	0.706	**19.26%**	0.452	0.501	**10.84%**	0.489	0.631	**29.04%**
2	0.611	0.651	**6.55%**	0.605	0.721	**19.17%**	0.462	0.521	**12.77%**	0.496	0.644	**29.84%**
3	0.615	0.662	**7.64%**	0.615	0.741	**20.49%**	0.481	0.511	**13.10%**	0.506	0.659	**30.24%**
4	0.621	0.677	**9.02%**	0.628	0.759	**20.86%**	0.502	0.572	**13.94%**	0.521	0.679	**30.33%**
Metrics	Precision@20			Recall@20			Precision@20			Recall@20		
Threshold	DecRS	GCRec	Improv.	DecRS	GCRec	Improv.	DecRS	GCRec	Improv.	DecRS	GCRec	Improv.
0	0.577	0.618	**7.11%**	0.581	0.675	**16.18%**	0.428	0.467	**9.11%**	0.532	0.602	**13.16%**
0.5	0.581	0.629	**8.26%**	0.585	0.689	**17.78%**	0.432	0.481	**11.34%**	0.536	0.615	**14.74%**
1	0.588	0.638	**8.50%**	0.588	0.706	**20.07%**	0.448	0.501	**11.83%**	0.544	0.631	**15.99%**
2	0.591	0.651	**10.15%**	0.593	0.721	**21.59%**	0.454	0.521	**14.76%**	0.551	0.644	**16.88%**
3	0.598	0.662	**10.70%**	0.601	0.741	**23.29%**	0.463	0.544	**17.49%**	0.562	0.659	**17.26%**
4	0.605	0.677	**11.90%**	0.608	0.759	**24.84%**	0.476	0.572	**20.17%**	0.566	0.679	**19.96%**

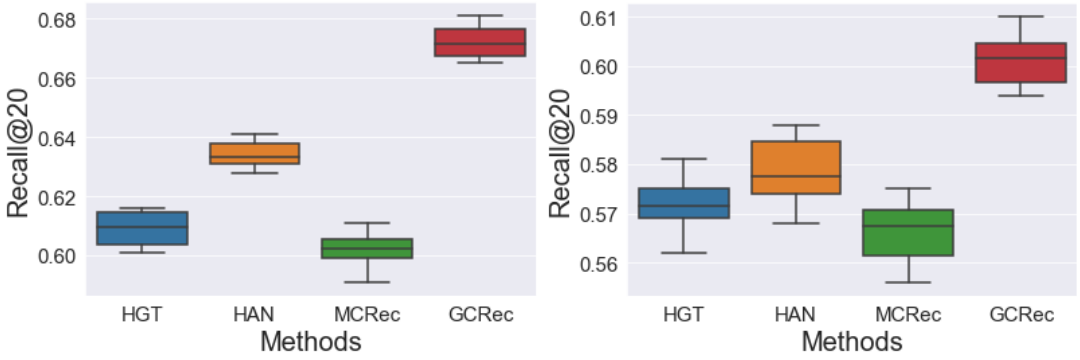
Table 4.6: GCRec: Inference strategy analysis across two datasets. The strongest baselines are marked with underline, and the improvement rates are marked as bold.

	MovieLens			**Douban-Movie**		
Metrics	Recall@20	Precision@20	NDCG@20	Recall@20	Precision@20	NDCG@20
GCRec(w/o)	0.615	0.602	0.518	0.567	0.447	0.459
GCRec	0.675	0.618	0.534	0.602	0.467	0.499
Improv.	**14.47**%	**2.66**%	**3.09**%	**6.17**%	**4.17**%	**8.71**%
DecRS	<u>0.581</u>	0.577	0.488	<u>0.532</u>	0.428	<u>0.401</u>
DICE	0.547	0.535	0.456	0.498	0.399	0.357
DENC	0.577	<u>0.598</u>	<u>0.511</u>	0.467	<u>0.439</u>	0.398
GCRec(w/o)	0.615	0.602	0.518	0.567	0.447	0.459
Improv.	**5.85**%	**0.67**%	**1.37**%	**6.58**%	**1.82**%	**14.46**%

and back-door adjustment in reducing confounding effects. The performance improvement is particularly pronounced on the sparser *Douban-Movie* dataset, demonstrating our inference strategy's capability to mitigate bias amplification through effective high-order aggregation. These results confirm that GCRec's combination of back-door adjustment and adaptive inference successfully addresses confounding effects while maintaining robust performance across different data distributions.

Attention Mechanisms Analysis: To evaluate our attention mechanism design, we compare GCRec's performance against state-of-the-art attention-based approaches. Three baselines are employed as they used meta-path, which also involves rich contextual information as ours: HGT [73] employs attention based on transformer architecture; HAN [184] employs GNNs to model context from meta-paths and hierarchical attention and MCRec [72] employs meta-path based context with the co-attention mechanism. Figure 4.12 presents

the Recall@20 results across 20 runs on the *MovieLens* and *Douban-Movie* datasets. It's worth noting that our GCRec consistently outperforms these baselines, achieving a 4.2% improvement over the strongest baseline (HAN) on *MovieLens*. This superior performance can be attributed to our multi-faceted attention design that separately models item attention (α_{kj}), social attention (β_{ko}), and user attention (μ_{jk}). Moreover, GCRec demonstrates greater stability with performance fluctuations under 0.015, suggesting that our differentiated attention approach effectively captures complementary contextual signals from multiple interaction types. These results validate the effectiveness of modeling distinct attention mechanisms for different relationship types rather than using a uniform attention strategy.



(a) Attention methods comparison on Recall@20 on *MovieLens*.

(b) Attention methods comparison on Recall@20 on *Douban-Movie*.

Figure 4.12: GCRec: Performance comparison over different attention-based baselines.

4.2.3.4 Summary

In this work, we propose GCRec, a novel causal framework that addresses RQ2 regarding enhancing recommender system robustness against spurious correlations in complex recommendation scenarios. By synergistically integrating GNNs with adaptive back-door adjustment, GCRec effectively mitigates confounding effects from both social networks and item groups while preserving beneficial bias amplification through preference-aware inference. Extensive experiments demonstrate GCRec's superior performance over state-of-the-art baselines in terms of recommendation accuracy and robustness across different user groups. Future work will explore combining front-door and back-door adjustments to further enhance model robustness in complex structures.

4.3 Breaking The Loop: Causal Learning To Mitigate Echo Chambers In Social Networks

4.3.1 Overview

In social networks, echo chambers form when users primarily encounter information that reinforces their existing beliefs, limiting exposure to diverse perspectives. These self-reinforcing information spaces can worsen societal issues such as polarization and declining public discourse. Traditional approaches attempt to address echo chambers by analyzing observable interaction patterns to identify their formative mechanisms. However, they overlook unobserved implicit factors, known as hidden confounders in causal inference, that significantly shape information diffusion patterns despite not being directly captured in the data. These hidden confounders simultaneously affect both content exposure and multiple users' behaviors across social networks, creating spurious correlations that mask true causal factors driving the echo chambers. Therefore, addressing hidden confounders becomes essential for developing more effective strategies to mitigate echo chambers and promote information diversity in social networks while maintaining user engagement.

4.3.1.1 Research Objective

This study aims to address RQ2 regarding enhancing recommender system robustness against spurious correlations in complex scenarios, particularly in the context of complex social networks. Specifically, we propose to integrate causal inference with transformer architecture to effectively identify and adjust for hidden confounders, that influence content exposure and users' behaviors together and create spurious correlations among them. This integration enables us to distinguish true causal mechanisms from spurious correlations in echo chamber formation, and develop targeted interventions to promote information diversity while maintaining user engagement. In this way, we aim to systematically break echo chambers by reshaping information flows based on causal understanding rather than being misled by spurious correlation patterns in social network interactions.

4.3.1.2 The Proposed Method

To achieve the aforementioned objectives, we propose a novel framework called **C**ausal Echo **D**iffusion **A**ttenuator (**CEDA**) with four synergistic components:

- ***User Dual Modelling*** builds comprehensive user embeddings by combining users'

attributes and structural information within social networks to fully capture user behavior patterns.

- ***Causal Transformer*** uses comprehensive user embeddings to estimate residual embeddings that account for hidden confounders, incorporating them into the Transformer's attention mechanism as causal adjustments to make unbiased user embeddings.
- ***Social Diffusion Predictor*** utilizes unbiased user embeddings to jointly optimize diffusion prediction accuracy and information diversity across the social network.
- ***Targeted Interventions*** strategically reshapes information flows to disrupt echo chambers based on the generated prediction and diversity insights.

The key contributions of this research are:

- To the best of our knowledge, we are the first to apply causal learning to address echo chambers in social networks, providing an innovative perspective for the social network domain.
- We propose a novel framework CEDA, which integrates causal inference with sequential recommendation techniques to mitigate echo chambers by addressing hidden confounders often overlooked but critically influencing information flow patterns.
- We develop a dual-perspective user modeling approach that combines user attributes and structural positions within diffusion sequences, offering a comprehensive representation of user behavior patterns.
- We design a residual embedding-based causal adjustment mechanism for the Transformer, which quantifies the effects of hidden confounders through behavioral discrepancies, enhancing the accuracy of information diffusion prediction and enabling targeted interventions to address echo chambers.

4.3.2 CEDA

4.3.2.1 Problem Definition

Let U be the set of users in social networks, where each user $u_j \in U$ is associated with a set of attributes a_j. Information propagates through the social network via sequential user interactions $S = \{u_1, ..., u_j, ..., u_J\}$, where each transition $\{u_{(j-1)}, u_j\}$ represents content propagated from user $u_{(j-1)}$ to user u_j. As mentioned, these interaction sequences are impacted

by hidden confounders that are not directly observable in the dataset, but affect both content exposure and user interaction patterns, as shown in Figure 4.13. Hidden confounders can create spurious correlations between observed variables, leading to biased estimates of true causal factors behind echo chamber formation. Our goal is to address hidden confounders in social networks to accurately predict information diffusion patterns and reveal true causal mechanisms underlying echo chambers, thereby enabling targeted interventions to effectively promote information diversity while maintaining user engagement.

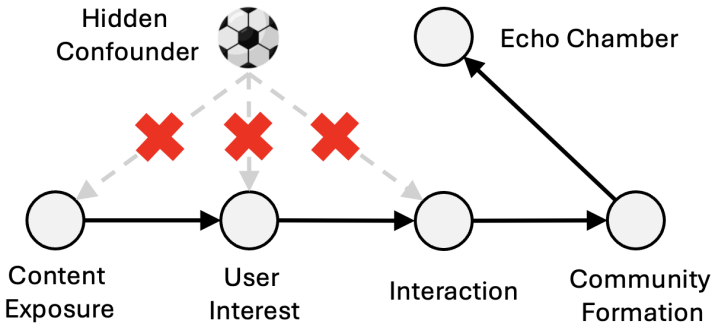


Figure 4.13: Our designed causal graph to illustrate the impact of hidden confounders on echo chamber formation in social networks.

Specifically, we propose a novel causal-based approach called CEDA that integrates causal inference into the social network analysis. Based on users' attributes a_j and their sequential positions in information diffusion paths S, we first learn a comprehensive user embedding e_j for each user u_j to capture user behavior patterns. Following causal inference principles where discrepancies between predicted and observed outcomes may indicate hidden confounder effects, we estimate residual embeddings r_j by computing the difference between predicted content sharing activity based on e_j and actual sharing records O_j. This discrepancy reveals hidden confounders' influence on sharing behaviors that are not directly reflected in user attributes and structural positions. The residual embeddings are then incorporated into the Transformer's attention mechanism as a causal adjustment to adjust attention weights for hidden confounders, thereby refining the comprehensive user embeddings e_j into unbiased user embeddings u_j. Then, we optimize our model through a joint objective to capture different aspects of information spread and diversity:

$$L = \lambda_1 L_r + \lambda_2 L_m(u_j, u_q) + \lambda_3(1 - L_d(u_j)) + \lambda_4(1 - L_c(u_j)) \tag{4.34}$$

where λ_1, λ_2, λ_3 and λ_4 are weighting parameters. L_r ensures the causal adjustment accounts for the effects of hidden confounders, $L_m(u_j, u_q)$ measures the Mean Absolute Error

(MAE) of diffusion probability predictions between user pairs, $L_d(u_j)$ measures the Intra-List Diversity (ILD) of user dissimilarity within diffusion sequences, and $L_c(u_j)$ measures the Category Coverage (CC) to quantify the diversity of content exposure in information diffusion across the network. By optimizing L, our model learns to make accurate predictions while promoting information diversity, enabling targeted interventions that effectively reshape information flows to break echo chambers.

4.3.2.2 Methodology

Figure 4.14 illustrates CEDA's framework, which systematically integrates causal inference principles to address echo chambers in social networks. To begin with, the first component ***User Dual Modelling*** aims to capture the complex user behavior patterns that drive information diffusion in social networks. User behaviors are fundamentally shaped by both individual attributes that determine content preferences and structural network positions that influence interaction opportunities [23, 60, 6, 201]. Therefore, we synergistically combine attribute information with positional data to produce a comprehensive user embedding for each user in the social network. Specifically, for each user u_j, we construct an attribute-based vector $\boldsymbol{a}_j \in \mathbb{R}^d$ using a retrieval function [137], which transforms categorical attributes into numerical encodings for compatibility. On the other hand, inspired by [116], we learn users' sequential position in information cascades through a positional encoding $\boldsymbol{p}_j \in \mathbb{R}^d$:

$$\boldsymbol{p}_j = \left[\sin\left(\frac{j}{10000^{2i/d}}\right), \cos\left(\frac{j}{10000^{2i/d}}\right)\right] \tag{4.35}$$

where j represents the sequential order of user in information cascades. $i \in [0, d/2-1]$ controls the frequency of sinusoidal functions, and $\boldsymbol{p}_j$ denotes the positional encoding vector for the j-th user. In other words, the positional encoding maps each position j to a d-dimensional vector through $d/2$ pairs of sinusoidal functions, with the dimension index $i \in [0, d/2-1]$ controlling the frequency of each sinusoidal pair. This can ensure a unique encoding for each sequential position while maintaining consistent relative distances between positions across different information diffusion sequences in the social network.

With both attribute-based representation $\boldsymbol{a}_j$ and positional encoding $\boldsymbol{p}_j$, we generate the comprehensive user embedding $\boldsymbol{e}_j$ through:

$$\boldsymbol{e}_j = (\boldsymbol{a}_j \oplus \boldsymbol{p}_j)\mathbf{W}^E \tag{4.36}$$

where $\mathbf{W}^E$ denotes a learnable weight matrix and $\oplus$ represents concatenation. The resulting comprehensive user embedding $\boldsymbol{e}_j$ captures both intrinsic user characteristics from

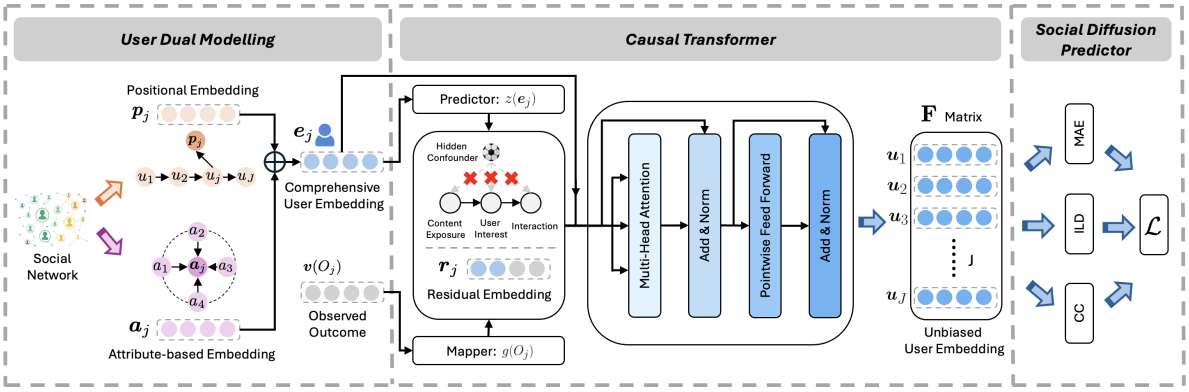


Figure 4.14: The overall framework of our proposed method CEDA.

attributes and structural information from network positions, providing a holistic view of user behavior in the social network.

While comprehensive user embeddings provide a foundation for predicting information diffusion, they fail to account for hidden confounders in social networks that can distort our understanding of echo chamber formation mechanisms. To address this limitation, our second component ***Causal Transformer*** integrates causal inference principles into the Transformer architecture through residual embeddings. The Transformer architecture, with its self-attention mechanism, excels at modeling sequential social interactions by capturing long-range dependencies in user behavior patterns [25, 199]. We enhance this capability by incorporating causal adjustments derived from behavioral discrepancies between predicted and observed outcomes [111, 9]. Mathematically, for each user u_j, we compute a residual embedding $\boldsymbol{r}_j$ as:

$$\boldsymbol{r}_j = \mathrm{MLP}\left(z\left(\boldsymbol{e}_j\right) - g\left(O_j\right)\right) \tag{4.37}$$

$$z(\boldsymbol{e}_j) = \mathrm{ReLU}(\boldsymbol{W}_z \boldsymbol{e}_j + \boldsymbol{b}_z) \tag{4.38}$$

$$g(O_j) = \mathrm{ReLU}(\boldsymbol{W}_g \boldsymbol{v}(O_j) + \boldsymbol{b}_g) \tag{4.39}$$

where $z(\cdot)$ predicts outcomes based on comprehensive user embedding $\boldsymbol{e}_j$, while $g(\cdot)$ transforms observed outcomes O_j into the same embedding space. MLP denotes the multi-layer perceptron. The ReLU activation function introduces non-linearity to capture complex relationships. The discrepancy between these functions reveals the influence of hidden confounders. $\boldsymbol{v}(O_j)$ maps scalar outcomes to high-dimensional vectors [137, 66], while $\mathbf{W}_z$, $\mathbf{W}_g$, $\boldsymbol{b}_z$, and $\boldsymbol{b}_g$ are learnable parameters. Inspired by [116], we train the $z()$ and $g()$ as the

following:

$$\mathscr{L}_r = \sum_{(u_j, O_j)} \left| z\left(\boldsymbol{e}_j\right) - g\left(O_j\right) \right|^2 + \lambda \left(|\mathbf{W}_z|_F^2 + |\mathbf{W}_g|_F^2 \right) \tag{4.40}$$

where λ controls regularization strength and $|\cdot|_F$ denotes the Frobenius norm. By optimizing the $\mathscr{L}_r$, we can quantify and adjust for hidden confounders' effects, leading to more accurate diffusion predictions and more effective echo chamber interventions

To more effectively model hidden confounders across different representational perspectives, we implement a multi-head attention mechanism [33, 140] within our Causal Transformer. This mechanism enables simultaneous processing of user interactions across multiple embedding subspaces while incorporating causal adjustments. For each attention head j, we compute:

$$\text{head}_j = \text{Attention}(\mathbf{E}\mathbf{W}_j^Q, \mathbf{E}\mathbf{W}_j^K, \mathbf{V}\mathbf{W}_j^V, \boldsymbol{r}_j) \tag{4.41}$$

$$= \text{softmax}\left(\frac{\mathbf{E}\mathbf{W}_j^Q (\mathbf{E}\mathbf{W}_j^K)^\top - \mathbf{E}\mathbf{W}_j^Q (\boldsymbol{r}_j \mathbf{1}_J^\top)^\top}{\sqrt{d_k}} \right) \mathbf{V}\mathbf{W}_j^V \tag{4.42}$$

$$\mathbf{O} = (\text{head}_1 \oplus \ldots \oplus \text{head}_j)\mathbf{W}^J \tag{4.43}$$

where $\mathbf{W}_j^Q, \mathbf{W}_j^K$, and $\mathbf{W}_j^V \in \mathbb{R}^{d \times d}$ are learnable projection matrices. The matrix $\mathbf{E} = [\boldsymbol{e}_1, \cdots, \boldsymbol{e}_J]^\top \in \mathbb{R}^{J \times d}$ contains the comprehensive user embeddings. The residual embedding $\boldsymbol{r}_j$ serves as a correction term, broadcast across all users via $\mathbf{1}_J \in \mathbb{R}^J$, adjusting attention weights to account for hidden confounders. $\mathbf{W}^J$ projects the concatenated attention heads into the output space. $\oplus$ is concatenation. The output $\mathbf{O}$ captures multi-faceted user interactions while accounting for hidden confounders, which are then passed to the next transformer sublayer.

To enhance the model's representational capacity beyond linear transformations, we incorporate a Pointwise Feed Forward Network (PFFN) [134, 205] after the multi-head attention mechanism. The PFFN applies successive transformations with nonlinear activations to generate refined user representations:

$$\mathbf{O}' = \text{LayerNorm}(\mathbf{O} + \text{Dropout}(\phi(\mathbf{O}))) \tag{4.44}$$

$$\mathbf{H} = \text{LeakyReLU}(\mathbf{O}'\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + b_2 \tag{4.45}$$

$$\mathbf{F} = \text{LayerNorm}(\mathbf{O}' + \text{Dropout}(\mathbf{H})) \tag{4.46}$$

where ϕ denotes the PFFN operation and LayerNorm [216] standardizes features to stabilize training. LeakyReLU [215] introduces nonlinearity between transformations. The resulting matrix $\mathbf{F} \in \mathbb{R}^{J \times d}$ contains unbiased user embeddings, with each row vector $\boldsymbol{u}_j \in \mathbb{R}^d$

representing the causally-adjusted unbiased representation for user u_j. This architecture effectively combines the benefits of attention mechanisms with nonlinear transformations while maintaining causal adjustment for addressing hidden confounders.

Building upon the unbiased user embeddings, we develop our third component ***Social Diffusion Predictor*** that optimizes both prediction accuracy and information diversity in social networks. Our approach integrates three crucial metrics to comprehensively evaluate and guide the diffusion process. First is the Mean Absolute Error (MAE), which quantifies the prediction accuracy of diffusion probabilities between user pairs [71]. Second is the Intra-List Diversity (ILD), which measures user variation within diffusion cascades [82, 77]. Third is the Category Coverage (CC), assessing the breadth of content exposure across the network [200, 130]. Through simultaneous optimization of these metrics, we develop a model that balances accurate diffusion prediction with content diversity, establishing a foundation for effective echo chamber intervention strategies.

First, we calculate MAE based on the average absolute difference between predicted and actual diffusion probabilities for each pair of users as

$$\mathscr{L}_m(\boldsymbol{u}_j, \boldsymbol{u}_q) = \frac{1}{|E|} \sum_{(j,q)\in E} |D_{jq} - \widehat{D}_{jq}| \tag{4.47}$$

where $D_{jq} = \frac{|I_{jq}|}{|I_j|}$ represents the true diffusion probability from user u_j to u_q, computed as the ratio of successful transmissions $|I_{jq}|$ to total transmissions $|I_j|$. The predicted probability $\widehat{D}_{jq}$ is calculated as:

$$\widehat{D}_{jq} = \sigma(\boldsymbol{M}^\top(\boldsymbol{u}_j \oplus \boldsymbol{u}_q)) \tag{4.48}$$

where σ is the sigmoid function and $\boldsymbol{M}$ is a learnable weight vector.

Second, we measure ILD based on the average pairwise dissimilarity between users within each diffusion cascade as

$$\mathscr{L}_d(\boldsymbol{u}_j) = \frac{1}{|C|} \sum_{c\in C} \frac{1}{|c|(|c|-1)} \sum_{j,q\in c} (1 - \cos(\boldsymbol{u}_j, \boldsymbol{u}_q)) \tag{4.49}$$

where C is the set of all diffusion cascades and c represents a specific cascade, with $|c|$ denoting the number of users in cascade c. The term $\cos(\boldsymbol{u}_j, \boldsymbol{u}_q)$ computes the cosine similarity between users u_j and u_q, measuring how similar their behavioral patterns are.

Third, we assess Category Coverage (CC) to measure the diversity of content categories that users engage with across the network. For each user u_j, we map his/her unbiased embedding $\boldsymbol{u}_j$ to a probability distribution over k predefined content categories as

$$\text{Categories}(\boldsymbol{u}_j) = \sigma(\mathbf{W}_c \boldsymbol{u}_j + \mathbf{b}_c) \tag{4.50}$$

where $\mathbf{W}_c \in \mathbb{R}^{k \times d}$ is the learnable parameter. The output Categories$(\boldsymbol{u}_j) \in \mathbb{R}^k$ represents the probabilities of user u_j participating in each of the predefined k categories. Following that, the overall category coverage across the whole network is then measured as:

$$\mathscr{L}_c(\boldsymbol{u}_j) = \frac{|\bigcup_{j=1}^{J} \text{Categories}(\boldsymbol{u}_j)|}{k} \tag{4.51}$$

where $\mathscr{L}_c(\boldsymbol{u}_j)$ measures CC as the ratio of unique categories that have user participation to the total number of predefined categories k.

Finally, we integrate causal residual loss, prediction accuracy, and diversity metrics into one composite loss to jointly optimize the model as:

$$\mathscr{L} = \lambda_1 \mathscr{L}_r + \lambda_2 \mathscr{L}_m(\boldsymbol{u}_j, \boldsymbol{u}_q) + \lambda_3(1 - \mathscr{L}_d(\boldsymbol{u}_j)) + \lambda_4(1 - \mathscr{L}_c(\boldsymbol{u}_j)) \tag{4.52}$$

where $\mathscr{L}$ is the composite loss, which allows us to achieve accurate diffusion predictions while addressing hidden confounders, enhancing diversity and mitigating echo chambers.

Leveraging the trained model, our last component ***Targeted Interventions*** implements targeted interventions to mitigate echo chambers while preserving user engagement. Generally, we aim to develop a systematic approach to identify optimal intervention points by analyzing both network topology and information diversity patterns. This allows for strategic adjustments of information flow throughout the network structure before echo chambers are formed. Specifically, our intervention strategy focuses on two key structural elements in the network:

- **Low Diversity Clusters:** User clusters with low Intra-List Diversity (ILD) scores indicate the presence of potential echo chambers, since these groups primarily share and interact with similar content. Mathematically, the clusters with low ILD scores are defined as:

  $$ILD_{low} = \{c \in C | \frac{1}{|c|} \sum_{u_j \in c} \mathscr{L}_d(\boldsymbol{u}_j) < \theta_m\} \tag{4.53}$$

  where ILD_{low} represents clusters with average ILD scores below threshold θ_m, and $|c|$ is the number of users in cluster c.

- **Bottleneck Users:** Users who are connected to multiple communities but showing limited content diversity in their sharing, where Category Coverage (CC) scores are low. Mathematically, bottleneck user is defined as:

  $$U_{botnec} = \{u_j \in \mathscr{U} | N(u_j) \geq \theta_n \wedge \mathscr{L}_c(\boldsymbol{u}_j) < \theta_c\} \tag{4.54}$$

  where θ_n is the neighbor threshold. $N(u_j)$ represents the number of neighboring communities for user u_j and θ_c is the content diversity threshold.

Based on the identified intervention points, we implement two complementary intervention strategies to enhance information diversity. For low diversity clusters, we employ diversity-aware content injection by strategically rewiring connections. This involves removing an input edge from the highest-degree user within each cluster while establishing new connections from external sources, thereby maximizing improvements in ILD scores. Concurrently, we implement cross-category bridging for bottleneck users by establishing strategic output edges to facilitate information flow across disparate communities. This targeted approach enhances Category Coverage scores while maintaining network connectivity. Through these coordinated interventions, we systematically reshape information diffusion patterns to disrupt echo chamber formation while preserving the network's fundamental social structure.

4.3.3 Experiments

4.3.3.1 Datasets

We evaluate our CEDA on three real-world social network datasets: *Twitter*[5], *Google+*[6], and *Facebook*[7]. As shown in Table 4.7, these datasets exhibit diverse network characteristics, enabling comprehensive evaluation across different social environments. The Twitter dataset comprises 973 social circles with 81,306 users and a moderate clustering coefficient of 0.5653. Google+ contains 132 circles encompassing 107,614 users with relatively sparse connections (clustering coefficient: 0.4901), while Facebook features 10 densely connected circles of 4,039 users with a higher clustering coefficient of 0.6055. To ensure the data quality, we filter interaction sequences shorter than 5 interactions and partition the remaining data into training (70%), validation (10%), and test (20%) sets. This preprocessing ensures meaningful evaluation of both short-term interactions and long-range dependencies in information diffusion patterns.

Table 4.7: CEDA: Statistical details of the three datasets.

Statistics	**Twitter**	**Google+**	**Facebook**
#Networks	973	132	10
#Users	81,306	107,614	4,039
Avg. CC	0.5653	0.4901	0.6055
Diameter	7	6	8

[5]https://snap.stanford.edu/data/ego-Twitter.html
[6]https://snap.stanford.edu/data/ego-Gplus.html
[7]https://snap.stanford.edu/data/ego-Facebook.html

4.3.3.2 Baselines and Evaluation

We implement our CEDA framework using PyTorch, setting the user representation dimension to 128 and optimizing with the Adam optimizer [87] (learning rate 0.001, batch size 256). The model employs a Causal Transformer with 16 attention heads and a 0.1 dropout rate to prevent overfitting. Loss function weights λ_1, λ_2, and λ_3 are fine-tuned through grid search in [0.1, 1.0]. To evaluate CEDA's performance, we use five complementary metrics: RMSE [71] for prediction accuracy, Precision@K and Recall@K [232, 38] for top-K recommendation quality, Gini coefficient [54] for measuring exposure fairness, and Simpson's Diversity Index (SDI) [151, 254] for user-side exposure diversity. While CEDA's core focus is on mitigating echo chambers by restructuring user-user diffusion paths, it directly benefits recommendation by generating more accurate and diverse user-item predictions through causal graph adjustments. For a fair comparison, we adapt baselines originally designed for echo chamber detection or diversity promotion by integrating their outputs—such as modified user graphs or exposure scores—into our sequential recommendation pipeline, converting them into user-item ranking scores based on aligned diffusion patterns. All baseline models are tuned using the same parameter ranges to ensure equitable evaluation.

We compare our CEDA with below state-of-the-art methods:

- **FRECH** [165] implements GCN architecture for learning implicit echo chamber patterns and recommending diverse connections.
- **CECD** [119] utilizes probabilistic modeling to analyze echo chamber formation and information flow.
- **OCR** [172] employs quadratic optimization to balance content diversity with user preferences.
- **ECS** [7] quantifies echo chamber effects through user embedding distances in the social network.
- **ECM** [145] models echo chamber dynamics using agent-based simulations in online social networks.
- **GRU4Rec** [70] leverages GRU networks for modeling sequential user behaviors.
- **NARM** [96] combines attention mechanisms with RNNs for comprehensive behavior modeling.

- **STAMP** [110] integrates both the short-term and long-term interest modeling through attention mechanisms.
- **LLMS** [62] utilizes large language models for semantic relationship enhancement to improve sequential recommendations.
- **ReFor** [97] uses a transformer-based model to learn language representations for sequential recommendation.
- **DCCF** [219] implements counterfactual reasoning with back-door adjustment for echo chamber mitigation.

4.3.3.3 Result Analysis

Performance Comparison: To comprehensively evaluate CEDA's effectiveness, we conduct extensive experiments on three real-world datasets. As shown in Table 4.8, CEDA outperforms all the baselines on all datasets. For example, CEDA achieves significant improvements in RMSE, reducing error rates by 13.38%, 10.05%, and 12.44% on Twitter, Google+ and Facebook datasets respectively, compared to the strongest baselines. The performance advantage extends to ranking metrics, with notable improvements in Precision@40 (Twitter: 14.22%, Google+: 5.99%, Facebook: 7.58%) and Recall@40 (Twitter: 11.65%, Google+: 2.09%, Facebook: 2.63%). Furthermore, CEDA demonstrates substantial effectiveness in mitigating echo chambers, as measured through the Gini coefficient and Simpson's Diversity Index (SDI). The model achieves consistent reductions in Gini coefficients (Twitter: 12.23%, Google+: 5.44%, Facebook: 10.36%), indicating more balanced information distribution. Concurrently, SDI improvements (Twitter: 23.63%, Google+: 11.23%, Facebook: 12.89%) reflect enhanced content diversity exposure across user groups. These results validate CEDA's capability to simultaneously improve prediction accuracy while effectively addressing echo chamber formation across diverse social network environments.

Ablation Study: To explore the individual contribution of each component in CEDA, we conduct ablation studies as shown in Figure 4.15. Through systematic component replacement, we examine three key variations: DM (replacing User Dual Modelling with attribute-based embedding), CT (bypassing Causal Transformer), and TI (substituting Targeted Interventions with random forest prediction), comparing against the complete CEDA model (ALL). The results reveal the Causal Transformer's paramount importance, with its removal causing substantial performance degradation in Precision@40 (Twitter: 36.76%, Facebook: 22.66%, Google+: 7.48%). The Targeted Interventions component demonstrates platform-dependent impact, showing significant effects on Twitter (29.20% decrease) and Facebook

Table 4.8: CEDA: Recommendation performance comparison: the best results are bolded, while the best baselines are underlined.

Dataset	Metric	GRU4Rec	NARM	STAMP	LLMS	ReFor	CECD	OCR	FRECH	DCCF	ECS	CEDA	Improv.%
	RMSE	0.3621	0.3598	0.3567	0.3512	0.3489	0.3556	0.4102	0.3645	0.3599	0.3532	**0.3022**	13.38%
	Pr@5	0.5312	0.5389	0.5456	0.5634	0.5601	0.5215	0.5301	0.5589	0.5625	0.5578	**0.6015**	6.76%
	Pr@10	0.6089	0.6156	0.6223	0.6301	0.6378	0.5987	0.6056	0.6312	0.6378	0.6467	**0.6898**	6.67%
	Pr@20	0.6756	0.6823	0.6901	0.6978	0.7056	0.6654	0.6721	0.6978	0.7045	0.7156	**0.7689**	7.45%
	Pr@40	0.7412	0.7489	0.7567	0.7645	0.7723	0.7312	0.7389	0.7601	0.7878	0.7745	**0.8998**	14.22%
Twitter	Re@5	0.4956	0.5023	0.5101	0.5178	0.5356	0.4856	0.4923	0.5145	0.5201	0.5289	**0.5789**	8.08%
	Re@10	0.5678	0.5745	0.5823	0.5901	0.5978	0.5578	0.5645	0.5889	0.5956	0.6045	**0.6587**	8.97%
	Re@20	0.6345	0.6412	0.6489	0.6567	0.6745	0.6245	0.6312	0.6567	0.6634	0.6723	**0.7378**	9.38%
	Re@40	0.7301	0.7378	0.7456	0.7634	0.7612	0.7401	0.6978	0.7212	0.7289	0.7367	**0.8523**	11.65%
	Gini	0.5923	0.5856	0.5789	0.5723	0.5656	0.6123	0.5987	0.5678	0.5534	0.5456	**0.4789**	12.23%
	SDI	0.4122	0.4044	0.3966	0.3877	0.3799	0.4322	0.4211	0.3844	0.3511	0.3622	**0.2766**	23.63%
	RMSE	0.3456	0.3423	0.3389	0.3356	0.3323	0.3401	0.3956	0.3489	0.3423	0.3378	**0.2989**	10.05%
	Pr@5	0.5445	0.5523	0.5601	0.5778	0.5756	0.5345	0.5432	0.5712	0.5789	0.5723	**0.6137**	6.21%
	Pr@10	0.6223	0.6301	0.6378	0.6456	0.6534	0.6123	0.6201	0.6456	0.6523	0.6578	**0.6964**	5.87%
	Pr@20	0.6889	0.6967	0.7045	0.7123	0.7201	0.6789	0.6867	0.7123	0.7189	0.7245	**0.7701**	6.29%
	Pr@40	0.7556	0.7634	0.7712	0.7789	0.7867	0.7456	0.7534	0.7745	0.8023	0.7889	**0.8504**	5.99%
Google+	Re@5	0.5089	0.5167	0.5245	0.5323	0.5501	0.4989	0.5056	0.5289	0.5345	0.5434	**0.5751**	4.54%
	Re@10	0.5812	0.5889	0.5967	0.6045	0.6123	0.5712	0.5789	0.6023	0.6101	0.6201	**0.6559**	5.77%
	Re@20	0.6489	0.6567	0.6645	0.6823	0.6801	0.6389	0.6456	0.6712	0.6789	0.6778	**0.7274**	6.61%
	Re@40	0.7445	0.7523	0.7601	0.7678	0.7756	0.7545	0.7123	0.7356	0.7434	0.7512	**0.7918**	2.09%
	Gini	0.5789	0.5723	0.5656	0.5589	0.5523	0.5989	0.5845	0.5534	0.5389	0.5312	**0.5023**	5.44%
	SDI	0.4088	0.3999	0.3911	0.3822	0.3733	0.4188	0.4077	0.3699	0.3366	0.3477	**0.2988**	11.23%
	RMSE	0.3689	0.3656	0.3623	0.3489	0.3556	0.3589	0.4234	0.3767	0.3701	0.3534	**0.3055**	12.44%
	Pr@5	0.5201	0.5278	0.5356	0.5634	0.5512	0.5101	0.5189	0.5467	0.5534	0.5489	**0.5957**	5.73%
	Pr@10	0.5967	0.6045	0.6123	0.6201	0.6378	0.5867	0.5945	0.6201	0.6267	0.6323	**0.6834**	7.15%
	Pr@20	0.6634	0.6712	0.6789	0.6867	0.6945	0.6534	0.6601	0.6856	0.6923	0.7001	**0.7578**	8.24%
	Pr@40	0.7289	0.7367	0.7445	0.7523	0.7601	0.7189	0.7267	0.7478	0.7545	0.7612	**0.8189**	7.58%
Facebook	Re@5	0.4845	0.4923	0.5001	0.5078	0.5256	0.4745	0.4812	0.5034	0.5089	0.5167	**0.5567**	5.92%
	Re@10	0.5567	0.5645	0.5723	0.5801	0.5878	0.5467	0.5534	0.5778	0.5845	0.5912	**0.6395**	8.17%
	Re@20	0.6234	0.6312	0.6389	0.6667	0.6545	0.6134	0.6201	0.6456	0.6523	0.6589	**0.7134**	7.00%
	Re@40	0.7189	0.7267	0.7345	0.7423	0.7601	0.7289	0.6867	0.7101	0.7278	0.7256	**0.7801**	2.63%
	Gini	0.6034	0.5967	0.5901	0.5834	0.5767	0.6234	0.6101	0.5789	0.5845	0.5867	**0.5189**	10.36%
	SDI	0.4333	0.4244	0.4155	0.4066	0.3977	0.4433	0.4322	0.3955	0.3733	0.3622	**0.3155**	12.89%

(15.33% decrease), while exhibiting modest influence on Google+ (2.59% decrease). User Dual Modelling similarly shows varying importance across platforms (Twitter: 19.53%, Google+: 24.26%, Facebook: 8.44%). Notably, the Causal Transformer proves essential for echo chamber mitigation, as evidenced by increased Gini coefficients upon its removal (Twitter: 9.0%, Google+: 5.44%, Facebook: 6.0%). These findings underscore the critical role of addressing hidden confounders through causal inference in maintaining effective echo chamber mitigation across diverse social network environments.

Parameters Analysis: To assess CEDA's sensitivity for various parameter configurations and optimize its performance, we conduct a comprehensive hyperparameter analysis on three real-world datasets. Generally, we focus on four parameters: first is the user embedding dimensions d, impacting the depth of user preference insights; second is the number of attention heads j, impacting the model's capacity to process multiple relevance signals; third is the batch size, impacting learning stability; and last is the learning rate, impacting convergence speed. As shown in Figure 4.16, Figure 4.17, and Figure 4.18, our results

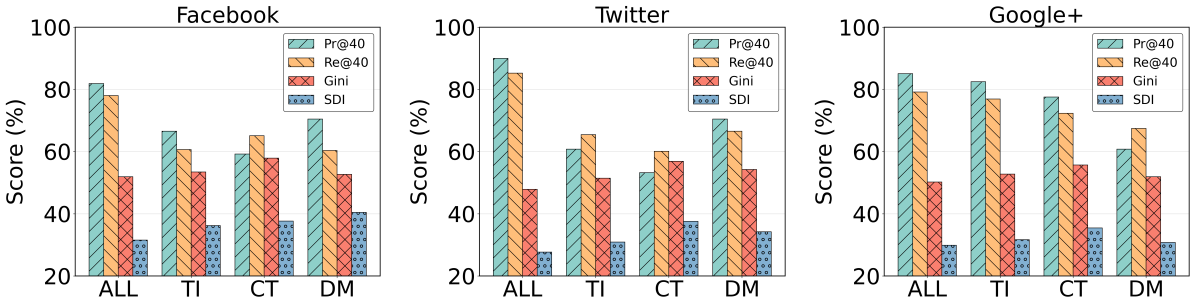


Figure 4.15: CEDA: Ablation study.

reveal distinct sensitivity patterns across different social network environments, emphasizing the importance of platform-specific optimization. Specifically, the optimal attention head configuration varies with network density. Twitter achieves peak performance (Precision@40: 89.98%) with 16 heads, while Google+ and Facebook perform optimally with 8 heads (85.04% and 81.89% respectively). Similarly, embedding dimensionality requirements correlate with network complexity, with Twitter and Facebook performing best at 128 dimensions, while Google+ requires 256 dimensions for optimal performance. Regarding training parameters, batch size optimization shows platform dependence, with Twitter and Facebook achieving optimal results at 256, while Google+ benefits from larger batches of 512. However, the learning rate demonstrates consistent behavior across platforms, with 0.001 yielding optimal performance universally. These findings highlight the necessity of platform specific hyperparameter tuning for effective echo chamber mitigation, particularly for parameters affecting model capacity and training dynamics.

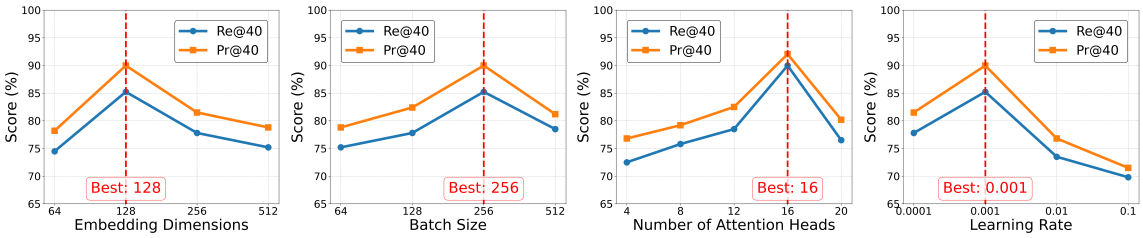


Figure 4.16: CDEA: Parameter sensitivity analysis on the Twitter dataset.

Understanding Intervention Strategies: We evaluate our intervention strategies' effectiveness through a detailed analysis of representative networks constructed from 20 randomly selected users on each platform. Initial network snapshots (Before) reveal distinct echo chamber formations: Twitter exhibited three chambers, Google+ contained four chambers, and Facebook displayed five chambers. Our method identifies optimal modification

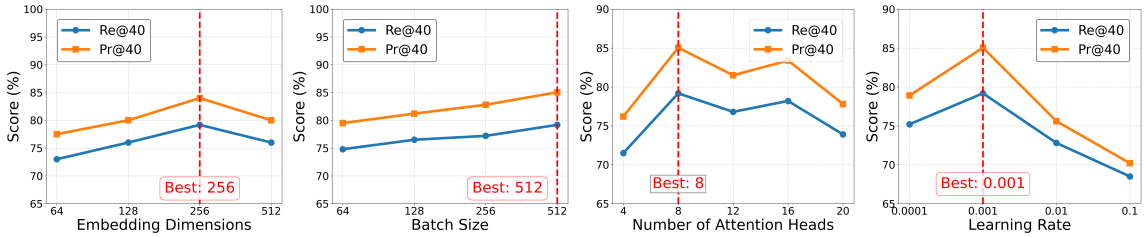


Figure 4.17: CDEA: Parameter sensitivity analysis on the Google+ dataset.

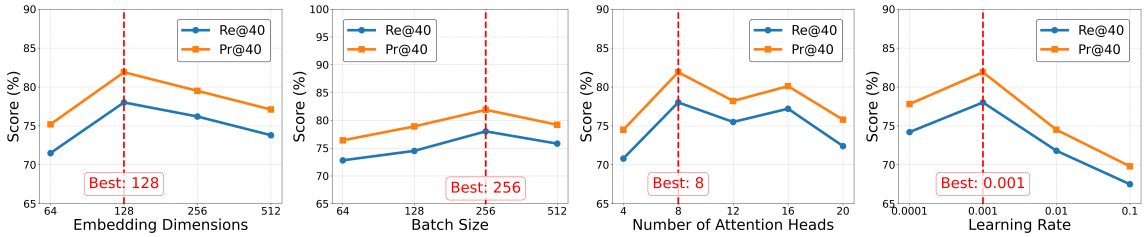


Figure 4.18: CDEA: Parameter sensitivity analysis on the Facebook dataset.

points through two mechanisms: detection of low diversity clusters using ILD scores and identification of bottleneck users based on their network positioning. We implement targeted modifications by strategically removing connections from high degree users within low diversity clusters (red prohibition signs) while establishing new cross cluster pathways through bottleneck users (purple nodes with purple connecting lines). Post intervention analysis (After) demonstrates significant structural improvements. The number of echo chambers decreased substantially: Twitter reduced from three to one, Google+ from four to two, and Facebook from four to zero. The newly established purple bridging connections successfully transform previously segregated communities into interconnected networks. These results validate CEDA's ability to leverage causal learning for identifying and addressing hidden confounders, enabling systematic echo chamber mitigation while promoting diverse information flow across social networks.

4.3.3.4 Summary

In this work, we propose CEDA, a novel causal framework that addresses RQ2 regarding enhancing recommender system robustness against spurious correlations in complex social network structures. By synergistically integrating causal inference with transformer architecture, CEDA effectively identifies and adjusts for hidden confounders that create spurious correlations in information diffusion patterns, enabling more accurate prediction of user behavior while actively mitigating echo chamber formation. Extensive experiments

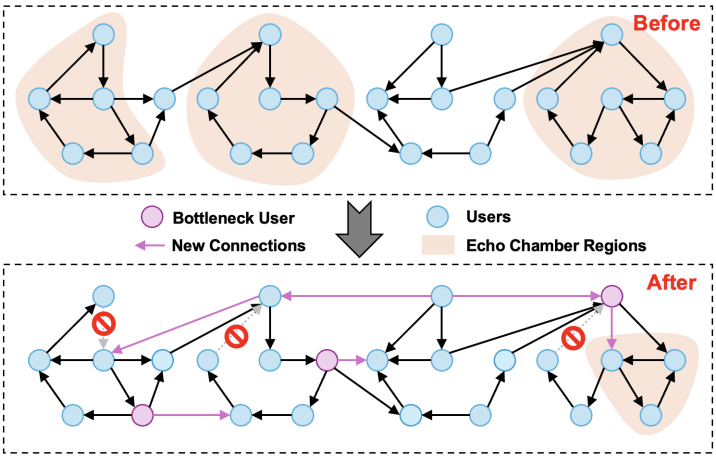


Figure 4.19: CDEA: Visualization echo chamber mitigation effects on Twitter.

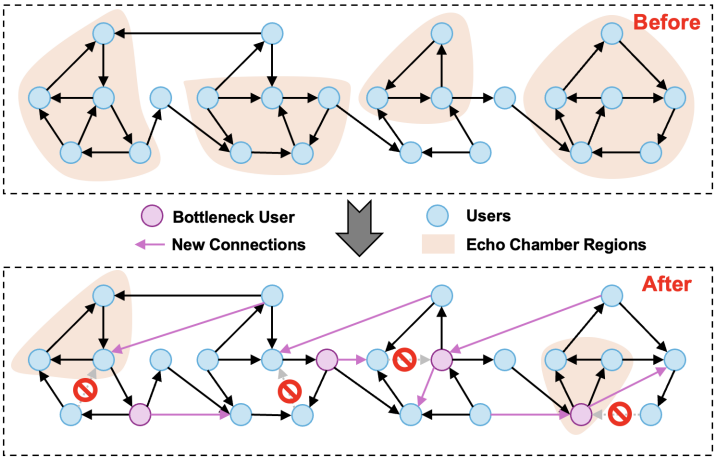


Figure 4.20: CDEA: Visualization of echo chamber mitigation effects on Google+.

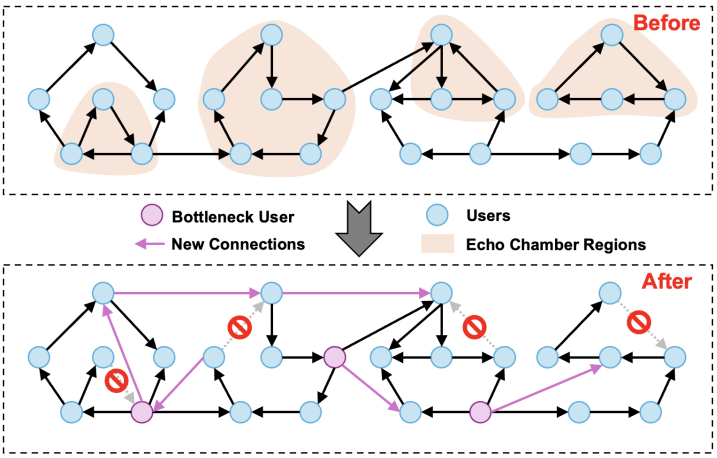


Figure 4.21: CDEA: Visualization of echo chamber mitigation effects on Facebook.

demonstrate CEDA's superior performance over state-of-the-art baselines in terms of both prediction accuracy and echo chamber mitigation across diverse social network environments. Future work will explore extending our causal framework to handle multimodal social network data, potentially uncovering additional confounding mechanisms that influence information diffusion patterns in broader social contexts.

CHAPTER 5

CAUSAL MODEL FOR EXPLAINABLE RECOMMENDATIONS

This chapter addresses RQ3 regarding how causal insights can improve the recommendation interpretability. By designing two causal models, we demonstrate how causal reasoning can effectively address the problem of spurious correlation to improve the interpretability of recommendation models. The first study introduces semantics-guided disentangled learning, leveraging causal structures to identify interpretable factors that drive recommendations, thereby generating transparent rationales for system decisions. Building upon this foundation, the second study employs counterfactual reasoning in conversational recommendations to not only improve the accuracy but also provide clear explanations by identifying minimal changes that would alter recommendation outcomes. Through comprehensive methodology, rigorous experimentation, and detailed findings, this chapter establishes the effectiveness of causal inference in creating recommender systems that are both accurate and interpretable to users.

5.1 Semantics-Guided Disentangled Learning for Recommendation

5.1.1 Overview

Traditional recommender systems mainly rely on historical interaction data to learn user preferences, which often fails to distinguish between true user interests and spurious correlations in the data [231, 183, 10]. While recent disentangled learning methods attempt to

address this by separating different behavioral factors, they primarily focus on numerical patterns without considering rich semantics that could help explain the true underlying causes of user preferences [101, 185, 253]. This limitation can lead to suboptimal recommendations which are difficult to explain and justify. For example, when a user interacts with multiple movies, conventional methods may struggle to identify whether these interactions stem from true interest in specific genres/directors or from conformity bias due to social influences. Without proper semantic disentanglement, the system cannot provide meaningful explanations for its recommendations.

5.1.1.1 Research Objective

This study aims to address RQ3 regarding how causal insights can improve recommendation explainability by developing a novel semantic-guided disentangled learning framework. Specifically, we propose to leverage heterogeneous information networks (HIN) to untangle the complex relationships between users, items and their attributes to reveal the true causal factors driving user preferences. Through this semantic disentanglement, we aim to not only improve recommendation accuracy but also generate interpretable explanations by identifying which semantic aspects (e.g., movie genres, directors) causally influence each user's decisions.

5.1.1.2 The Proposed Method

To achieve the aforementioned objectives, we propose a novel framework called **Se**mantics-guided **D**isentangled **L**earning for **R**ecommendation (**SeDLR**) with three synergistic components:

- ***Graph Disentangling*** decomposes the user-item interaction graph into multiple intent-aware subgraphs based on semantic aspects extracted from the HIN, capturing different factors influencing user preferences independently.
- ***Semantic-aware Intent Representation*** leverages meta-path schemes in the HIN to learn enriched semantic embeddings that ground interaction patterns in interpretable semantic concepts for generating meaningful explanations.
- ***Monte Carlo Edge-drop*** identifies causally significant semantic aspects by systematically dropping edges in the HIN structure and observing their impacts on recommendations to generate precise explanations.

The key contributions of this research are summarized as follows:

- We propose a novel framework called SeDLR that integrates HINs with disentangled representation learning to jointly address the challenges of separating spurious correlations and providing semantic explanations in RSs.
- We develop a principled approach that leverages meta-path schemes to ground disentangled user intents in interpretable semantic concepts, enabling meaningful explanations for recommendations.
- We design an innovative Monte Carlo edge-drop strategy that can identify causally significant semantic aspects through systematic edge pruning, providing precise explanations for why specific items are recommended.

5.1.2 SeDLR

5.1.2.1 Problem Definition

Let $\mathscr{U}$ and $\mathscr{I}$ be the sets of users and items, respectively. Each user $u \in \mathscr{U}$ and item $i \in \mathscr{I}$ associated with multiple attributes $\mathscr{A} = \{a_1, a_2, ..., a_K\}$ extracted from the heterogeneous information networks (HINs). Each attribute $a_k \in \mathscr{A}$ represents interpretable properties like movie genres and directors. The historical user-item interactions are denoted as $\mathscr{O} = \{(u,i)|u \in \mathscr{U}, i \in \mathscr{I}\}$, where each tuple (u,i) indicates user u has interacted with item i.

Our goal is to accurately predict user preferences while providing semantic explanations by disentangling true user interests from spurious correlations. Mathematically, we decompose the holistic user-item interaction graph $\mathscr{G} = (\mathscr{U} \cup \mathscr{I}, \mathscr{O}, \mathscr{A})$ into multiple intent-aware subgraphs $\{\mathscr{G}_1, ..., \mathscr{G}_Q\}$, where each subgraph $\mathscr{G}_q$ captures semantically meaningful interaction patterns related to specific attributes. Through meta-paths $\mathscr{P} = \{p_1, ..., p_M\}$ in the HIN that connects users and items via different attribute combinations, we learn interpretable semantic embeddings $\mathbf{e}_q$ for each subgraph. For a target user-item pair (u,i), we generate both a prediction score $\hat{y}_{ui}$ indicating the likelihood of interaction and a set of semantic explanations $\mathscr{E} = \{a_k|a_k \in \mathscr{A}\}$ highlighting which attributes causally influenced the recommendation. The key challenges lie in effectively leveraging rich semantics to identify and separate spurious correlations from true causal relationships in user behavior patterns, while grounding disentangled representations in interpretable semantic concepts that enable meaningful explanations. By addressing these challenges through our proposed SeDLR, we aim to achieve more robust and explainable recommendations.

5.1.2.2 Methodology

Figure 5.1 shows the overall framework of our proposed SeDLR. To begin with, we decompose the holistic user-item interaction graph into multiple intent-aware subgraphs. For each user u and item i, we divide their embeddings into q chunks representing different intents:

$$\boldsymbol{u} = \left(\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_q\right), \quad \boldsymbol{i} = \left(\boldsymbol{i}_1, \boldsymbol{i}_2, \ldots, \boldsymbol{i}_q\right) \tag{5.1}$$

where $\boldsymbol{u}_q$ and $\boldsymbol{i}_q$ are the chunked representation for q-th intent on interaction of user/item. To model relationships between intents and interactions, we construct a score vector $S(u,i) = \left(S_1(u,i), S_2(u,i), \cdots, S_q(u,i)\right)$, where each element represents the likelihood of an interaction being driven by that intent. These score vectors effectively form adjacency matrices for intent-aware subgraphs. We then employ a graph disentangling layer that propagates information through these subgraphs using neighbor aggregation:

$$e_q^{u(1)} = g(\boldsymbol{u}_q, \{\boldsymbol{i}_q \mid i \in \mathcal{N}_u\}) \tag{5.2}$$

where $e_q^{u(1)}$ aggregates first-order neighbor information and $\mathcal{N}_u$ represents historically interacted items. Through iterative updates, the model refines both intent-aware embeddings and graph structures. Next, to ground these disentangled intents in interpretable concepts, we leverage meta-paths in the HIN to learn enriched semantic embeddings:

$$\tilde{S}_q^n(u,i) = \frac{\exp S_q^n(u,i)}{\sum_{q'=1}^{q} \exp S_{q'}^n(u,i)} \tag{5.3}$$

where $\tilde{S}_q^n(u,i)$ represents the normalized importance score for the q-th intent between user u and item i at iteration n. $S_q^n(u,i)$ denotes the raw score for the q-th intent. The denominator sums over all q intents to normalize the scores through softmax, ensuring they form a valid probability distribution. These scores enable construction of the Laplacian matrix $\mathcal{M}_q^n$ that guides information propagation while preserving semantic meaning:

$$\mathcal{M}_q^n(u,i) = \frac{\tilde{S}_q^n(u,i)}{\sqrt{D_q^n(u) \cdot D_q^n(i)}} \tag{5.4}$$

where D_q^n represents node degrees. Following that, the final embedding propagation aggregates semantically-enriched information across the graph:

$$\boldsymbol{u}_q^n = \sum_{i \in N_u} \mathcal{M}_q^n(u,i) \cdot \boldsymbol{i}_q^0 \tag{5.5}$$

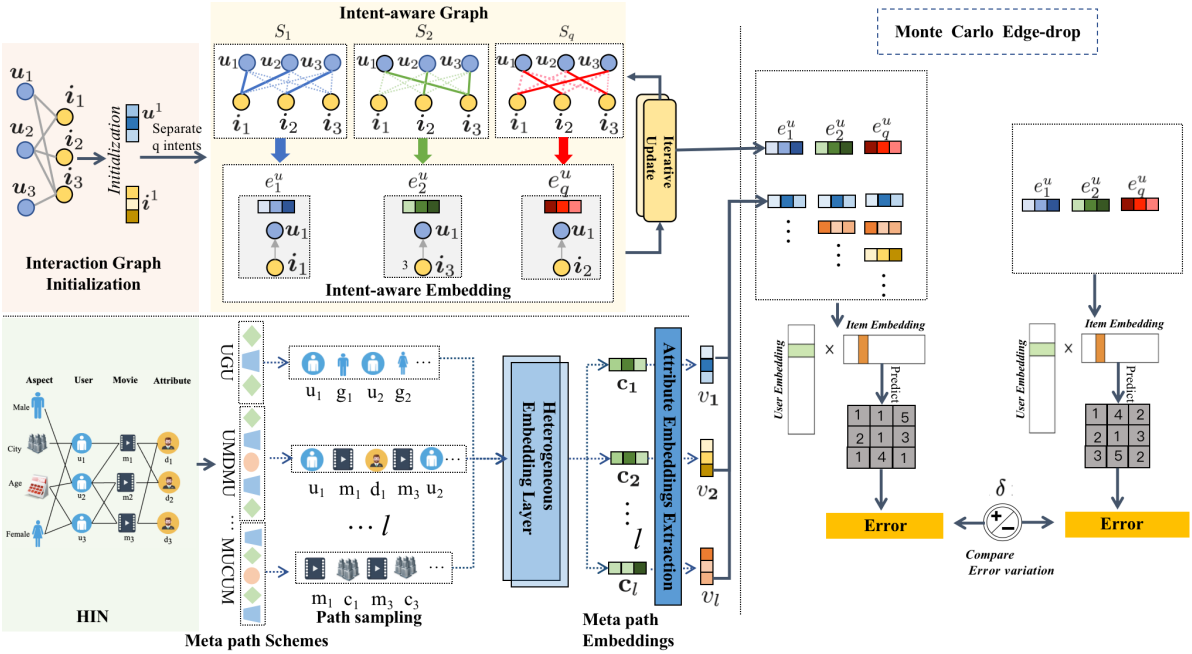


Figure 5.1: The overall framework of our proposed method SeDLR.

where $\boldsymbol{u}_q^n$ represents the user embedding for the q-th intent at iteration n. $\boldsymbol{i}_q^0$ is the initial item embedding for intent q. In this way, we compute a weighted sum of item embeddings based on their semantic relevance to capture user preferences for each intent.

Thereafter, we iteratively update the intent-aware graph by adjusting interaction scores between users and items:

$$S_q^{n+1}(u,i) = S_q^n(u,i) + \boldsymbol{u}_q^{n\top} \tanh\left(\boldsymbol{i}_q^0\right) \tag{5.6}$$

where $S_q^{n+1}(u,i)$ represents the updated score for intent q between user u and item i. The term $\boldsymbol{u}_q^{n\top} \tanh\left(\boldsymbol{i}q^0\right)$ measures their semantic affinity using the nonlinear tanh activation. This iterative process strengthens connections between items driven by similar intents, ultimately producing disentangled representations $e_q^{u(1)} = \boldsymbol{u}_q^n$ and the refined intent-aware graph $\tilde{S}_q^n$. We then stack multiple layers to capture rich semantics from high-order connectivity. For each layer r, we aggregate information from previous layers as:

$$e_q^{u(r)} = g\left(e_q^{u(r-1)}, \left\{e_q^{i\,(r-1)} \mid i \in \mathcal{N}_u\right\}\right) \tag{5.7}$$

where $e_q^{u(r-1)}$ represents the propagated information from $(r$-1)-hop neighbors for user u on intent q. Each layer maintains its intent-aware adjacency matrix $\tilde{S}_q^r$. The final representations are obtained by concatenating across all layers and intents: $e^u = (e1^u, \cdots, e_q^u)$ for users and $e^i = (e1^i, \cdots, e_q^i)$ for items, where $e_q^u = (e_q^{u(0)}, e_q^{u(1)}, \cdots, e_q^{u(r)})$. Following that, we

leverage meta-paths in the HINs to extract rich semantic aspects for refining intent representations. A meta-path scheme defines a composite path connecting different node types through various relationships, capturing complex higher-level semantics. For instance, a User-Movie-User (UMU) path $\mathrm{U}_{u_1}-\mathrm{M}_{m_1}-\mathrm{U}_{u_2}$ reveals behavioral similarities between users u_1 and u_2, where u_2's preferences may influence u_1's intents. By incorporating these semantic aspects modeled from meta-paths, we can better understand and explain the true causal factors driving user preferences. Formally, we generate path instances $\rho = \{u_1, u_2, \cdots, u_l\}$ for each pre-defined meta-path $\mathbf{p}$ based random walks. These paths capture both semantic and structural relationships between different node types. We then learn embeddings for these path instances using a CNN followed by max-pooling:

$$\mathbf{c}_{\mathbf{p}} = \text{max-pooling}\left(\left\{CNN\left(\{\mathbf{X}_i^{\rho}\};\Theta\right)\right\}_{i=1}^{L}\right) \tag{5.8}$$

where $\mathbf{X}_i^{\rho}$ is the embeddings for L path instances of meta-path $\mathbf{p}$. From these meta-path embeddings, we derive semantic representations for each user u through embedding lookup:

$$v_{\mathbf{p}} = \mathbf{c}_{\mathbf{p}}^{\top} \cdot \boldsymbol{u} \tag{5.9}$$

where $\boldsymbol{u}$ is the one-hot encoding of user u. The resulting $v_{\mathbf{p}}$ serves as the aspect embedding for user u under meta-path $\mathbf{p}$, capturing rich semantic context to guide intent learning. Following that, we perform semantic-aware intent learning by incorporating the semantic embeddings $v_{\mathbf{p}}$ with intent representations e^u and e^i using a Factorization Machine (FM):

$$\boldsymbol{h}_{\mathbf{p}} = e^u \odot v_{\mathbf{p}} \tag{5.10}$$

where $\boldsymbol{h}_{\mathbf{p}}$ represents the semantic-enriched intent representation and $\odot$ denotes element-wise product. We then generate the final prediction score by combining user-item interactions with semantic intents:

$$\hat{\boldsymbol{y}}_{ui} = \alpha \boldsymbol{u}^{\top} \boldsymbol{i} + (1-\alpha)\boldsymbol{h}_{\mathbf{p}}^{\top} \boldsymbol{i} \tag{5.11}$$

where α balances the contribution of each component. The model parameters are optimized using the Bayesian Personalized Ranking (BPR) loss:

$$\mathscr{L}_{\text{BPR}} = \sum_{u,i,j \in \mathscr{D}} -\ln \sigma\left(\hat{\boldsymbol{y}}_{ui} - \hat{\boldsymbol{y}}_{uj}\right) + \lambda \|E\|_2^2 \tag{5.12}$$

where $\mathscr{D}$ contains user-item interaction triples and E denotes the embedding matrices. This unified approach enables SeDLR to learn disentangled yet semantically meaningful representations for explainable recommendations.

Table 5.1: SeDLR: Statistical details of the two datasets.

Dataset (Density)	Node	Relation (A-B)	Avg. Degree of A/B
Walmart Recruit (0.11%)	User (U): 5,647	U-G: 5,645	U/G: 1 / 2822.5
	Gender (G): 2	U-C: 5,645	U/C: 1 / 564.5
	City (C): 10	U-T: 23,053	U/T: 4.1 / 1.1
	Transaction (T): 20,878	U-U: 0	U/U: 0 / 0
	Category Type (CT): 5	T-A: 23,053	T/A: 1.1 / 4.0
	Amount (A): 5,764	T-CT: 23,053	T/CT: 1.1 / 4610.6
Douban Book (0.27%)	User (U): 13,024	U-Bo: 792,062	U/Bo: 60.8 / 35.4
	Book (Bo): 22,347	U-U: 169,150	U/U: 13.0 / 13.0
	Group (Gr): 2,936	U-Gr: 1,189,271	U/Gr: 91.3 / 405.1
	Author (Au): 10,805	Bo-Au: 21,907	Bo/Au: 1.0 / 2.0
	Publisher (P): 1,815	Bo-P: 21,773	Bo/P: 1.0 / 12.0
	Year (Y): 64	Bo-Y: 21,192	Bo/Y: 1.0 / 331.1

Finally, we propose a Monte Carlo edge-drop strategy to explain recommendations by quantifying the causal impact of different semantic aspects. After obtaining the prediction model $f(\cdot)$, we systematically remove each edge from meta-path $\mathbf{p}$ and generate new predictions $\hat{\boldsymbol{y}}^s_{ui}$. The importance of attribute b is determined by comparing the prediction difference $|\hat{\boldsymbol{y}}^s_{ui} - \hat{\boldsymbol{y}}_{ui}|$ against a threshold δ. If this difference exceeds δ, we identify b as a causally significant aspect that substantially influences the recommendation, providing interpretable explanations for the model's decisions.

5.1.3 Experiments

5.1.3.1 Datasets

We evaluate our proposed SeDLR on two real-world datasets: *Walmart Recruit*[1], and *Douban Book*[2]. As shown in Table 5.1, Walmart Recruit contains retail transactions from 2011 to 2013 with rich HIN context including price, discount, user demographics, and item categories. Douban-Book comprises comprehensive book ratings with three user attributes and four book attributes. For both datasets, we binarize feedback by treating ratings higher than 5 as positive interactions while randomly sampling unobserved items as negative instances to ensure balanced training.

[1]https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting

[2]https://github.com/librahu/HIN-Datasets-for-Recommendation-and-Network-Embedding/tree/master/Douban%20Book

5.1.3.2 Baselines and Evaluation

We implement our model on a Linux server with RTX3070 GPU. For training, we split both datasets into 80%/10%/10% proportions for train/validation/test. The embedding dimension is selected from {16,32,64,128} using Xavier initialization, with learning rate tuned in {0.001,0.01,0.05,0.1}. We train for a maximum 1000 epochs with early stopping. The default parameters for SeDLR are: embedding size 128, iteration number n=3, intent number q=4, and learning rate 0.01. We evaluate SeDLR using three popular metrics Recall@K, NDCG@K, and Precision@K where $K \in \{1, 10, 20, 40\}$. To benchmark our model, we compare it with six representative baselines

- **IF-BPR [235]**: A meta-path-based recommendation model that uses adaptive weighting on heterogeneous information networks (HINs) and optimizes a Bayesian personalized ranking objective.
- **MCRec [72]**: A multi-channel heterogeneous graph model that captures different semantic relationships in HINs through attention-based channel aggregation.
- **NeuMF [66]**: Neural Matrix Factorization combines generalized matrix factorization and multi-layer perceptrons to model complex nonlinear user-item interactions.
- **NGCF [183]**: Neural Graph Collaborative Filtering propagates embeddings through the user-item graph, capturing higher-order collaborative signals.
- **DGCF [185]**: Disentangled Graph Collaborative Filtering learns multiple independent factors from interaction graphs to improve recommendation diversity and interpretability.
- **M-VAE [115]**: A multi-aspect variational autoencoder that disentangles user preferences into independent latent components for explainable recommendation.

These baselines offer diverse modeling capabilities in semantic modeling, graph propagation, and latent disentanglement, providing a comprehensive benchmark for evaluating the effectiveness of SeDLR.

5.1.3.3 Result Analysis

Performance Comparison: To verify the effectiveness of our proposed SeDLR, we compare it with many state-of-the-art models on Top-K recommendations. Table 5.1.3.3 demonstrates SeDLR's superior performance across both two datasets. For instance, SeDLR achieves

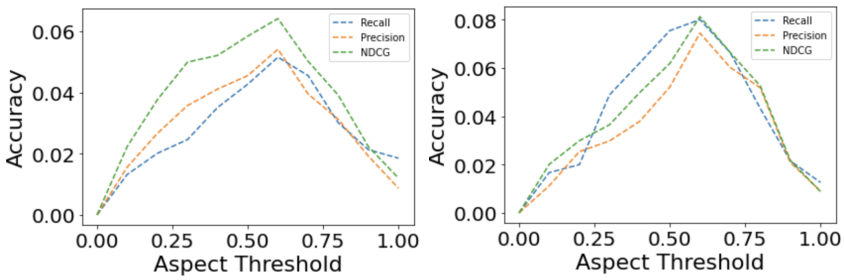
significant improvements in NDCG@20, outperforming the strongest baselines by 27.7% on Walmart Recruit and 15.2% on Douban Book. The consistent improvements over 10% validate the effectiveness of our Monte Carlo edge-drop strategy in identifying causally significant aspects. Moreover, the strong performance of disentangled methods compared to other approaches confirms the benefits of separating distinct user intents for more accurate recommendations.

Table 5.2: SeDLR: Recommendation performance comparison: the best results are marked as bold, strongest baselines are marked with underline.

Datasets	**Metrics**	**NeuMF**	**NGCF**	**DGCF**	**M-VAE**	**IF-BPR**	**MCRec**	**SeDLR**	**Improv.**
Walmart Recruit	Recall@1	0.0376	0.0299	0.0421	0.0391	0.0385	0.0381	**0.0476**	13.1%
	Recall@10	0.0401	0.0387	0.0447	0.0472	0.0419	0.0437	**0.0512**	8.5%
	Recall@20	0.0451	0.0430	0.0516	0.0509	0.0479	0.0448	**0.0552**	7.0%
	Recall@40	0.0612	0.0582	0.0572	0.0519	0.0556	0.0622	**0.0672**	8.0%
	Precision@1	0.0301	0.0315	0.0357	0.0322	0.0316	0.0351	**0.0417**	16.8%
	Precision@10	0.0457	0.0385	0.0477	0.0369	0.0399	0.0426	**0.0516**	8.2%
	Precision@20	0.0528	0.0497	0.0519	0.0489	0.0462	0.0512	**0.0556**	5.3%
	Precision@40	0.0609	0.0599	0.0712	0.0603	0.0591	0.0621	**0.0776**	9.0%
	NDCG@1	0.0201	0.0315	0.0362	0.0288	0.0291	0.0343	**0.0415**	14.6%
	NDCG@10	0.0341	0.0392	0.0448	0.0429	0.0409	0.0422	**0.0512**	14.3%
	NDCG@20	0.0396	0.0499	0.0513	0.0489	0.0502	0.0511	**0.0591**	15.2%
	NDCG@40	0.0670	0.0689	0.0711	0.0676	0.0709	0.0712	**0.0823**	15.6%
Douban Book	Recall@1	0.0267	0.0205	0.0333	0.0301	0.0329	0.0324	**0.0387**	16.2%
	Recall@10	0.0311	0.0377	0.0411	0.0339	0.0362	0.0401	**0.0458**	11.4%
	Recall@20	0.0339	0.0252	0.0431	0.0309	0.0396	0.0478	**0.0515**	7.7%
	Recall@40	0.0641	0.0707	0.0749	0.0691	0.0628	0.0481	**0.0801**	6.9%
	Precision@1	0.0302	0.0344	0.0351	0.0325	0.0281	0.0294	**0.0401**	14.2%
	Precision@10	0.0391	0.0402	0.0415	0.0378	0.0356	0.0352	**0.0476**	14.7%
	Precision@20	0.0420	0.0495	0.0538	0.0322	0.0376	0.0309	**0.0541**	0.6%
	Precision@40	0.0599	0.0618	0.0725	0.0425	0.0564	0.0468	**0.0745**	2.8%
	NDCG@1	0.0301	0.0295	0.0327	0.0341	0.0205	0.0202	**0.0395**	15.8%
	NDCG@10	0.0356	0.0441	0.0457	0.0401	0.0398	0.0268	**0.0552**	20.8%
	NDCG@20	0.0391	0.0301	0.0502	0.0425	0.0463	0.0294	**0.0641**	27.7%
	NDCG@40	0.0682	0.0691	0.0663	0.0645	0.0601	0.0507	**0.0813**	19.2%

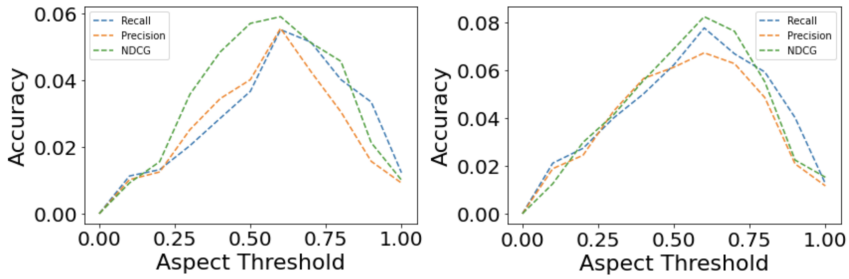
Aspect Threshold Influence: To analyze the impact of aspect threshold δ in our Monte Carlo edge-drop strategy, we conduct extensive experiments as shown in Figure 5.2. We find that the performance peaks at $\delta = 0.6$ across all metrics and datasets. The initial improvement from $\delta = 0$ to 0.6 demonstrates the benefits of filtering less significant aspects, while the decline beyond $\delta = 0.6$ indicates that removing too many aspects leads to insufficient input features. This analysis establishes $\delta = 0.6$ as the optimal threshold for balancing aspect filtering in HIN-based recommendations.

Model Explainability: To provide a deeper understanding of SeDLR's explainability, we conduct case studies on the Walmart Recruit dataset, as illustrated in Figure 5.3. In both examples, SeDLR disentangles and highlights different user- and context-related aspects



(a) The impact of Threshold @K20 on *Douban Book*.

(b) The impact of Threshold @K40 on *Douban Book*.



(c) The impact of Threshold @K20 on *Walmart Recruit*.

(d) The impact of Threshold @K40 on *Walmart Recruit*.

Figure 5.2: SeDLR: The influence of aspect threshold in Monte Carlo edge-drop strategy.

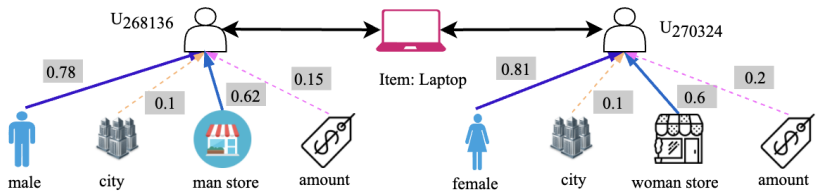


Figure 5.3: SeDLR: Two case studies from Walmart Recruit.

that contribute to the prediction of a laptop purchase. For user 268136, the model identifies "male" (score 0.78) and "man store" (score 0.62) as significant factors, while for user 278124, "female" (score 0.81) and "woman store" (score 0.66) are most influential, both above the threshold $\delta = 0.6$. This demonstrates that SeDLR generates personalized and context-aware semantic explanations: for each user, the model adapts its reasoning to the user's specific profile and shopping environment, rather than associating the item with a particular gender. Notably, the model provides similarly high purchase probabilities for both users (0.8 and 0.81), which suggests that SeDLR's explanations do not reveal a gender-related bias toward the laptop item. Instead, the explainability reflects how relevant user attributes and store context together influence each recommendation. This validates that SeDLR offers nuanced, fair, and meaningful explanations for diverse user groups.

5.1.3.4 Ablation Study

To assess the contribution of each module in SeDLR, we conduct an ablation study on the Walmart Recruit dataset. As shown in Table 5.3, removing either the disentangled embedding component or the latent reasoning module leads to a noticeable decrease in recommendation performance across all evaluation metrics. Without disentangled embeddings, the model struggles to effectively capture diverse user intents. Similarly, eliminating the latent reasoning loop reduces the model's ability to align user-item semantics during decision-making. When both modules are removed, performance degrades substantially, approaching the level of conventional models like NeuMF. These results confirm that both modules are essential to the effectiveness of SeDLR, and their joint use is critical for capturing semantic and intent-aware representations.

Table 5.3: Ablation Study of SeDLR on Walmart Recruit Dataset.

Variant	**Description**	**Recall@10**	**NDCG@10**	**Precision@10**
Full SeDL	Full model with disentangled embeddings and latent reasoning	0.0512	0.0512	0.0516
w/o Disentangled Embeddings	Replace disentangled embeddings with standard latent representations	0.0473	0.0464	0.0493
w/o Latent Reasoning	Disable latent reasoning module (intent refinement loop)	0.0451	0.0442	0.0461
w/o Both Modules	Remove both modules, retain backbone only (similar to NeuMF)	0.0401	0.0341	0.0457

5.1.3.5 Summary

In this work, we propose SeDLR, a novel framework that addresses RQ3 regarding improving recommendation explainability through causal insights. By synergistically integrating HIN-based semantic modeling with disentangled representation learning, SeDLR effectively identifies true causal factors driving user preferences while providing interpretable explanations grounded in semantics. Through Monte Carlo edge-drop analysis, we can quantify the causal significance of different semantic aspects, enabling more transparent explanations for recommendations. Extensive experiments demonstrate SeDLR's superior performance over state-of-the-art baselines in terms of both recommendation accuracy and explainability. Future work will explore fine-grained semantic disentanglement at the item-level to further enhance model explainability.

5.2 Counterfactual Explainable Conversational Recommendation

5.2.1 Overview

Conversational Recommender Systems (CRSs) fundamentally differ from traditional recommender systems by interacting with users in a conversational session to predict their preferences for personalized recommendations [94, 49, 157], as shown in Figure 5.4. Although current CRSs have achieved favorable recommendation performance, their explainability is still in its infancy stage [80, 217, 93]. Most CRSs tend to provide coarse explanations that simply list matching attributes between users and items, without exploring the deeper causal relationships that drive recommendation decisions. This limitation makes it difficult for users to understand why specific items are recommended or rejected, potentially reducing user satisfaction.

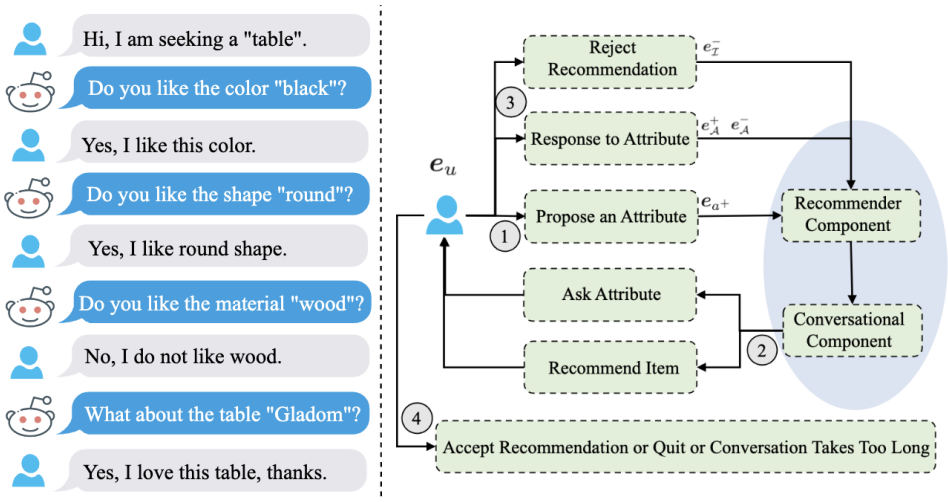


Figure 5.4: The general workflow of CRSs.

5.2.1.1 Research Objective

This study aims to address RQ3 regarding how causal insights can improve recommendation explainability by developing a novel counterfactual framework. While existing CRSs provide basic explanations based on matching attributes, they fail to explore the underlying causal mechanisms that drive user decisions. To bridge this gap, we propose a counterfactual approach that identifies minimal attribute changes necessary to alter recommendation outcomes, thereby revealing the true causal relationships between item attributes

and user preferences. Our framework serves two complementary goals: First, it enhances explainability by generating granular, causal explanations that help users understand why specific items are recommended, accepted or rejected. Second, it improves the accuracy of recommendations by learning from samples created based on these counterfactual insights. Since the system can better capture user preferences by understanding which attribute modifications might avoid rejection.

5.2.1.2 The Proposed Method

To achieve the above objectives, we propose a **C**ounterfactual **E**xplainable **C**onversational **R**ecommender (**CECR**) with four synergistic components:

- ***Offline Preference Learning*** constructs initial user and item representations from historical interaction data to establish a foundation for understanding user preferences and item characteristics in conversations.
- ***Preference Refinement*** dynamically updates user preferences based on real-time conversational feedback, incorporating both positive and negative signals.
- ***Counterfactual Reasoning*** identifies minimal attribute changes that would alter recommendation decisions, providing interpretable causal explanations for why specific items are recommended or rejected.
- ***Recommendation*** leverages the generated counterfactual explanations to create additional training samples, augmenting the training data to improve both recommendation accuracy and model robustness.

The key contributions of this research are summarized as follows:

- To the best of our knowledge, we are the first to incorporate counterfactual reasoning into conversational recommender systems. Our approach provides fine-grained, interpretable explanations that reveal how minimal attribute changes would affect recommendation outcomes, addressing a critical gap in CRS explainability.
- We propose a dynamic preference refinement mechanism that effectively synthesizes historical interactions with real-time conversational feedback. This dual-source approach enables more accurate and adaptive preference modeling, allowing the CRS to capture both the long-term and short-term user preferences.

- We develop a novel counterfactual reasoning framework that generates counterfactual explanations to improve interpretability while creating counterfactual samples to improve recommendation performance.

5.2.2 CECR

5.2.2.1 Problem Definition

Let U, I and A be the user set, item set and attribute set, respectively. Each user $u \in U$ provides positive and negative feedback during the conversational process. The positive feedback includes the positive attribute set A_u^+ containing attributes that user likes. The negative feedback comprises the negative attribute set A_u^- and rejected item set I_u^-, containing attributes and items that user dislikes. For the multi-round conversation, we aim to achieve two key objectives. First, we aim to generate accurate recommendations by predicting the items $i \in I$ that match user u's preferences expressed through online feedback on attributes $a \in A$. Second, we aim to provide counterfactual explanations $E = \{a_k | a_k \in A\}$ identifying which attributes causally influenced each recommendation decision. To achieve these objectives, we learn from both positive feedback A_u^+ and negative feedback A_u^- and I_u^- during conversations to gradually understand user preferences. Specifically, we identify minimal attribute changes Δ that would alter recommendation outcomes, enabling counterfactual reasoning as an explanation of why specific items are recommended or rejected. These counterfactual insights are then used to generate additional training samples to improve the system's recommendation accuracy.

5.2.2.2 Methodology

As shown in Figure 5.5, the framework of CECR consists of four synergistic components designed to enhance both recommendation accuracy and explainability through counterfactual reasoning. To begin with, our first component ***GNNs-based Offline Learning*** learns offline item, attribute and user representation based on their historical interaction data, which can serve as starting points for conversations. Specifically, we first resort a Multi-Layer Perceptron (MLP) to jointly consider the interaction information with rating feedback as

$$\mathbf{z}_{u,i} = MLP\left(\left[\boldsymbol{e}_i^{(in)} \oplus \boldsymbol{e}_{y_{u,i}}\right]\right) \tag{5.13}$$

where $\oplus$ is the concatenation operation. $\mathbf{z}_{u,i}$ is the rating-aware representation of user u for item i, which integrates the initial item embedding $\boldsymbol{e}_i^{(in)}$ with the user's rating encod-

ing $\boldsymbol{e}_{y_{u,i}}$. Following that, we model the offline user representation $\boldsymbol{e}_u$ by considering both item-space interactions and social-space relationships. Specifically, we construct the user representation through two complementary components: the item-space user latent factor $\boldsymbol{e}_u^U$ capturing user-item interactions, and the social-space user latent factor $\boldsymbol{e}_u^S$ representing user-user social relations. For the item-space factor, we aggregate a user's historical interactions through an attention-weighted mechanism:

$$\boldsymbol{e}_u^U = \sigma(\mathbf{W} \cdot \sum_{i \in \mathscr{I}_u} \alpha_{ui} \mathbf{z}_{u,i} + \mathbf{b}) \tag{5.14}$$

where σ is a rectified linear unit as the non-linear activation function. The attention weight α_{ui} captures the varying importance of different item interactions:

$$\begin{aligned} \alpha_{ui}^* &= \mathbf{W}_2^\alpha \cdot \sigma\left(\mathbf{W}_1^\alpha \cdot \left[\mathbf{z}_{u,i} \oplus \boldsymbol{e}_u^{(in)}\right] + \mathbf{b}_1\right) + \mathbf{b}_2 \\ \alpha_{ui} &= \frac{\exp(\alpha_{ui}^*)}{\sum_{i \in \mathscr{I}_u} \exp(\alpha_{ui}^*)} \end{aligned} \tag{5.15}$$

where $\boldsymbol{e}_u^{(in)}$ represents the initial user attributes, and α_{ui} is normalized by the attentive score α_{ui}^* to remove dimensional effect between data features.

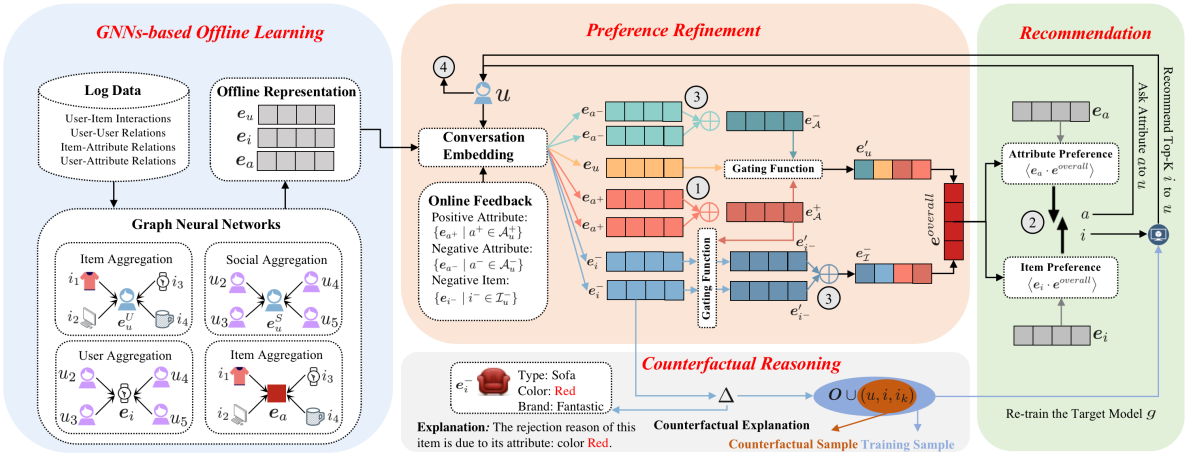


Figure 5.5: The overall framework of our proposed method CECR.

Following social theory [150], we further incorporate social influence by defining the social-space user latent factor $\boldsymbol{e}_u^S$ that captures preference signals from a user's social connections. Through neighbor aggregation over the social graph $\mathscr{S}$, we compute this social-space representation as:

$$\boldsymbol{e}_u^S = \sigma(\mathbf{W} \cdot \sum_{o \in N(u)} \beta_{ko} \boldsymbol{u}_o^U + \mathbf{b}) \tag{5.16}$$

where $\{\boldsymbol{u}_o^U, \forall o \in N(u)\}$ denotes the embeddings of u's directly connections. Social attention weights β_{ko} are learned through a two-layer neural network:

$$\begin{aligned} \beta_{ko}^* &= \mathbf{W}_2^\beta \cdot \sigma\left(\mathbf{W}_1^\beta \cdot \left[\boldsymbol{u}_o^U \oplus \boldsymbol{e}_u^{(in)}\right] + \mathbf{b}_1\right) + \mathbf{b}_2 \\ \beta_{ko} &= \frac{\exp\left(\beta_{ko}^*\right)}{\sum_{o \in N(u)} \exp\left(\beta_{ko}^*\right)} \end{aligned} \tag{5.17}$$

where β_{ko}^* is the attentive score and β_{ko} represents the social strengths between users. Finally, we combine the item-space and social-space factors through a MLP to obtain the overall user representation:

$$\begin{aligned} \boldsymbol{e}_u^{(0)} &= MLP([\boldsymbol{e}_u^U \oplus \boldsymbol{e}_u^S]) \\ \boldsymbol{e}_u^{(1)} &= \sigma\left(\mathbf{W}_2^U \cdot \boldsymbol{e}_u^{(0)} + \mathbf{b}_2\right) \\ &\cdots \\ \boldsymbol{e}_u &= \sigma\left(\mathbf{W}_\ell^U \cdot \boldsymbol{e}_u^{(\ell-1)} + \mathbf{b}_\ell\right) \end{aligned} \tag{5.18}$$

where ℓ is hidden layer number of MLP and $\boldsymbol{e}_u^{(\ell-1)}$ is the user representation on the $(\ell-1)$ layer. The overall user representation $\boldsymbol{e}_u$ effectively captures both high-order interaction patterns and social influence signals, providing a comprehensive model of user preferences. Next, we construct the offline item representation $\boldsymbol{e}_i$ by capturing diverse user feedback patterns for each item. Since different users may interact with the same item differently, incorporating this feedback diversity is crucial for characterizing item features comprehensively. Mathematically, we aggregate user interactions through an attention-weighted mechanism:

$$\boldsymbol{e}_i = \sigma(\mathbf{W}^i \cdot \sum_{u \in B(i)} \mu_{iu} \mathbf{z}_{i,u} + \mathbf{b}) \tag{5.19}$$

where $\mathbf{z}_{i,u}$ is rating-aware representation of item i for user u. The user attention weights μ_{iu} are learned through a two-layer neural network:

$$\begin{aligned} \mu_{iu}^* &= \mathbf{W}_2^\mu \cdot \sigma\left(\mathbf{W}_1^\mu \cdot \left[\mathbf{z}_{i,u} \oplus \boldsymbol{e}_i^{(in)}\right] + \mathbf{b}_1\right) + \mathbf{b}_2 \\ \mu_{iu} &= \frac{\exp\left(\mu_{iu}^*\right)}{\sum_{u \in B_{(i)}} \exp\left(\mu_{iu}^*\right)} \end{aligned} \tag{5.20}$$

where μ_{iu}^* is the attentive score and $\boldsymbol{e}_i$ it the overall item representation for item i containing information from high-order interaction patterns.

Likewise, we learn the offline attribute representation $\boldsymbol{e}_a$ by aggregating information from items sharing the same attribute. Since items with common attributes typically belong to similar categories, they exhibit correlated interaction patterns. We aggregate these

items using their rating-aware representations $\mathbf{z}_{a,i}$ and incorporate user attention weights ψ_{iu} similar to the item-level attention mechanism. This process generates attribute representations that capture rich semantic patterns from similarly-attributed items. Through this comprehensive approach, we obtain GNN-based representations for users ($\boldsymbol{e}_u$), items ($\boldsymbol{e}_i$), and attributes ($\boldsymbol{e}_a$) that form the foundation for subsequent recommendation and explanation tasks.

Next, the second component ***Preference Refinement*** dynamically updates user preferences by synthesizing current conversational feedback with historical interaction patterns. Let $u \in \mathcal{U}$ denote a user, with $\mathcal{A}_u^+$ and $\mathcal{A}_u^-$ representing the sets of positive and negative attributes from online feedback respectively. We model the user's feedback representation through:

$$\mathbf{R}_{\mathcal{A}_u^+} = \left[\mathbf{e}_{a^+} \vdots \mathbf{e}_{a^+}\right], \mathbf{R}_{\mathcal{A}_u^-} = \left[\mathbf{e}_{a^-} \vdots \mathbf{e}_{a^-}\right], \mathbf{R}_{\mathcal{I}_u^-} = \left[\mathbf{e}_i^- \vdots \mathbf{e}_i^-\right] \tag{5.21}$$

where $\mathbf{R}_{\mathcal{A}_u^+}$, $\mathbf{R}_{\mathcal{A}_u^-}$, and $\mathbf{R}_{\mathcal{I}_u^-}$ represent matrix encodings of accepted attributes, rejected attributes, and rejected items. To effectively aggregate positive feedback signals, we employ attention mechanisms:

$$\mathbf{R}^+ = \text{softmax}\left(\frac{\mathbf{R}_{\mathcal{A}_u^+}\mathbf{W}_j(\mathbf{R}_{\mathcal{A}_u^+})^T\mathbf{W}_k}{\sqrt{d}}\right)\mathbf{R}_{\mathcal{A}_u^+}\mathbf{W}_f \tag{5.22}$$

where $\mathbf{R}^+$ captures contextual information across positive feedback. $\mathbf{W}_j$, $\mathbf{W}_k$, and $\mathbf{W}_f$ are learnable projection matrices. For negative feedback including both rejected attributes and items, we construct $\mathbf{R}_{ne} = \mathbf{R}_{\mathcal{A}_u^-} \oplus \mathbf{R}_{\mathcal{I}_u^-}$ and compute:

$$\mathbf{R}^- = \text{softmax}\left(\frac{\mathbf{R}_{ne}\mathbf{W}_j(\mathbf{R}_{ne})^T\mathbf{W}_k}{\sqrt{d}}\right)\mathbf{R}_{ne}\mathbf{W}_f \tag{5.23}$$

where $\mathbf{R}^-$ captures contextual information across negative feedback. The final refined user preference is then updated as:

$$\mathbf{e}_u^t = \mathbf{e}_u + \mathbf{e}_u^+ - \mathbf{e}_u^- \tag{5.24}$$

where $\mathbf{e}_u^t$ denotes the user preference at round t, combining long-term preferences $\mathbf{e}_u$ with positive $\mathbf{e}_u^+$ and negative $\mathbf{e}_u^-$ feedback signals from the current conversation. This enables our model to maintain an adaptive balance between historical patterns and real-time user feedback for more accurate preference modeling.

Now, we start updating the user representations in each conversation round. We first aggregate positive and negative feedback signals to capture attribute-level preferences:

$$\boldsymbol{e}_{\mathcal{A}}^- = \frac{1}{\left|\mathcal{A}_u^-\right|}\sum_{a^- \in \mathcal{A}_u^-}\boldsymbol{e}_{a^-}, \boldsymbol{e}_{\mathcal{A}}^+ = \frac{1}{\left|\mathcal{A}_u^+\right|}\sum_{a^+ \in \mathcal{A}_u^+}\boldsymbol{e}_{a^+} \tag{5.25}$$

where $\mathbf{e}^-_{\mathscr{A}}$ and $\mathbf{e}^+_{\mathscr{A}}$ represent aggregated negative and positive attribute feedback. To adaptively balance long-term and short-term preferences, we introduce gating mechanisms:

$$\begin{aligned} \boldsymbol{C}^-_u &= \sigma\left(\mathbf{W}_3 \cdot \text{ Concat } \left(\boldsymbol{e}^-_{\mathscr{A}}, \boldsymbol{e}_u, \boldsymbol{e}^-_{\mathscr{A}} \odot \boldsymbol{e}_u\right) + \mathbf{b}\right) \\ \boldsymbol{C}^+_u &= \sigma\left(\mathbf{W}_4 \cdot \text{ Concat } \left(\boldsymbol{e}^+_{\mathscr{A}}, \boldsymbol{e}_u, \boldsymbol{e}^+_{\mathscr{A}} \odot \boldsymbol{e}_u\right) + \mathbf{b}\right) \\ \boldsymbol{e}'_u &= \boldsymbol{e}_u \odot (\boldsymbol{C}^-_u - \boldsymbol{C}^+_u) \end{aligned} \tag{5.26}$$

where $\boldsymbol{C}^+_u$ and $\boldsymbol{C}^-_u$ are the gating functions [217] to control the information propagated via feedback based on positive and negative attribute signals. $\odot$ means the element-wise product. The final updated user representation $\boldsymbol{e}'_u$ is computed as:

$$\mathbf{e}'_u = \mathbf{e}_u \odot (\mathbf{C}^-_u - \mathbf{C}^+_u) \tag{5.27}$$

where $\boldsymbol{e}'_u$ is the updated user representation. Likewise, we begin to update item representation during each conversation round. Given that users often reject items due to specific attributes rather than the entire item, we need a refined approach to handle negative feedback at the item level while preserving positive attribute associations. Let $\mathbf{e}^-_i$ denote the representation of a rejected item. We compute the updated item representation through:

$$\begin{aligned} \boldsymbol{C}^-_i &= \sigma\left(\mathbf{W}_5 \cdot \text{ Concat } \left(\boldsymbol{e}^+_{\mathscr{A}}, \boldsymbol{e}^-_i, \boldsymbol{e}^+_{\mathscr{A}} \odot \boldsymbol{e}^-_i\right) + \mathbf{b}\right) \\ \boldsymbol{e}'_{i^-} &= \boldsymbol{e}^-_i \odot \boldsymbol{C}^-_i \end{aligned} \tag{5.28}$$

where $\boldsymbol{e}'_{i^-}$ it the updated item representation computed through the gating function $\boldsymbol{C}^-_i$. The final aggregated representation of all rejected items is then computed as:

$$\boldsymbol{e}^-_{\mathscr{I}} = \frac{1}{\left|\mathscr{I}^-_u\right|} \sum_{i^- \in \mathscr{I}^-_u} \boldsymbol{e}'_{i^-} \tag{5.29}$$

where $\mathbf{e}^-_{\mathscr{I}}$ captures the collective negative feedback at the item level while accounting for the influence of positive attributes. This enables our model to learn finer-grained preference patterns by distinguishing between rejected attributes and preserving valuable positive signals within rejected items, thereby enhancing recommendation quality. Next, we formulate the overall user preference representation $\mathbf{e}^{overall}$ by combining the refined user preferences and item-level feedback as:

$$\mathbf{e}^{overall} = \mathbf{e}'_u - \mathbf{e}^-_{\mathscr{I}} \tag{5.30}$$

where $\mathbf{e}'_u$ represents the updated user preferences incorporating both long-term and short-term signals, while $\mathbf{e}^-_{\mathscr{I}}$ captures aggregated negative feedback at the item level.

The third component ***Counterfactual Reasoning*** aims to address data sparsity while enhancing model explainability through counterfactual analysis. We first employ a multi-layer neural network $g(\cdot)$ to estimate the ranking score s_{ui} between user u and item i:

$$\begin{aligned} s_{ui}^{(1)} &= \mathbf{W}_1 \cdot \text{ReLU}\left(g(\boldsymbol{e}'_u, \boldsymbol{e}_i)\right) + \mathbf{b}_1 \\ s_{ui}^{(2)} &= \mathbf{W}_2 \cdot \text{ReLU} \cdot s_{ui}^{(1)} + \mathbf{b}_2 \\ &\cdots \\ s_{ui} &= \mathbf{W}_M \cdot \text{ReLU} \cdot s_{ui}^{(M-1)} + \mathbf{b}_M \end{aligned} \tag{5.31}$$

where the ReLU is the rectified linear unit activation function for all layers ($m \in [1, M]$). Explanation Strength (ES) and Explanation Complexity (EC) are two aspects used to describe counterfactual explainability [158]. According to the Occam's Razor Principle [46], we seek the strong (i.e., high ES) yet simple (i.e., low EC) counterfactual explanation, which is a small change vector on item: $\Delta = \{\delta_0, \delta_1, \cdots, \delta_d\}$. The elements in Δ are initialized to zero and constrained within attribute-specific ranges (e.g., [-1,1] for binary attributes). The counterfactual explanation is generated by iteratively updating Δ in the embedding space to find the minimal attribute changes that would modify the recommendation result: $\mathbf{e}'_i = \mathbf{e}_i + \Delta$, where $\mathbf{e}'_i$ represents the counterfactually modified item embedding. This approach enables the identification of fine-grained attribute-level changes required to affect recommendations, providing interpretable explanations while generating valuable counterfactual samples to enhance model robustness.

We quantify the effectiveness of counterfactual explanations through two key metrics: Explanation Strength (ES) and Explanation Complexity (EC). The ES measures how significantly a small change Δ influences the model's decision:

$$S(\Delta) = s_{u,i} - s_{u,i_\Delta} \tag{5.32}$$

where s_{u,i_Δ} represents the ranking score after applying change Δ. We establish a threshold ϵ based on the score difference between items:

$$S(\Delta) > \epsilon = s_{u,i} - s_{u,i_k} \tag{5.33}$$

where s_{u,i_k} is the k-th item's ranking score. When $S(\Delta) > \epsilon$, the explanation demonstrates strong influence, indicating the change significantly impacts recommendations. For EC, we consider both the number of modified attributes ($|\Delta|_0$) and magnitude of changes ($|\Delta|_2^2$). Following [16, 15], we approximate the non-convex $|\Delta|_0$ using the ℓ_1-norm $|\Delta|_1$. The optimization objective becomes:

$$\underset{\Delta}{\text{minimize}} \|\Delta\|_2^2 + \|\Delta\|_1 + \gamma\|\Delta\|_0 + \lambda \log[\sigma(s_{u,i_\Delta}, s_{u,i_k})] \tag{5.34}$$

where γ and λ are tuning parameters. $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function. We iteratively update Δ through:

$$\Delta^{(t+1)} = \Delta^{(t)} - \Phi \left[\frac{\lambda \log[\sigma(\epsilon)]}{\Delta^{(t)}} + \frac{\|\Delta^{(t)}\|_2^2 + \gamma \|\Delta^{(t)}\|_1}{\Delta^{(t)}} \right] \tag{5.35}$$

where Φ represents the learning rate. $\Delta^{(t)}$ is the change vector at iteration t, which can serve as the counterfactual explanation after t iterations to improve the explainability of the CRS task, providing the fine-grained recommendation reasons.

With the counterfactual explanation $\Delta^{(t)}$, we then address data sparsity through counterfactual sample augmentation. Let $\mathbf{O}$ denote the set of original training samples, where each triplet (u, i_1, i_2) indicates user u prefers item i_1 over i_2. We optimize the model through a modified Bayesian Personalized Ranking (BPR) loss:

$$\begin{aligned} &\text{minimize}[-\sum_{(u,i_1,i_2)\in \boldsymbol{O}} \log[\sigma(\epsilon)] + \Omega\|g\|^2] \\ &\text{s.t., Generate } (u, i, i_k) \cup \boldsymbol{O}. \text{ if has } \Delta^{(t)} \end{aligned} \tag{5.36}$$

where $\Omega|g|^2$ represents the regularization term. When a counterfactual explanation $\Delta^{(t)}$ is available for a rejected item i, we generate a new training sample (u, i, i_k) where i_k represents the k-th ranked item that displaced item i from the Top-K recommendations after applying $\Delta^{(t)}$. By augmenting the training data with these counterfactual samples $\mathbf{O} \cup (u, i, i_k)$, we enhance the model's robustness and performance on sparse datasets while maintaining interpretability through the counterfactual reasoning process.

The final component ***Recommendation*** incorporates counterfactual samples into the training process and then makes predictions. Given an item $i \in \mathscr{I}$ or attribute $a \in \mathscr{A}$, we compute preference scores using the overall user representation $\mathbf{e}^{overall}$:

$$\begin{aligned} y\left(i \mid u, \boldsymbol{e}_{\mathscr{I}}^{-}\right) &= \left\langle \boldsymbol{e}_i \cdot \boldsymbol{e}^{overall} \right\rangle, \\ y\left(a \mid u, \boldsymbol{e}_{\mathscr{I}}^{-}\right) &= \left\langle \boldsymbol{e}_a \cdot \boldsymbol{e}^{overall} \right\rangle \end{aligned} \tag{5.37}$$

where the system selects between recommending items or querying attributes by comparing their respective preference scores $y(i|u, \mathbf{e}_{\mathscr{I}}^{-})$ and $y(a|u, \mathbf{e}_{\mathscr{I}}^{-})$. Overall, the training process consists of three phases: First, we train the base model g on the original dataset $\mathbf{O}$. Second, when users reject recommendations, we generate counterfactual samples (u, i, i_k). Third, we retrain the model on the augmented dataset $\mathbf{O} \cup (u, i, i_k)$. During conversations, the system recommends Top-K items only when the "recommend item" action is selected; otherwise, it continues attribute-based exploration.

To facilitate efficient development and evaluation of conversational recommender systems without requiring extensive real-user data collection, we employ a user simulator that interacts with the system in a principled manner. The simulator takes as input the user's historical preferences and system queries (either attribute-based questions or item recommendations) and generates realistic feedback that mimics actual user responses like acceptance or rejection of recommendations. To enable dynamic interaction, the simulator supports both passive response generation and active preference probing. When the recommender system poses questions about specific attributes or items, the simulator draws upon the stored user preference model to generate appropriate responses. These simulated interactions allow the system to iteratively refine its recommendation strategy by gathering additional preference information through strategic question-asking, just as it would in real conversational scenarios. This simulation-based approach provides several key advantages: it enables rapid prototyping and evaluation of different conversational strategies, ensures reproducibility of experiments, and allows systematic assessment of the system's performance under various user behavior patterns.

5.2.3 Experiments

5.2.3.1 Datasets

We evaluate our model on three benchmark datasets: *Yelp*[3], *Douban-Book*[4] and *MovieLens*[5]. As shown in Table 5.4, each dataset contains rich user-item interaction data along with diverse attributes: Yelp captures user feedback on businesses, with user attributes (age, gender, location) and business attributes (category, price range, reviews). Douban-Book comprises book ratings and social connections, including user metadata (location, groups, occupation) and book information (author, publisher, year). MovieLens contains movie ratings with user demographics (age, gender, occupation) and film properties (genre, year). To ensure data quality, we filter out users with fewer than 10 interactions and remove infrequent attributes. The remaining data is split into training/validation/test sets with ratios of 70%/20%/10%.

[3]https://www.yelp.com/dataset/
[4]https://github.com/librahu/HIN-Datasets-for-Recommendation-and-Network-Embedding
[5]https://github.com/librahu/HIN-Datasets-for-Recommendation-and-Network-Embedding

Table 5.4: CECR: Statistical details of the three datasets.

Dataset	#User	#Item	#Interaction	#Density
Yelp	27,675	70,311	1,368,606	0.07%
Douban-Book	13,024	22,347	792,062	0.27%
MovieLens	943	1,682	100,000	6.30%

5.2.3.2 Baselines and Evaluation

For recommendation evaluation, we employ Success Rate (SR@r), which measures the success rate within the first r conversation rounds, and Average Turns (AT), which calculates the average number of rounds until conversation completion [93, 217]. For explainability evaluation, we use both user-oriented and model-oriented metrics. The user-oriented metrics include Precision, which measures explanation accuracy by calculating the ratio of correctly identified attributes in the explanation, and Recall, which evaluates explanation completeness by measuring the percentage of user-liked attributes included. The model-oriented metrics consist of Probability of Necessity (PN), assessing how necessary an attribute is for recommendations, and Probability of Sufficiency (PS), evaluating how sufficient an attribute is for making recommendations. Mathematically, Precision and Recall are defined as

$$\text{Precision} = \frac{\sum_{a=1}^{d} p_{u,i}^{(a)} \cdot I(\delta_a)}{\sum_{a=1}^{d} I(\delta_a)}, \ \text{Recall} = \frac{\sum_{a=1}^{d} p_{u,i}^{(a)} \cdot I(\delta_a)}{\sum_{a=1}^{d} p_{u,i}^{(a)}} \tag{5.38}$$

where $p_{u,i}^{(a)}$ indicates user u's feedback on attribute a of item i, and $I(\delta_a)$ is an indicator function that equals 1 when $\delta_a \neq 0$. Moreover, PN and PS are defined as

$$\begin{aligned} \text{PN} &= \frac{\sum_{u \in \mathcal{U}} \sum_{i \in R_{u,K}} \text{PN}_{ui}}{\sum_{u \in \mathcal{U}} \sum_{i \in R_{u,K}} I(\mathcal{A}_{ui} \neq \emptyset)}, \\ \text{PS} &= \frac{\sum_{u \in \mathcal{U}} \sum_{i \in R_{u,K}} \text{PS}_{ui}}{\sum_{u \in \mathcal{U}} \sum_{i \in R_{u,K}} I(\mathcal{A}_{ui} \neq \emptyset)} \end{aligned} \tag{5.39}$$

where $R_{u,K}$ denotes the Top-K recommendation list for user u, PNui and PS$_{ui}$ are the necessity and sufficiency scores for recommending item i to user u, and $I(\mathcal{A}_{ui} \neq \emptyset)$ is an indicator function that equals 1 when explanation attributes exist for the user-item pair. In this way, we can still create counterfactual item i' and counterfactual recommendation list $R'_{u,K}$. For implementation, we set the embedding dimension to 128 with regularization parameter 1e-5. The learning rate is tuned in [0.0001, 0.1], and Top-K is set to 10 for all models. We employ Adam optimizer for training due to its superior convergence speed and low

memory requirements. For a fair comparison, baseline models are tuned using the same hyperparameter ranges as ours.

We compare our proposed CECR to the following state-of-the-art CRSs:

EAR [93]: implements a three-stage reinforcement learning framework for estimating user preferences in CRSs through estimation, action and reflection stages.

UNICORN [41]: adopts a unified approach for multi-stage decision optimization to enhance the CRS performance and scalability.

FPAN [217]: designs gating modules for negative and positive feedback processing, and then adapts user preferences based on online feedback in CRS.

KBRD [24] : leverages knowledge graphs to incorporate users' knowledge-grounded information for more accurate preference modeling.

CPR [94] : reformulates conversational recommendation as a path-finding problem, employing reinforcement learning to discover meaningful explanations.

CAEC [89] : employs a BERT-based architecture to generate explanations in CRS without requiring prior knowledge about users.

SAUR [246]: enhances questioning capabilities through Multi-Memory Networks while utilizing search history for explanation generation.

Random: a basic baseline that simply selects attributes randomly as explanations.

Note that we did not include our own conversational model CCR in this comparison, even though both models are evaluated on shared datasets, because they are designed for fundamentally different settings. Our previous CCR is built for improving predictive accuracy while CECR in this work is designed for improving the interpretability. Therefore, their architectures and objectives are not directly comparable.

5.2.3.3 Result Analysis

Performance Comparison: Table 5.5 shows CECR's superior performance across all datasets and metrics. CECR achieves significant improvements in SR@10 over the strongest baselines: 5.75% on Yelp, 6.10% on Douban-Book, and 6.87% on MovieLens. Moreover, CECR requires fewer conversation rounds to reach successful recommendations, as evidenced by the lowest AT values among all models. Meanwhile, figure 5.6 reveals the relationship between success rate and conversation rounds. All models show rapid improvement in SR during the first 5 rounds, followed by a slower increase until reaching peak performance around round 10. This pattern reflects the natural progression of conversational recommendation: initial rounds effectively narrow down candidate items, while later rounds face diminishing returns as the search space becomes more focused. Notably, CECR outper-

Table 5.5: CECR: Recommendation performance comparison: the best results are bolded, while the best baselines are underlined.

Dataset	Yelp				Douban-Book				MovieLens			
ModelMetrics	SR@5	SR@10	SR@20	AT	SR@5	SR@10	SR@20	AT	SR@5	SR@10	SR@20	AT
EAR	0.793	0.818	0.859	4.88	0.812	0.833	0.866	4.82	0.844	0.871	0.943	4.31
FPAN	0.811	0.851	0.911	4.66	0.833	0.867	0.918	4.51	0.851	0.888	0.935	4.11
UNICORN	0.805	0.852	0.903	5.01	0.836	0.869	0.914	4.89	0.843	0.865	0.937	4.17
KBRD	0.799	0.833	0.909	4.81	0.821	0.848	0.915	4.66	0.841	0.869	0.941	4.21
CECR	**0.852**	**0.901**	**0.961**	**4.31**	**0.866**	**0.922**	**0.973**	**4.22**	**0.891**	**0.949**	**0.982**	**3.77**
Improv. %	5.06%	5.75%	5.49%	7.51%	3.59%	6.1%	6%	6.43%	4.7%	6.87%	4.14%	8.27%

forms all baselines, reaching 0.901, 0.949 and 0.922 at round 10 on *Yelp*, *MovieLens* and *Douban-Book*, respectively. Thus, we conclude that the more rounds the higher the SR is, and the increasing trend can slow down in 5 rounds due to the reduced room for improvement.

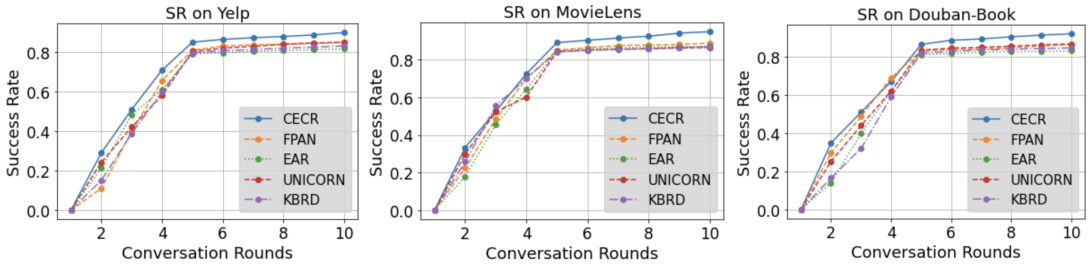


Figure 5.6: CECR: Success rate comparison.

Ablation Study: Table 5.7 presents our ablation study examining component contributions on SR@20 and AT metrics across MovieLens and Yelp datasets. We evaluate four key components: $GNNs$ means removing aggregation function, using only first-layer representations. $PreRi$ means removing preference refinement, ignoring online feedback. $CouRe$ means removing counterfactual sampling, training only on original data. $RecPe$ means removing preference score comparison, using random decisions. CCR is the complete CECR model. The results demonstrate that preference refinement and counterfactual reasoning are the most crucial components. Removing preference refinement ($PreRi$) significantly degrades SR@20 to 0.599/0.657 on MovieLens/Yelp, while removing counterfactual reasoning ($CouRe$) reduces it to 0.613/0.686. Both components' removal also substantially increases AT, requiring approximately 7 rounds for convergence. These findings validate that CECR's superior performance stems from its ability to leverage counterfactual techniques to distill positive information from users' online feedback.

Explainability Analysis: Figure 5.6 demonstrates CECR's superior explainability compared to state-of-the-art baselines across both user-oriented and model-oriented metrics.

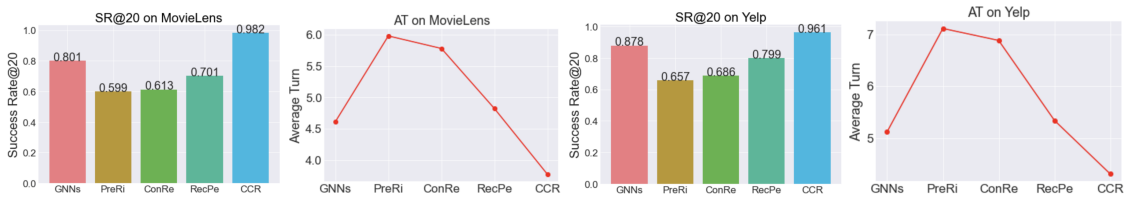


Figure 5.7: CECR: ablation study for our CECR.

Table 5.6: CECR: Explainability comparison: the best results are bolded, while the best baselines are underlined.

Dataset	Yelp				MovieLens			
ModelMetrics	Pr%	Re%	PN%	PS%	Pr%	Re%	PN%	PS%
CPR	5.67	44.35	9.66	3.47	9.31	55.13	2.35	4.25
CAEC	20.81	36.44	79.81	82.15	28.14	46.67	85.55	89.97
SAUR	22.34	4.71	88.26	91.59	26.48	6.74	91.51	93.14
Random	0.25	0.31	1.68	3.45	0.33	0.51	2.22	5.14
CECR	**29.51**	**50.37**	**94.11**	**96.39**	**35.14**	**59.91**	**96.42**	**97.12**
Improv. %	14.96	12.66	4.96	3.12	19.89	5.59	4.41	3.05

The baseline Random's poor performance confirms the importance of meaningful explanations. On Yelp and MovieLens respectively, CECR achieves significant improvements: 14.96% and 19.89% in Precision, 12.66% and 5.59% in Recall, 4.96% and 4.41% in NP, 3.12% and 3.05% in Sufficiency over the strongest baselines. The consistently higher Probability of Sufficiency versus Probability of Necessity scores indicate that CECR effectively identifies attributes that are truly relevant for recommendations. These results validate CECR's ability to generate explanations that both align with user preferences and accurately reflect the system's decision-making process.

Case Study: Figure 5.8 presents a case study from MovieLens that validates CECR's explanation capabilities. CECR demonstrates two key advantages in explanation generation: First, for the accepted movie "The Mummy", CECR provides fine-grained explanations based on attribute matching, highlighting how the movie aligns with user preferences for "Adventure", "Action" and "Thriller" genres. This transparency helps users better understand the recommendation rationale. Second, for the rejected movie "Uncharted", CECR generates system-oriented counterfactual explanations, identifying "Mystery" as the key attribute leading to rejection. The explanation suggests that modifying this genre attribute could potentially change the recommendation outcome. These insights not only guide future recommendations but also help generate counterfactual training samples. Additionally, CECR achieves successful recommendations in fewer conversation rounds compared to traditional CRS by leveraging counterfactual reasoning to identify promising candidates from

negative feedback. Thus, CECR can provide not only fine-grained user-oriented explanations for accepted items but also system-oriented counterfactual explanations for rejected items, as well as generate counterfactual samples to supplement the training set.

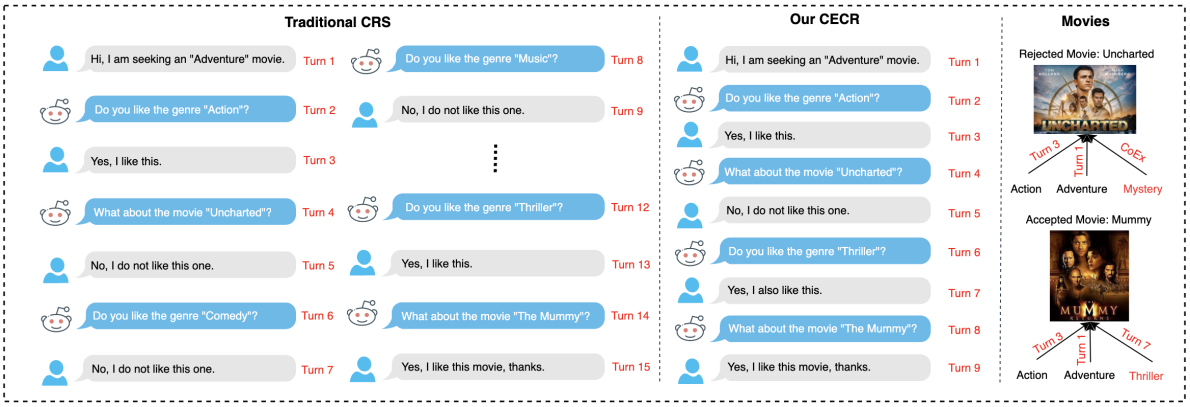
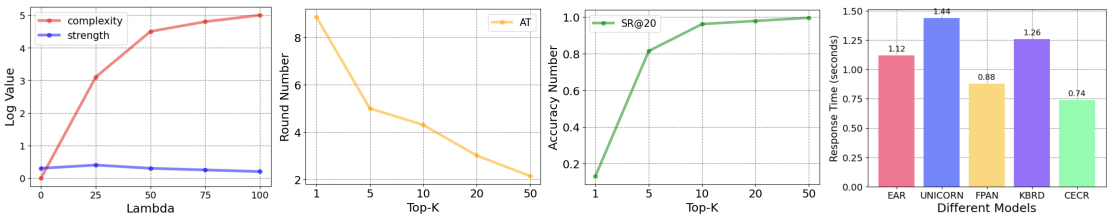


Figure 5.8: CECR: A case study to demonstrate the model explainability.

Influence of Parameters: To understand CECR's sensitivity to key parameters, we analyze the effects of λ and Top-K on Yelp. λ balances explanation strength and complexity, where larger values prioritize stronger explanations at the cost of increased complexity. Figure 5.9 shows that increasing λ enables CECR to generate explanations for more items, though with higher complexity. The explanation strength remains stable across different λ values due to ϵ's control over ranking scores. Regarding Top-K, larger values improve both SR@20 and AT metrics, as recommending more items naturally increases the likelihood of successful recommendations. Based on these findings, we set $\lambda = 6$ as optimal, while K can be adjusted according to specific application needs.



(a) Impacts of λ for counterfactual explanation.

(b) Impacts of Top-K for AT.

(c) Impacts of Top-K for SR@20.

(d) Time complexity comparison.

Figure 5.9: CECR: The analysis of different parameters and time complexity.

Time Complexity: To verify the efficiency of our proposed CECR, we compare the response time taken for recommendation generation of CECR to baselines, denoting the time

complexity of recommendation generation. As shown in Figure 5.9, CECR has the fastest response time of 0.74 seconds through its effective preference refinement mechanism. This surpasses all baselines: EAR (1.12s), FPAN (0.88s), UNICORN (1.44s), and KBRD (1.26s). The performance differences primarily stem from algorithmic complexity. While EAR and FPAN maintain reasonable speeds through gating mechanisms, UNICORN and KBRD require longer processing due to their multi-stage optimization and knowledge graph operations respectively. These results validate CECR's ability to maintain computational efficiency while delivering high-quality recommendations.

5.2.3.4 Summary

In this work, we propose CECR, a novel framework that addresses RQ3 regarding improving recommendation explainability through causal insights. By synergistically integrating counterfactual reasoning with conversational dynamics, CECR effectively generates both user-oriented and system-oriented explanations while enhancing recommendation performance through counterfactual sampling. CECR provides transparent explanations through two complementary mechanisms: attribute-based reasoning for accepted items and counterfactual analysis for rejected items. This dual approach not only helps users understand why items are recommended but also enables the system to learn from rejection feedback. The counterfactual samples further augment the training data, creating a virtuous cycle of continual improvement. Future work will explore incorporating large language models and multimedia processing capabilities to enable more natural and diverse explanations across different data modalities.

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1 Conclusion

This thesis has presented a systematic investigation into the challenges posed by spurious correlations in recommender systems through the causal inference techniques. Through three interconnected research questions, the thesis has established a comprehensive causal-based framework for developing more robust, unbiased, and interpretable recommender systems.

First, we develop foundational causal models for mitigating confounding effects caused by spurious correlations in recommendations (RQ1). Through the novel integration of causal adjustment and propensity scoring in Chapter 3, the LDPE and CCR frameworks successfully distinguish true user preferences from misleading spurious correlations. These causal models demonstrated significant performance improvements over state-of-the-art baselines across multiple real-world datasets, validating the effectiveness of causal approaches in generating unbiased recommendations.

Second, we extend the causal principles to enhance model robustness against spurious correlations in complex recommendation scenarios (RQ2). Building upon the debiasing techniques established in previous Chapter 3, Chapter 4 present three novel approaches targeting distinct complex recommendation contexts: CGSR for blocking shortcut paths in session graphs, GCRec for addressing multiple confounders in complex high-order interaction networks, and CEDA for mitigating echo chamber effects in complex social networks. Through the systematic integration of GNNs with causal interventions, these ad-

vanced causal models effectively handled spurious correlations in complex scenarios while maintaining computational efficiency.

Third, we leverage causal insights from causal models to improve recommendation explainability (RQ3). In Chapter 5, our proposed causal models SeDLR and CECR employed disentangled learning and counterfactual reasoning to generate transparent explanations grounded in causality. This systematic integration of causal techniques enhanced the model explainability by revealing genuine preference patterns beneath spurious correlations, thereby providing users with meaningful insights into recommendation decisions.

In summary, this thesis has advanced the field of recommender systems through the systematic integration of causal inference techniques, effectively addressing spurious correlations across three dimensions: foundational debiasing, enhanced robustness in complex scenarios, and improved explainability. Our proposed causal models consistently outperformed state-of-the-art baselines while providing interpretable recommendations, demonstrating the effectiveness of causal inference for practical recommendation applications.

6.2 Future Work

In future work, we may focus on two promising research directions to further improve the field of recommender systems. First, leveraging causal inference to enhance LLM-based recommendations represents an opportunity to understand and mitigate the new forms of spurious correlations that emerge from large language models and their training data. This direction involves developing novel causal structures to capture semantic-level confounding effects while using counterfactual reasoning to generate more accurate explanations. Second, extending the causal framework to specialized domains, like healthcare and education, requires careful consideration of domain-specific spurious correlations arising from institutional practices and measurement challenges. Through systematic development of domain-specific causal structures and intervention techniques, we strive to enhance causality-based recommender systems, enabling them to accurately identify and mitigate spurious correlations across various real-world applications.

BIBLIOGRAPHY

[1] J. ADEBAYO, M. MUELLY, H. ABELSON, AND B. KIM, *Post hoc explanations may be ineffective for detecting unknown spurious correlation*, in International conference on learning representations, 2022.

[2] P. ADITYA, I. BUDI, AND Q. MUNAJAT, *A comparative analysis of memory-based and model-based collaborative filtering on the implementation of recommender system for e-commerce in indonesia: A case study pt x*, in 2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS), IEEE, 2016, pp. 303–308.

[3] G. ADOMAVICIUS AND A. TUZHILIN, *Context-aware recommender systems*, in Recommender systems handbook, Springer, 2010, pp. 217–253.

[4] A. F. AGARAP, *Deep learning using rectified linear units (relu)*, arXiv preprint arXiv:1803.08375, (2018).

[5] M. AI, Z. CHEN, J. WANG, J. SHANG, T. TAO, AND Z. LI, *Improve roi with causal learning and conformal prediction*, in 2024 IEEE 40th International Conference on Data Engineering (ICDE), IEEE, 2024, pp. 598–610.

[6] M. A. AL-KAHTANI AND R. SANDHU, *A model for attribute-based user-role assignment*, in 18th Annual Computer Security Applications Conference, 2002. Proceedings., IEEE, 2002, pp. 353–362.

[7] F. ALATAWI, P. SHETH, AND H. LIU, *Quantifying the echo chamber effect: An embedding distance-based approach*, in Proceedings of the International Conference on Advances in Social Networks Analysis and Mining, 2023, pp. 38–45.

[8] E. ASLANIAN, M. RADMANESH, AND M. JALILI, *Hybrid recommender systems based on content feature relationship*, IEEE transactions on industrial informatics, (2016).

[9] C. F. BARNES, S. A. RIZVI, AND N. M. NASRABADI, *Advances in residual vector quantization: A review*, IEEE transactions on image processing, 5 (1996), pp. 226–262.

[10] R. V. D. BERG, T. N. KIPF, AND M. WELLING, *Graph convolutional matrix completion*, arXiv preprint arXiv:1706.02263, (2017).

[11] S. BHADANI, *Biases in recommendation system*, in Proceedings of the 15th ACM Conference on Recommender Systems, 2021, pp. 855–859.

[12] S. BONNER AND F. VASILE, *Causal embeddings for recommendation*, in Proceedings of the 12th ACM conference on recommender systems, 2018, pp. 104–112.

[13] M. BRAUNHOFER, V. CODINA, AND F. RICCI, *Switching hybrid for cold-starting context-aware recommender systems*, in Proceedings of the 8th ACM Conference on Recommender systems, 2014, pp. 349–352.

[14] R. BURKE, *Hybrid recommender systems: Survey and experiments*, User modeling and user-adapted interaction, 12 (2002), pp. 331–370.

[15] E. J. CANDES, J. K. ROMBERG, AND T. TAO, *Stable signal recovery from incomplete and inaccurate measurements*, Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences, 59 (2006), pp. 1207–1223.

[16] E. J. CANDES AND T. TAO, *Decoding by linear programming*, IEEE transactions on information theory, 51 (2005), pp. 4203–4215.

[17] E. ÇANO AND M. MORISIO, *Hybrid recommender systems: A systematic literature review*, Intelligent data analysis, 21 (2017), pp. 1487–1524.

[18] E. CAVENAGHI, A. ZANGA, F. STELLA, AND M. ZANKER, *Towards a causal decision-making framework for recommender systems*, ACM Transactions on Recommender Systems, 2 (2024), pp. 1–34.

[19] Y. CHANG, X. WANG, J. WANG, Y. WU, L. YANG, K. ZHU, H. CHEN, X. YI, C. WANG, Y. WANG, ET AL., *A survey on evaluation of large language models*, ACM Transactions on Intelligent Systems and Technology, 15 (2024), pp. 1–45.

[20] J. CHEN, H. DONG, Y. QIU, X. HE, X. XIN, L. CHEN, G. LIN, AND K. YANG, *Autodebias: Learning to debias for recommendation*, in Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 21–30.

[21] L. CHEN, G. CHEN, AND F. WANG, *Recommender systems based on user reviews: the state of the art*, User Modeling and User-Adapted Interaction, 25 (2015), pp. 99–154.

[22] L. CHEN, D. YAN, AND F. WANG, *User evaluations on sentiment-based recommendation explanations*, ACM Transactions on Interactive Intelligent Systems (TiiS), 9 (2019), pp. 1–38.

[23] Q. CHEN, J. LI, Z. GUO, G. LI, AND Z. DENG, *Attribute-enhanced dual channel representation learning for session-based recommendation*, in Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, 2023, pp. 3793–3797.

[24] Q. CHEN, J. LIN, Y. ZHANG, M. DING, Y. CEN, H. YANG, AND J. TANG, *Towards knowledge-based recommender dialog system*, arXiv preprint arXiv:1908.05391, (2019).

[25] Q. CHEN, H. ZHAO, W. LI, P. HUANG, AND W. OU, *Behavior sequence transformer for e-commerce recommendation in alibaba*, in Proceedings of the 1st international workshop on deep learning practice for high-dimensional sparse data, 2019, pp. 1–4.

[26] R. CHEN, Q. HUA, Y.-S. CHANG, B. WANG, L. ZHANG, AND X. KONG, *A survey of collaborative filtering-based recommender systems: From traditional methods to hybrid methods based on social networks*, IEEE access, 6 (2018), pp. 64301–64320.

[27] T. CHEN AND R. C.-W. WONG, *Handling information loss of graph neural networks for session-based recommendation*, in Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 1172–1180.

[28] X. CHEN, S. JIA, AND Y. XIANG, *A review: Knowledge reasoning over knowledge graph*, Expert systems with applications, 141 (2020), p. 112948.

[29] Y. CHEN, J. CAO, Y. WANG, J. WU, H. CHEN, AND G. XU, *Causal variational inference for deconfounded multi-behavior recommendation*, ACM Transactions on Information Systems, (2025).

[30] Z. CHEN, X. WANG, X. XIE, M. PARSANA, A. SONI, X. AO, AND E. CHEN, *Towards explainable conversational recommendation*, in Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence, 2021, pp. 2994–3000.

[31] H. W. CHUNG, L. HOU, S. LONGPRE, B. ZOPH, Y. TAY, W. FEDUS, Y. LI, X. WANG, M. DEHGHANI, S. BRAHMA, ET AL., *Scaling instruction-finetuned language models*, Journal of Machine Learning Research, 25 (2024), pp. 1–53.

[32] J. CHUNG, C. GULCEHRE, K. CHO, AND Y. BENGIO, *Empirical evaluation of gated recurrent neural networks on sequence modeling*, arXiv preprint arXiv:1412.3555, (2014).

[33] J.-B. CORDONNIER, A. LOUKAS, AND M. JAGGI, *Multi-head attention: Collaborate instead of concatenate*, arXiv preprint arXiv:2006.16362, (2020).

[34] J. CORREA AND E. BAREINBOIM, *A calculus for stochastic interventions: Causal effect identification and surrogate experiments*, in Proceedings of the AAAI conference on artificial intelligence, vol. 34, 2020, pp. 10093–10100.

[35] P. COVINGTON, J. ADAMS, AND E. SARGIN, *Deep neural networks for youtube recommendations*, in Proceedings of the 10th ACM conference on recommender systems, 2016, pp. 191–198.

[36] Q. CUI, S. WU, Q. LIU, W. ZHONG, AND L. WANG, *Mv-rnn: A multi-view recurrent neural network for sequential recommendation*, IEEE Transactions on Knowledge and Data Engineering, 32 (2018), pp. 317–331.

[37] K. DAMAK, S. KHENISSI, AND O. NASRAOUI, *Debiasing the cloze task in sequential recommendation with bidirectional transformers*, in Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2022, pp. 273–282.

[38] J. DAVIS AND M. GOADRICH, *The relationship between precision-recall and roc curves*, in Proceedings of the 23rd international conference on Machine learning, 2006, pp. 233–240.

[39] L. DENG, D. LIAN, C. WU, AND E. CHEN, *Graph convolution network based recommender systems: Learning guarantee and item mixture powered strategy*, Advances in Neural Information Processing Systems, 35 (2022), pp. 3900–3912.

[40] Y. DENG, Y. LI, B. DING, AND W. LAM, *Leveraging long short-term user preference in conversational recommendation via multi-agent reinforcement learning*, IEEE Transactions on Knowledge and Data Engineering, 35 (2022), pp. 11541–11555.

[41] Y. DENG, Y. LI, F. SUN, B. DING, AND W. LAM, *Unified conversational recommendation policy learning via graph-based reinforcement learning*, in Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 1431–1441.

[42] M. DESHPANDE AND G. KARYPIS, *Item-based top-n recommendation algorithms*, ACM Transactions on Information Systems (TOIS), 22 (2004), pp. 143–177.

[43] P. DING AND F. LI, *Causal inference*, Statistical Science, 33 (2018), pp. 214–237.

[44] S. DING, F. FENG, X. HE, Y. LIAO, J. SHI, AND Y. ZHANG, *Causal incremental graph convolution for recommender system retraining*, IEEE Transactions on Neural Networks and Learning Systems, 35 (2022), pp. 4718–4728.

[45] H.-Q. DO, T.-H. LE, AND B. YOON, *Dynamic weighted hybrid recommender systems*, in 2020 22nd International Conference on Advanced Communication Technology (ICACT), IEEE, 2020, pp. 644–650.

[46] P. DOMINGOS, *The role of occam's razor in knowledge discovery*, Data mining and knowledge discovery, 3 (1999), pp. 409–425.

[47] T. DONKERS, B. LOEPP, AND J. ZIEGLER, *Sequential user-based recurrent neural network recommendations*, in Proceedings of the eleventh ACM conference on recommender systems, 2017, pp. 152–160.

[48] W. FAN, Y. MA, Q. LI, Y. HE, E. ZHAO, J. TANG, AND D. YIN, *Graph neural networks for social recommendation*, in The World Wide Web Conference, 2019, pp. 417–426.

[49] C. GAO, W. LEI, X. HE, M. DE RIJKE, AND T.-S. CHUA, *Advances and challenges in conversational recommender systems: A survey*, AI Open, 2 (2021), pp. 100–126.

[50] C. GAO, Y. ZHENG, N. LI, Y. LI, Y. QIN, J. PIAO, Y. QUAN, J. CHANG, D. JIN, X. HE, ET AL., *A survey of graph neural networks for recommender systems: Challenges, methods, and directions*, ACM Transactions on Recommender Systems, 1 (2023), pp. 1–51.

[51] C. GAO, Y. ZHENG, W. WANG, F. FENG, X. HE, AND Y. LI, *Causal inference in recommender systems: A survey and future directions*, ACM Transactions on Information Systems, 42 (2024), pp. 1–32.

[52] M. GÖKSEDEF AND Ş. GÜNDÜZ-ÖĞÜDÜCÜ, *Combination of web page recommender systems*, Expert systems with applications, 37 (2010), pp. 2911–2922.

[53] C. A. GOMEZ-URIBE AND N. HUNT, *The netflix recommender system: Algorithms, business value, and innovation*, ACM Transactions on Management Information Systems (TMIS), 6 (2015), pp. 1–19.

[54] S. GOSWAMI, C. MURTHY, AND A. K. DAS, *Sparsity measure of a network graph: Gini index*, Information Sciences, 462 (2018), pp. 16–39.

[55] N. GRGIC-HLACA, M. B. ZAFAR, K. P. GUMMADI, AND A. WELLER, *The case for process fairness in learning: Feature selection for fair decision making*, in NIPS symposium on machine learning and the law, vol. 1, 2016, p. 2.

[56] J. GU, Z. WANG, J. KUEN, L. MA, A. SHAHROUDY, B. SHUAI, T. LIU, X. WANG, G. WANG, J. CAI, ET AL., *Recent advances in convolutional neural networks*, Pattern recognition, 77 (2018), pp. 354–377.

[57] L. GUO, H. YIN, Q. WANG, T. CHEN, A. ZHOU, AND N. QUOC VIET HUNG, *Streaming session-based recommendation*, in Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, 2019, pp. 1569–1577.

[58] Y. HAGMAYER, S. A. SLOMAN, D. A. LAGNADO, AND M. R. WALDMANN, *Causal reasoning through intervention*, Causal learning: Psychology, philosophy, and computation, (2007), pp. 86–100.

[59] M. HAMMAR, R. KARLSSON, AND B. J. NILSSON, *Using maximum coverage to optimize recommendation systems in e-commerce*, in Proceedings of the 7th ACM conference on Recommender systems, 2013, pp. 265–272.

[60] K. HAN, J. GUO, C. ZHANG, AND M. ZHU, *Attribute-aware attention model for fine-grained representation learning*, in Proceedings of the 26th ACM international conference on Multimedia, 2018, pp. 2040–2048.

[61] N. HARIRI, B. MOBASHER, R. BURKE, AND Y. ZHENG, *Context-aware recommendation based on review mining.*, in ITWP@ IJCAI, 2011.

[62] J. HARTE, W. ZORGDRAGER, P. LOURIDAS, A. KATSIFODIMOS, D. JANNACH, AND M. FRAGKOULIS, *Leveraging large language models for sequential recommendation*, in Proceedings of the 17th ACM Conference on Recommender Systems, 2023, pp. 1096–1102.

[63] P. HASE, H. XIE, AND M. BANSAL, *The out-of-distribution problem in explainability and search methods for feature importance explanations*, Advances in neural information processing systems, 34 (2021), pp. 3650–3666.

[64] A. HAWALAH AND M. FASLI, *Dynamic user profiles for web personalisation*, Expert Systems with Applications, 42 (2015), pp. 2547–2569.

[65] X. HE, T. CHEN, M.-Y. KAN, AND X. CHEN, *Trirank: Review-aware explainable recommendation by modeling aspects*, in Proceedings of the 24th ACM international on conference on information and knowledge management, 2015, pp. 1661–1670.

[66] X. HE, L. LIAO, H. ZHANG, L. NIE, X. HU, AND T.-S. CHUA, *Neural collaborative filtering*, in Proceedings of the 26th international conference on world wide web, 2017, pp. 173–182.

[67] X. He, Y. Zhang, F. Feng, C. Song, L. Yi, G. Ling, and Y. Zhang, *Addressing confounding feature issue for causal recommendation*, ACM Transactions on Information Systems, 41 (2023), pp. 1–23.

[68] B. C. Herd and S. Miles, *Detecting causal relationships in simulation models using intervention-based counterfactual analysis*, ACM Transactions on Intelligent Systems and Technology (TIST), 10 (2019), pp. 1–25.

[69] T. Hester, M. Vecerik, O. Pietquin, M. Lanctot, T. Schaul, B. Piot, D. Horgan, J. Quan, A. Sendonaris, I. Osband, et al., *Deep q-learning from demonstrations*, in Proceedings of the AAAI conference on artificial intelligence, vol. 32, 2018.

[70] B. Hidasi, A. Karatzoglou, L. Baltrunas, and D. Tikk, *Session-based recommendations with recurrent neural networks*, arXiv preprint arXiv:1511.06939, (2015).

[71] T. O. Hodson, *Root mean square error (rmse) or mean absolute error (mae): When to use them or not*, Geoscientific Model Development Discussions, 2022 (2022), pp. 1–10.

[72] B. Hu, C. Shi, W. X. Zhao, and P. S. Yu, *Leveraging meta-path based context for top-n recommendation with a neural co-attention model*, in Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018, pp. 1531–1540.

[73] Z. Hu, Y. Dong, K. Wang, and Y. Sun, *Heterogeneous graph transformer*, in Proceedings of The Web Conference 2020, 2020, pp. 2704–2710.

[74] Z. Hu, Z. Zhao, X. Yi, T. Yao, L. Hong, Y. Sun, and E. Chi, *Improving multi-task generalization via regularizing spurious correlation*, Advances in Neural Information Processing Systems, 35 (2022), pp. 11450–11466.

[75] C. Huang, H. Xu, Y. Xu, P. Dai, L. Xia, M. Lu, L. Bo, H. Xing, X. Lai, and Y. Ye, *Knowledge-aware coupled graph neural network for social recommendation*, in Proceedings of the AAAI conference on artificial intelligence, vol. 35, 2021, pp. 4115–4122.

[76] Y. Huang, H. Liu, W. Li, Z. Wang, X. Hu, and W. Wang, *Lifestyles in amazon: Evidence from online reviews enhanced recommender system*, International Journal of Market Research, 62 (2020), pp. 689–706.

[77] N. J. Hurley, *Personalised ranking with diversity*, in Proceedings of the 7th ACM Conference on Recommender Systems, 2013, pp. 379–382.

[78] P. IZMAILOV, P. KIRICHENKO, N. GRUVER, AND A. G. WILSON, *On feature learning in the presence of spurious correlations,* Advances in Neural Information Processing Systems, 35 (2022), pp. 38516–38532.

[79] A. JABER, A. RIBEIRO, J. ZHANG, AND E. BAREINBOIM, *Causal identification under markov equivalence: calculus, algorithm, and completeness,* Advances in Neural Information Processing Systems, 35 (2022), pp. 3679–3690.

[80] D. JANNACH, A. MANZOOR, W. CAI, AND L. CHEN, *A survey on conversational recommender systems,* ACM Computing Surveys (CSUR), 54 (2021), pp. 1–36.

[81] U. JAVED, K. SHAUKAT, I. A. HAMEED, F. IQBAL, T. M. ALAM, AND S. LUO, *A review of content-based and context-based recommendation systems,* International Journal of Emerging Technologies in Learning (iJET), 16 (2021), pp. 274–306.

[82] M. JESSE, C. BAUER, AND D. JANNACH, *Intra-list similarity and human diversity perceptions of recommendations: the details matter,* User Modeling and User-Adapted Interaction, 33 (2023), pp. 769–802.

[83] K.-Y. JUNG, D.-H. PARK, AND J.-H. LEE, *Hybrid collaborative filtering and content-based filtering for improved recommender system,* in International Conference on Computational Science, Springer, 2004, pp. 295–302.

[84] W.-C. KANG AND J. MCAULEY, *Self-attentive sequential recommendation,* in 2018 IEEE international conference on data mining (ICDM), IEEE, 2018, pp. 197–206.

[85] G. KARYPIS, *Evaluation of item-based top-n recommendation algorithms,* in Proceedings of the tenth international conference on Information and knowledge management, 2001, pp. 247–254.

[86] B. M. KIM, Q. LI, C. S. PARK, S. G. KIM, AND J. Y. KIM, *A new approach for combining content-based and collaborative filters,* Journal of Intelligent Information Systems, 27 (2006), pp. 79–91.

[87] D. P. KINGMA, *Adam: A method for stochastic optimization,* arXiv preprint arXiv:1412.6980, (2014).

[88] T. N. KIPF AND M. WELLING, *Semi-supervised classification with graph convolutional networks,* arXiv preprint arXiv:1609.02907, (2016).

[89] N. KONDYLIDIS, J. ZOU, AND E. KANOULAS, *Category aware explainable conversational recommendation,* arXiv preprint arXiv:2103.08733, (2021).

[90] Y. KOREN, *Collaborative filtering with temporal dynamics*, in Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, 2009, pp. 447–456.

[91] A. KUMAR, A. DESHPANDE, AND A. SHARMA, *Causal effect regularization: automated detection and removal of spurious correlations*, Advances in Neural Information Processing Systems, 36 (2024).

[92] W. LEI, X. HE, M. DE RIJKE, AND T.-S. CHUA, *Conversational recommendation: Formulation, methods, and evaluation*, in Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020, pp. 2425–2428.

[93] W. LEI, X. HE, Y. MIAO, Q. WU, R. HONG, M.-Y. KAN, AND T.-S. CHUA, *Estimation-action-reflection: Towards deep interaction between conversational and recommender systems*, in Proceedings of the 13th International Conference on Web Search and Data Mining, 2020, pp. 304–312.

[94] W. LEI, G. ZHANG, X. HE, Y. MIAO, X. WANG, L. CHEN, AND T.-S. CHUA, *Interactive path reasoning on graph for conversational recommendation*, in Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining, 2020, pp. 2073–2083.

[95] A. LI, Z. CHENG, F. LIU, Z. GAO, W. GUAN, AND Y. PENG, *Disentangled graph neural networks for session-based recommendation*, arXiv preprint arXiv:2201.03482, (2022).

[96] J. LI, P. REN, Z. CHEN, Z. REN, T. LIAN, AND J. MA, *Neural attentive session-based recommendation*, in Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, 2017, pp. 1419–1428.

[97] J. LI, M. WANG, J. LI, J. FU, X. SHEN, J. SHANG, AND J. MCAULEY, *Text is all you need: Learning language representations for sequential recommendation*, in Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2023, pp. 1258–1267.

[98] L. LI, Y. ZHANG, D. LIU, AND L. CHEN, *Large language models for generative recommendation: A survey and visionary discussions*, arXiv preprint arXiv:2309.01157, (2023).

[99] Q. LI, X. WANG, Z. WANG, AND G. XU, *Be causal: De-biasing social network confounding in recommendation*, ACM Transactions on Knowledge Discovery from Data, 17 (2023), pp. 1–23.

[100] Q. LI, X. WANG, AND G. XU, *Be causal: De-biasing social network confounding in recommendation*, arXiv preprint arXiv:2105.07775, (2021).

[101] Q. LI, Z. WANG, G. LI, J. PANG, AND G. XU, *Hilbert sinkhorn divergence for optimal transport*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 3835–3844.

[102] R. LI, S. EBRAHIMI KAHOU, H. SCHULZ, V. MICHALSKI, L. CHARLIN, AND C. PAL, *Towards deep conversational recommendations*, Advances in neural information processing systems, 31 (2018).

[103] Y. LI, T. CHEN, AND Z. HUANG, *Attribute-aware explainable complementary clothing recommendation*, World Wide Web, 24 (2021), pp. 1885–1901.

[104] Z. LI, F. LIU, W. YANG, S. PENG, AND J. ZHOU, *A survey of convolutional neural networks: analysis, applications, and prospects*, IEEE transactions on neural networks and learning systems, 33 (2021), pp. 6999–7019.

[105] Z. LI, C. YANG, Y. CHEN, X. WANG, H. CHEN, G. XU, L. YAO, AND M. SHENG, *Graph and sequential neural networks in session-based recommendation: A survey*, ACM Computing Surveys, (2024).

[106] A. LIN, Z. ZHU, J. WANG, AND J. CAVERLEE, *Enhancing user personalization in conversational recommenders*, arXiv preprint arXiv:2302.06656, (2023).

[107] J. LIN, R. SHAN, C. ZHU, K. DU, B. CHEN, S. QUAN, R. TANG, Y. YU, AND W. ZHANG, *Rella: Retrieval-enhanced large language models for lifelong sequential behavior comprehension in recommendation*, arXiv preprint arXiv:2308.11131, (2023).

[108] C. LIU, X. LI, G. CAI, Z. DONG, H. ZHU, AND L. SHANG, *Noninvasive self-attention for side information fusion in sequential recommendation*, in Proceedings of the AAAI conference on artificial intelligence, vol. 35, 2021, pp. 4249–4256.

[109] D. LIU, J. LI, B. DU, J. CHANG, R. GAO, AND Y. WU, *A hybrid neural network approach to combine textual information and rating information for item recommendation*, Knowledge and Information Systems, 63 (2021), pp. 621–646.

[110] Q. LIU, Y. ZENG, R. MOKHOSI, AND H. ZHANG, *Stamp: short-term attention/memory priority model for session-based recommendation*, in Proceedings of the 24th ACM

SIGKDD international conference on knowledge discovery & data mining, 2018, pp. 1831–1839.

[111] S. LIU, X. SONG, Z. MA, E. D. GANAA, AND X. SHEN, *More: multi-output residual embedding for multi-label classification*, Pattern Recognition, 126 (2022), p. 108584.

[112] Z. LIU, Y. FANG, AND M. WU, *Estimating propensity for causality-based recommendation without exposure data*, Advances in Neural Information Processing Systems, 36 (2024).

[113] H. LUO, F. ZHUANG, R. XIE, H. ZHU, D. WANG, Z. AN, AND Y. XU, *A survey on causal inference for recommendation*, The Innovation, (2024).

[114] H. MA, *An experimental study on implicit social recommendation*, in Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, 2013, pp. 73–82.

[115] J. MA, C. ZHOU, P. CUI, H. YANG, AND W. ZHU, *Learning disentangled representations for recommendation*, arXiv preprint arXiv:1910.14238, (2019).

[116] S. MCDONNELL, O. NADA, M. R. ABID, AND E. AMJADIAN, *Cyberbert: a deep dynamic-state session-based recommender system for cyber threat recognition*, in 2021 IEEE aerospace conference (50100), IEEE, 2021, pp. 1–12.

[117] S. E. MIDDLETON, N. R. SHADBOLT, AND D. C. DE ROURE, *Ontological user profiling in recommender systems*, ACM Transactions on Information Systems (TOIS), 22 (2004), pp. 54–88.

[118] Y. MING, H. YIN, AND Y. LI, *On the impact of spurious correlation for out-of-distribution detection*, in Proceedings of the AAAI conference on artificial intelligence, vol. 36, 2022, pp. 10051–10059.

[119] M. MINICI, F. CINUS, C. MONTI, F. BONCHI, AND G. MANCO, *Cascade-based echo chamber detection*, in Proceedings of the 31st ACM International Conference on Information & Knowledge Management, 2022, pp. 1511–1520.

[120] M. MORIK, A. SINGH, J. HONG, AND T. JOACHIMS, *Controlling fairness and bias in dynamic learning-to-rank*, in Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020, pp. 429–438.

[121] S. MU, Y. LI, W. X. ZHAO, J. WANG, B. DING, AND J.-R. WEN, *Alleviating spurious correlations in knowledge-aware recommendations through counterfactual gener-*

ator, in Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022, pp. 1401–1411.

[122] A. H. NABIZADEH, J. P. LEAL, H. N. RAFSANJANI, AND R. R. SHAH, *Learning path personalization and recommendation methods: A survey of the state-of-the-art*, Expert Systems with Applications, 159 (2020), p. 113596.

[123] M. M. NAJAFABADI, F. VILLANUSTRE, T. M. KHOSHGOFTAAR, N. SELIYA, R. WALD, AND E. MUHAREMAGIC, *Deep learning applications and challenges in big data analytics*, Journal of big data, 2 (2015), pp. 1–21.

[124] J. NI, Z. HUANG, Y. HU, AND C. LIN, *A two-stage embedding model for recommendation with multimodal auxiliary information*, Information Sciences, 582 (2022), pp. 22–37.

[125] M. O'MAHONY, N. HURLEY, N. KUSHMERICK, AND G. SILVESTRE, *Collaborative recommendation: A robustness analysis*, ACM Transactions on Internet Technology (TOIT), 4 (2004), pp. 344–377.

[126] S. PAN, D. LI, H. GU, T. LU, X. LUO, AND N. GU, *Accurate and explainable recommendation via review rationalization*, in Proceedings of the ACM Web Conference 2022, 2022, pp. 3092–3101.

[127] Y. PANG, L. WU, Q. SHEN, Y. ZHANG, Z. WEI, F. XU, E. CHANG, B. LONG, AND J. PEI, *Heterogeneous global graph neural networks for personalized session-based recommendation*, in Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, 2022, pp. 775–783.

[128] M. J. PAZZANI AND D. BILLSUS, *Content-based recommendation systems*, in The adaptive web: methods and strategies of web personalization, Springer, 2007, pp. 325–341.

[129] J. PEARL, *Causal inference in statistics: An overview*, (2009).

[130] S. A. PUTHIYA PARAMBATH, N. USUNIER, AND Y. GRANDVALET, *A coverage-based approach to recommendation diversity on similarity graph*, in Proceedings of the 10th ACM Conference on Recommender Systems, 2016, pp. 15–22.

[131] R. QIU, S. WANG, Z. CHEN, H. YIN, AND Z. HUANG, *Causalrec: Causal inference for visual debiasing in visually-aware recommendation*, in Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 3844–3852.

[132] M. QUADRANA, P. CREMONESI, AND D. JANNACH, *Sequence-aware recommender systems*, ACM computing surveys (CSUR), 51 (2018), pp. 1–36.

[133] M. QUADRANA, A. KARATZOGLOU, B. HIDASI, AND P. CREMONESI, *Personalizing session-based recommendations with hierarchical recurrent neural networks*, in proceedings of the Eleventh ACM Conference on Recommender Systems, 2017, pp. 130–137.

[134] R. REN, Z. LIU, Y. LI, W. X. ZHAO, H. WANG, B. DING, AND J.-R. WEN, *Sequential recommendation with self-attentive multi-adversarial network*, in Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval, 2020, pp. 89–98.

[135] X. REN, W. WEI, L. XIA, L. SU, S. CHENG, J. WANG, D. YIN, AND C. HUANG, *Representation learning with large language models for recommendation*, arXiv preprint arXiv:2310.15950, (2023).

[136] X. REN, H. YIN, T. CHEN, H. WANG, Z. HUANG, AND K. ZHENG, *Learning to ask appropriate questions in conversational recommendation*, in Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval, 2021, pp. 808–817.

[137] S. RENDLE, *Factorization machines*, in 2010 IEEE International conference on data mining, IEEE, 2010, pp. 995–1000.

[138] S. RENDLE, C. FREUDENTHALER, Z. GANTNER, AND L. SCHMIDT-THIEME, *Bpr: Bayesian personalized ranking from implicit feedback*, arXiv preprint arXiv:1205.2618, (2012).

[139] S. RENDLE, W. KRICHENE, L. ZHANG, AND J. ANDERSON, *Neural collaborative filtering vs. matrix factorization revisited*, in Proceedings of the 14th ACM Conference on Recommender Systems, 2020, pp. 240–248.

[140] S. REZA, M. C. FERREIRA, J. J. MACHADO, AND J. M. R. TAVARES, *A multi-head attention-based transformer model for traffic flow forecasting with a comparative analysis to recurrent neural networks*, Expert Systems with Applications, 202 (2022), p. 117275.

[141] Y. SAITO AND T. JOACHIMS, *Counterfactual learning and evaluation for recommender systems: Foundations, implementations, and recent advances*, in Proceedings of the 15th ACM Conference on Recommender Systems, 2021, pp. 828–830.

[142] Y. SAITO, S. YAGINUMA, Y. NISHINO, H. SAKATA, AND K. NAKATA, *Unbiased recommender learning from missing-not-at-random implicit feedback*, in Proceedings

of the 13th International Conference on Web Search and Data Mining, 2020, pp. 501–509.

[143] I. H. SARKER, *Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions*, SN computer science, 2 (2021), p. 420.

[144] B. SARWAR, G. KARYPIS, J. KONSTAN, AND J. RIEDL, *Item-based collaborative filtering recommendation algorithms*, in Proceedings of the 10th international conference on World Wide Web, 2001, pp. 285–295.

[145] K. SASAHARA, W. CHEN, H. PENG, G. L. CIAMPAGLIA, A. FLAMMINI, AND F. MENCZER, *Social influence and unfollowing accelerate the emergence of echo chambers*, Journal of Computational Social Science, 4 (2021), pp. 381–402.

[146] M. SATO, S. TAKEMORI, J. SINGH, AND T. OHKUMA, *Unbiased learning for the causal effect of recommendation*, in Proceedings of the 14th ACM conference on recommender systems, 2020, pp. 378–387.

[147] J. B. SCHAFER, J. A. KONSTAN, AND J. RIEDL, *E-commerce recommendation applications*, Data mining and knowledge discovery, 5 (2001), pp. 115–153.

[148] Y. SHI, M. LARSON, AND A. HANJALIC, *Collaborative filtering beyond the user-item matrix: A survey of the state of the art and future challenges*, ACM Computing Surveys (CSUR), 47 (2014), pp. 1–45.

[149] H. A. SIMON, *Spurious correlation: A causal interpretation*, Journal of the American statistical Association, 49 (1954), pp. 467–479.

[150] M. R. SMITH AND G. P. ALPERT, *Explaining police bias: A theory of social conditioning and illusory correlation*, Criminal justice and behavior, 34 (2007), pp. 1262–1283.

[151] P. SOMERFIELD, K. CLARKE, AND R. WARWICK, *Simpson index*, in Encyclopedia of ecology, Elsevier, 2008, pp. 3252–3255.

[152] J. SON AND S. B. KIM, *Content-based filtering for recommendation systems using multiattribute networks*, Expert Systems with Applications, 89 (2017), pp. 404–412.

[153] W. SONG, Z. XIAO, Y. WANG, L. CHARLIN, M. ZHANG, AND J. TANG, *Session-based social recommendation via dynamic graph attention networks*, in Proceedings of the Twelfth ACM international conference on web search and data mining, 2019, pp. 555–563.

[154] G. SUGANESHWARI AND S. SYED IBRAHIM, *A survey on collaborative filtering based recommendation system*, in Proceedings of the 3rd international symposium on

big data and cloud computing challenges (ISBCC–16'), Springer, 2016, pp. 503–518.

[155] Y. SUI, X. WANG, J. WU, M. LIN, X. HE, AND T.-S. CHUA, *Causal attention for interpretable and generalizable graph classification*, in Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2022, pp. 1696–1705.

[156] F. SUN, J. LIU, J. WU, C. PEI, X. LIN, W. OU, AND P. JIANG, *Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer*, in Proceedings of the 28th ACM international conference on information and knowledge management, 2019, pp. 1441–1450.

[157] Y. SUN AND Y. ZHANG, *Conversational recommender system*, in The 41st international acm sigir conference on research & development in information retrieval, 2018, pp. 235–244.

[158] J. TAN, S. XU, Y. GE, Y. LI, X. CHEN, AND Y. ZHANG, *Counterfactual explainable recommendation*, in Proceedings of the 30th ACM International Conference on Information & Knowledge Management, 2021, pp. 1784–1793.

[159] J. TANG, X. HU, AND H. LIU, *Social recommendation: a review*, Social Network Analysis and Mining, 3 (2013), pp. 1113–1133.

[160] S. TANG, Q. LI, D. WANG, C. GAO, W. XIAO, D. ZHAO, Y. JIANG, Q. MA, AND A. ZHANG, *Counterfactual video recommendation for duration debiasing*, in Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2023, pp. 4894–4903.

[161] Y. TAO, M. GAO, J. YU, Z. WANG, Q. XIONG, AND X. WANG, *Predictive and contrastive: Dual-auxiliary learning for recommendation*, IEEE Transactions on Computational Social Systems, 10 (2022), pp. 2254–2265.

[162] R. TAORI, I. GULRAJANI, T. ZHANG, Y. DUBOIS, X. LI, C. GUESTRIN, P. LIANG, AND T. B. HASHIMOTO, *Stanford alpaca: An instruction-following llama model*, 2023.

[163] P. B. THORAT, R. M. GOUDAR, AND S. BARVE, *Survey on collaborative filtering, content-based filtering and hybrid recommendation system*, International Journal of Computer Applications, 110 (2015), pp. 31–36.

[164] S. TIKKA, A. HYTTINEN, AND J. KARVANEN, *Identifying causal effects via context-specific independence relations*, Advances in neural information processing systems, 32 (2019).

[165] A. TOMMASEL, J. M. RODRIGUEZ, AND D. GODOY, *I want to break free! recommending friends from outside the echo chamber*, in Proceedings of the 15th ACM Conference on Recommender Systems, 2021, pp. 23–33.

[166] H. TOUVRON, L. MARTIN, K. STONE, P. ALBERT, A. ALMAHAIRI, Y. BABAEI, N. BASHLYKOV, S. BATRA, P. BHARGAVA, S. BHOSALE, ET AL., *Llama 2: Open foundation and fine-tuned chat models*, arXiv preprint arXiv:2307.09288, (2023).

[167] K. H. TRAN, A. GHAZIMATIN, AND R. SAHA ROY, *Counterfactual explanations for neural recommenders*, in Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 1627–1631.

[168] V. A. TRAN, G. SALHA-GALVAN, B. SGUERRA, AND R. HENNEQUIN, *Attention mixtures for time-aware sequential recommendation*, in Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2023, pp. 1821–1826.

[169] A. VALL, M. DORFER, H. EGHBAL-ZADEH, M. SCHEDL, K. BURJORJEE, AND G. WIDMER, *Feature-combination hybrid recommender systems for automated music playlist continuation*, User Modeling and User-Adapted Interaction, 29 (2019), pp. 527–572.

[170] T. VAN ERVEN AND P. HARREMOS, *Rényi divergence and kullback-leibler divergence*, IEEE Transactions on Information Theory, 60 (2014), pp. 3797–3820.

[171] R. VAN METEREN AND M. VAN SOMEREN, *Using content-based filtering for recommendation*, in Proceedings of the machine learning in the new information age: MLnet/ECML2000 workshop, vol. 30, Barcelona, 2000, pp. 47–56.

[172] A. VENDEVILLE, A. GIOVANIDIS, E. PAPANASTASIOU, AND B. GUEDJ, *Opening up echo chambers via optimal content recommendation*, in International Conference on Complex Networks and Their Applications, Springer, 2022, pp. 74–85.

[173] N. M. VILLEGAS, C. SÁNCHEZ, J. DÍAZ-CELY, AND G. TAMURA, *Characterizing context-aware recommender systems: A systematic literature review*, Knowledge-Based Systems, 140 (2018), pp. 173–200.

[174] D. WANG, Y. LIANG, D. XU, X. FENG, AND R. GUAN, *A content-based recommender system for computer science publications*, Knowledge-based systems, 157 (2018), pp. 1–9.

[175] H. WANG, N. WANG, AND D.-Y. YEUNG, *Collaborative deep learning for recommender systems*, in Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, 2015, pp. 1235–1244.

[176] J. WANG AND Y. ZHANG, *Opportunity model for e-commerce recommendation: right product; right time*, in Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, 2013, pp. 303–312.

[177] S. WANG, L. CAO, Y. WANG, Q. Z. SHENG, M. A. ORGUN, AND D. LIAN, *A survey on session-based recommender systems*, ACM Computing Surveys (CSUR), 54 (2021), pp. 1–38.

[178] S. WANG, L. HU, Y. WANG, L. CAO, Q. Z. SHENG, AND M. ORGUN, *Sequential recommender systems: challenges, progress and prospects*, arXiv preprint arXiv:2001.04830, (2019).

[179] S. WANG, Q. ZHANG, L. HU, X. ZHANG, Y. WANG, AND C. AGGARWAL, *Sequential/session-based recommendations: Challenges, approaches, applications and opportunities*, in Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval, 2022, pp. 3425–3428.

[180] T. WANG, J. HUANG, H. ZHANG, AND Q. SUN, *Visual commonsense r-cnn*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10760–10770.

[181] W. WANG, F. FENG, X. HE, X. WANG, AND T.-S. CHUA, *Deconfounded recommendation for alleviating bias amplification*, in Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining, 2021, pp. 1717–1725.

[182] W. WANG, F. FENG, X. HE, H. ZHANG, AND T.-S. CHUA, *Clicks can be cheating: Counterfactual recommendation for mitigating clickbait issue*, in Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 1288–1297.

[183] X. WANG, X. HE, M. WANG, F. FENG, AND T.-S. CHUA, *Neural graph collaborative filtering*, in Proceedings of the 42nd international ACM SIGIR conference on Research and development in Information Retrieval, 2019, pp. 165–174.

[184] X. WANG, H. JI, C. SHI, B. WANG, Y. YE, P. CUI, AND P. S. YU, *Heterogeneous graph attention network*, in The world wide web conference, 2019, pp. 2022–2032.

[185] X. WANG, H. JIN, A. ZHANG, X. HE, T. XU, AND T.-S. CHUA, *Disentangled graph collaborative filtering*, in Proceedings of the 43rd International ACM SIGIR Con-

ference on Research and Development in Information Retrieval, 2020, pp. 1001–1010.

[186] X. WANG, Q. LI, D. YU, P. CUI, Z. WANG, AND G. XU, *Causal disentanglement for semantic-aware intent learning in recommendation*, IEEE Transactions on Knowledge and Data Engineering, 35 (2022), pp. 9836–9849.

[187] X. WANG, Q. LI, D. YU, W. HUANG, AND G. XU, *Causal neural graph collaborative filtering*, arXiv preprint arXiv:2307.04384, (2023).

[188] X. WANG, Q. LI, D. YU, Z. WANG, H. CHEN, AND G. XU, *Mgpolicy: Meta graph enhanced off-policy learning for recommendations*, in Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022, pp. 1369–1378.

[189] X. WANG, D. WANG, C. XU, X. HE, Y. CAO, AND T.-S. CHUA, *Explainable reasoning over knowledge graphs for recommendation*, in Proceedings of the AAAI conference on artificial intelligence, vol. 33, 2019, pp. 5329–5336.

[190] X. WANG AND Y. WANG, *Improving content-based and hybrid music recommendation using deep learning*, in Proceedings of the 22nd ACM international conference on Multimedia, 2014, pp. 627–636.

[191] Y. WANG, H. GUO, B. CHEN, W. LIU, Z. LIU, Q. ZHANG, Z. HE, H. ZHENG, W. YAO, M. ZHANG, ET AL., *Causalint: Causal inspired intervention for multi-scenario recommendation*, in Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2022, pp. 4090–4099.

[192] Y. WANG, D. LIANG, L. CHARLIN, AND D. M. BLEI, *Causal inference for recommender systems*, in Proceedings of the 14th ACM Conference on Recommender Systems, 2020, pp. 426–431.

[193] Z. WANG, S. SHEN, Z. WANG, B. CHEN, X. CHEN, AND J.-R. WEN, *Unbiased sequential recommendation with latent confounders*, in Proceedings of the ACM Web Conference 2022, 2022, pp. 2195–2204.

[194] Z. WANG, W. WEI, G. CONG, X.-L. LI, X.-L. MAO, AND M. QIU, *Global context enhanced graph neural networks for session-based recommendation*, in Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval, 2020, pp. 169–178.

[195] Z. WANG, J. ZHANG, H. XU, X. CHEN, Y. ZHANG, W. X. ZHAO, AND J.-R. WEN, *Counterfactual data-augmented sequential recommendation*, in Proceedings of the

44th international ACM SIGIR conference on research and development in information retrieval, 2021, pp. 347–356.

[196] T. WEI, F. FENG, J. CHEN, Z. WU, J. YI, AND X. HE, *Model-agnostic counterfactual reasoning for eliminating popularity bias in recommender system*, in Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining, 2021, pp. 1791–1800.

[197] Y. WEI, X. WANG, L. NIE, S. LI, D. WANG, AND T.-S. CHUA, *Causal inference for knowledge graph based recommendation*, IEEE Transactions on Knowledge and Data Engineering, 35 (2022), pp. 11153–11164.

[198] L. WU, X. HE, X. WANG, K. ZHANG, AND M. WANG, *A survey on accuracy-oriented neural recommendation: From collaborative filtering to information-rich recommendation*, IEEE Transactions on Knowledge and Data Engineering, 35 (2022), pp. 4425–4445.

[199] L. WU, S. LI, C.-J. HSIEH, AND J. SHARPNACK, *Sse-pt: Sequential recommendation via personalized transformer*, in Proceedings of the 14th ACM conference on recommender systems, 2020, pp. 328–337.

[200] L. WU, Q. LIU, E. CHEN, N. J. YUAN, G. GUO, AND X. XIE, *Relevance meets coverage: A unified framework to generate diversified recommendations*, ACM Transactions on Intelligent Systems and Technology (TIST), 7 (2016), pp. 1–30.

[201] L. WU, C. QUAN, C. LI, Q. WANG, B. ZHENG, AND X. LUO, *A context-aware user-item representation learning for item recommendation*, ACM Transactions on Information Systems (TOIS), 37 (2019), pp. 1–29.

[202] S. WU, F. SUN, W. ZHANG, X. XIE, AND B. CUI, *Graph neural networks in recommender systems: a survey*, ACM Computing Surveys, 55 (2022), pp. 1–37.

[203] S. WU, Y. TANG, Y. ZHU, L. WANG, X. XIE, AND T. TAN, *Session-based recommendation with graph neural networks*, in Proceedings of the AAAI conference on artificial intelligence, vol. 33, 2019, pp. 346–353.

[204] T. WU, S. HE, J. LIU, S. SUN, K. LIU, Q.-L. HAN, AND Y. TANG, *A brief overview of chatgpt: The history, status quo and potential future development*, IEEE/CAA Journal of Automatica Sinica, 10 (2023), pp. 1122–1136.

[205] Y. WU, G. ZHAO, M. LI, Z. ZHANG, AND X. QIAN, *Reason generation for point of interest recommendation via a hierarchical attention-based transformer model*, IEEE Transactions on Multimedia, (2023).

[206] Z. WU, S. PAN, F. CHEN, G. LONG, C. ZHANG, AND S. Y. PHILIP, *A comprehensive survey on graph neural networks*, IEEE transactions on neural networks and learning systems, 32 (2020), pp. 4–24.

[207] L. XIA, C. HUANG, Y. XU, P. DAI, X. ZHANG, H. YANG, J. PEI, AND L. BO, *Knowledge-enhanced hierarchical graph transformer network for multi-behavior recommendation*, in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, 2021, pp. 4486–4493.

[208] Y. XIAN, Z. FU, S. MUTHUKRISHNAN, G. DE MELO, AND Y. ZHANG, *Reinforcement knowledge graph reasoning for explainable recommendation*, in Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval, 2019, pp. 285–294.

[209] Y. XIAN, T. ZHAO, J. LI, J. CHAN, A. KAN, J. MA, X. L. DONG, C. FALOUTSOS, G. KARYPIS, S. MUTHUKRISHNAN, ET AL., *Ex3: Explainable attribute-aware item-set recommendations*, in Proceedings of the 15th ACM Conference on Recommender Systems, 2021, pp. 484–494.

[210] M. XIE, H. YIN, H. WANG, F. XU, W. CHEN, AND S. WANG, *Learning graph-based poi embedding for location-based recommendation*, in Proceedings of the 25th ACM international on conference on information and knowledge management, 2016, pp. 15–24.

[211] X. XIN, J. YANG, H. WANG, J. MA, P. REN, H. LUO, X. SHI, Z. CHEN, AND Z. REN, *On the user behavior leakage from recommender system exposure*, ACM Transactions on Information Systems, 41 (2023), pp. 1–25.

[212] C. XU, J. XU, X. CHEN, Z. DONG, AND J.-R. WEN, *Dually enhanced propensity score estimation in sequential recommendation*, in Proceedings of the 31st ACM International Conference on Information & Knowledge Management, 2022, pp. 2260–2269.

[213] H. XU, Y. XU, C. LI, AND F. ZHUANG, *Causal structure representation learning of unobserved confounders in latent space for recommendation*, ACM Transactions on Information Systems, (2025).

[214] J. XU, X. HE, AND H. LI, *Deep learning for matching in search and recommendation*, in The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, 2018, pp. 1365–1368.

[215] J. XU, Z. LI, B. DU, M. ZHANG, AND J. LIU, *Reluplex made more practical: Leaky relu*, in 2020 IEEE Symposium on Computers and communications (ISCC), IEEE, 2020, pp. 1–7.

[216] J. XU, X. SUN, Z. ZHANG, G. ZHAO, AND J. LIN, *Understanding and improving layer normalization*, Advances in neural information processing systems, 32 (2019).

[217] K. XU, J. YANG, J. XU, S. GAO, J. GUO, AND J.-R. WEN, *Adapting user preference to online feedback in multi-round conversational recommendation*, in Proceedings of the 14th ACM international conference on web search and data mining, 2021, pp. 364–372.

[218] M. XU, F. LIU, AND W. XU, *A survey on sequential recommendation*, in 2019 6th International Conference on Information Science and Control Engineering (ICISCE), IEEE, 2019, pp. 106–111.

[219] S. XU, J. TAN, Z. FU, J. JI, S. HEINECKE, AND Y. ZHANG, *Dynamic causal collaborative filtering*, in Proceedings of the 31st ACM International Conference on Information & Knowledge Management, 2022, pp. 2301–2310.

[220] S. XU, J. TAN, S. HEINECKE, V. J. LI, AND Y. ZHANG, *Deconfounded causal collaborative filtering*, ACM Transactions on Recommender Systems, 1 (2023), pp. 1–25.

[221] H.-J. XUE, X. DAI, J. ZHANG, S. HUANG, AND J. CHEN, *Deep matrix factorization models for recommender systems.*, in IJCAI, vol. 17, Melbourne, Australia, 2017, pp. 3203–3209.

[222] R. K. YADAV, J. LEI, O.-C. GRANMO, AND M. GOODWIN, *Robust interpretable text classification against spurious correlations using and-rules with negation*, in IJCAI International Joint Conference on Artificial Intelligence, International Joint Conferences on Artificial Intelligence, 2022.

[223] D. YANG, J. HE, H. QIN, Y. XIAO, AND W. WANG, *A graph-based recommendation across heterogeneous domains*, in proceedings of the 24th ACM International on Conference on Information and Knowledge Management, 2015, pp. 463–472.

[224] H. YANG, X. ZHAO, Y. LI, H. CHEN, AND G. XU, *An empirical study towards prompt-tuning for graph contrastive pre-training in recommendations*, Advances in Neural Information Processing Systems, 36 (2024).

[225] J. YANG, Y. DING, Y. WANG, P. REN, Z. CHEN, F. CAI, J. MA, R. ZHANG, Z. REN, AND X. XIN, *Debiasing sequential recommenders through distributionally robust op-*

timization over system exposure, in Proceedings of the 17th ACM International Conference on Web Search and Data Mining, 2024, pp. 882–890.

[226] X. YANG, F. FENG, W. JI, M. WANG, AND T.-S. CHUA, *Deconfounded video moment retrieval with causal intervention*, in Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval, 2021, pp. 1–10.

[227] Y. YANG, M. LI, X. HU, G. PAN, W. HUANG, J. WANG, AND Y. WANG, *Exploring exposure bias in recommender systems from causality perspective*, in 2021 IEEE 21st International Conference on Software Quality, Reliability and Security Companion (QRS-C), IEEE, 2021, pp. 425–432.

[228] Z. YANG, M. DING, X. ZOU, J. TANG, B. XU, C. ZHOU, AND H. YANG, *Region or global? a principle for negative sampling in graph-based recommendation*, IEEE Transactions on Knowledge and Data Engineering, 35 (2022), pp. 6264–6277.

[229] L. YAO, Z. CHU, S. LI, Y. LI, J. GAO, AND A. ZHANG, *A survey on causal inference*, ACM Transactions on Knowledge Discovery from Data (TKDD), 15 (2021), pp. 1–46.

[230] J. YI, X. REN, AND Z. CHEN, *Multi-auxiliary augmented collaborative variational auto-encoder for tag recommendation*, ACM Transactions on Information Systems, 41 (2023), pp. 1–25.

[231] R. YING, R. HE, K. CHEN, P. EKSOMBATCHAI, W. L. HAMILTON, AND J. LESKOVEC, *Graph convolutional neural networks for web-scale recommender systems*, in Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018, pp. 974–983.

[232] D. YU, Q. LI, X. WANG, Q. LI, AND G. XU, *Counterfactual explainable conversational recommendation*, IEEE Transactions on Knowledge and Data Engineering, (2023).

[233] D. YU, Q. LI, X. WANG, AND G. XU, *Deconfounded recommendation via causal intervention*, Neurocomputing, 529 (2023), pp. 128–139.

[234] D. YU, Q. LI, H. YIN, AND G. XU, *Causality-guided graph learning for session-based recommendation*, in Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, 2023, pp. 3083–3093.

[235] J. YU, M. GAO, J. LI, H. YIN, AND H. LIU, *Adaptive implicit friends identification over heterogeneous network for social recommendation*, in Proceedings of the

27th ACM international conference on information and knowledge management, 2018, pp. 357–366.

[236] K. YU, A. SCHWAIGHOFER, V. TRESP, X. XU, AND H.-P. KRIEGEL, *Probabilistic memory-based collaborative filtering*, IEEE Transactions on Knowledge and Data Engineering, 16 (2004), pp. 56–69.

[237] Y. YU, X. SI, C. HU, AND J. ZHANG, *A review of recurrent neural networks: Lstm cells and network architectures*, Neural computation, 31 (2019), pp. 1235–1270.

[238] H. YUAN, H. YU, S. GUI, AND S. JI, *Explainability in graph neural networks: A taxonomic survey*, IEEE Transactions on Pattern Analysis and Machine Intelligence, (2022).

[239] D. ZHANG, H. ZHANG, J. TANG, X.-S. HUA, AND Q. SUN, *Causal intervention for weakly-supervised semantic segmentation*, Advances in Neural Information Processing Systems, 33 (2020), pp. 655–666.

[240] J. ZHANG, C. GAO, D. JIN, AND Y. LI, *Group-buying recommendation for social e-commerce*, in 2021 IEEE 37th International Conference on Data Engineering (ICDE), IEEE, 2021, pp. 1536–1547.

[241] M. ZHANG, S. WU, M. GAO, X. JIANG, K. XU, AND L. WANG, *Personalized graph neural networks with attention mechanism for session-aware recommendation*, IEEE Transactions on Knowledge and Data Engineering, (2020).

[242] S. ZHANG, Z. JIANG, J. YAO, F. FENG, K. KUANG, Z. ZHAO, S. LI, H. YANG, T.-S. CHUA, AND F. WU, *Causal distillation for alleviating performance heterogeneity in recommender systems*, IEEE Transactions on Knowledge and Data Engineering, 36 (2023), pp. 459–474.

[243] S. ZHANG, D. YAO, Z. ZHAO, T.-S. CHUA, AND F. WU, *Causerec: Counterfactual user sequence synthesis for sequential recommendation*, in Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 367–377.

[244] S. ZHANG, L. YAO, A. SUN, AND Y. TAY, *Deep learning based recommender system: A survey and new perspectives*, ACM computing surveys (CSUR), 52 (2019), pp. 1–38.

[245] W. ZHANG, Z. CHEN, H. ZHA, AND J. WANG, *Learning from substitutable and complementary relations for graph-based sequential product recommendation*, ACM Transactions on Information Systems (TOIS), 40 (2021), pp. 1–28.

[246] Y. ZHANG, X. CHEN, Q. AI, L. YANG, AND W. B. CROFT, *Towards conversational search and recommendation: System ask, user respond*, in Proceedings of the 27th acm international conference on information and knowledge management, 2018, pp. 177–186.

[247] Y. ZHANG, X. CHEN, ET AL., *Explainable recommendation: A survey and new perspectives*, Foundations and Trends® in Information Retrieval, 14 (2020), pp. 1–101.

[248] Y. ZHANG, F. FENG, X. HE, T. WEI, C. SONG, G. LING, AND Y. ZHANG, *Causal intervention for leveraging popularity bias in recommendation*, in Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval, 2021, pp. 11–20.

[249] Y. ZHANG, L. WU, Q. SHEN, Y. PANG, Z. WEI, F. XU, E. CHANG, AND B. LONG, *Graph learning augmented heterogeneous graph neural network for social recommendation*, ACM Transactions on Recommender Systems, 1 (2023), pp. 1–22.

[250] Z. ZHANG, Q. DAI, X. CHEN, Z. DONG, AND R. TANG, *Robust causal inference for recommender system to overcome noisy confounders*, in Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2023, pp. 2349–2353.

[251] Z. ZHANG, M. LIN, E. DAI, AND S. WANG, *Rethinking graph backdoor attacks: A distribution-preserving perspective*, in Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2024, pp. 4386–4397.

[252] K. ZHAO, X. WANG, Y. ZHANG, L. ZHAO, Z. LIU, C. XING, AND X. XIE, *Leveraging demonstrations for reinforcement recommendation reasoning over knowledge graphs*, in Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval, 2020, pp. 239–248.

[253] Y. ZHENG, C. GAO, X. LI, X. HE, Y. LI, AND D. JIN, *Disentangling user interest and conformity for recommendation with causal embedding*, in Proceedings of the Web Conference 2021, 2021, pp. 2980–2991.

[254] J. ZHOU, E. AGICHTEIN, AND S. KALLUMADI, *Diversifying multi-aspect search results using simpson's diversity index*, in Proceedings of the 29th ACM International conference on information & knowledge management, 2020, pp. 2345–2348.

[255] X. ZHU, Y. ZHANG, X. YANG, D. WANG, AND F. FENG, *Mitigating hidden confounding effects for causal recommendation*, IEEE Transactions on Knowledge and Data Engineering, (2024).

[256] Y. ZHU, L. WU, Q. GUO, L. HONG, AND J. LI, *Collaborative large language model for recommender systems*, arXiv preprint arXiv:2311.01343, (2023).

[257] Y. ZHU, J. YI, J. XIE, AND Z. CHEN, *Deep causal reasoning for recommendations*, ACM Transactions on Intelligent Systems and Technology, 15 (2024), pp. 1–25.