CAUSAL REPRESENTATION LEARNING FOR SEQUENTIAL USER BEHAVIORS: FROM GENERIC TO NEWS RECOMMENDATION

XINGMING CHEN

PhD

The Hong Kong Polytechnic University

2026

The Hong Kong Polytechnic University

Department of Computing

Causal Representation Learning for Sequential User Behaviors: From Generic to News Recommendation

Xingming Chen

A thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

July 2025

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgment has been made in the text.

Signature: ____________________

Name of Student: Xingming Chen

Abstract

The rapid expansion of digital platforms has rendered personalized recommendation systems essential for connecting users with pertinent information in increasingly vast and dynamic content environments. Among these systems, sequential recommendation (SR)—which models users' evolving interests based on their behavioral histories—serves as a cornerstone for a wide range of applications, including e-commerce and news delivery. Despite considerable advancements, the majority of existing methodologies are grounded in statistical associations, often overlooking the underlying causal mechanisms that govern user behaviors. This limitation frequently results in spurious correlations and various forms of bias, thereby undermining the robustness and interpretability of recommendation outcomes.

This thesis systematically investigates causal representation learning for sequential user behaviors, with particular emphasis on bridging generic sequential recommendation and the distinct challenges inherent to news recommendation (NR). In this context, three fundamental research challenges are identified and addressed through a unified research agenda.

First, most existing causality-based recommendation approaches rely heavily on explicit external signals, such as item popularity or display position, to deconfound user-item correlations. However, such features are frequently unavailable or insufficient in real-world scenarios. This raises the critical problem of how to uncover hidden confounders and spurious factors that induce bias in user behavior modeling when

explicit signals are absent. To address this, a causality-driven sequential user modeling approach is introduced, which constructs causal graphs to identify and isolate unobservable confounding factors within user interaction histories. This framework effectively distinguishes genuine user interests from spurious correlations, thereby mitigating the impact of implicit biases and shortcut pathways in generic sequential recommendation tasks.

The second challenge stems from the inadequacy of relying on predefined causal graphs in complex and dynamic recommendation environments. As user behaviors and item spaces grow increasingly diverse, static causal structures fail to capture the dynamic generative processes underlying user decision-making. To overcome this limitation, the focus shifts to news recommendation—a domain characterized by rapidly changing topics and low content similarity between consecutive items. To meet this challenge, a Causal Behavior Pattern Inference (CBPI) framework is developed, which learns latent causal graphs directly from user interaction data. This approach enables the model to flexibly adapt to the evolving landscape of user interests and news content.

The third challenge lies in the necessity of capturing multi-level causality in user behavior, which includes both the dependencies among observed interactions and the generative influence of latent user interests on these behaviors. Existing causal discovery (CD) methods are typically restricted to a single variable layer, rendering them inadequate for modeling the hierarchical and intertwined nature of real-world user decision processes. To this end, a hierarchical causal-inspired structure modeling method for news recommendation is proposed, which jointly uncovers causal relationships between latent interests and observable behaviors, resulting in more interpretable and robust recommendations.

By systematically addressing these three challenges, this thesis presents a unified causal representation learning framework that reconstructs user and item representations based on the underlying causal processes governing user behaviors. Exten-

sive experiments on large-scale real-world datasets demonstrate that the proposed causality-aware approaches significantly enhance recommendation accuracy, diversity, and robustness, thereby establishing a new direction for the development of personalized recommendation systems that are both effective and explainable.

Publications Arising from the Thesis

1. Xingming Chen, Qing Li, “Causality-driven User Modeling for Sequential Recommendations over Time”, in *Proceedings of the Temporal Web Analytics (TempWeb) at The Web Conference (WWW)* (2024).

2. Xingming Chen, wenqi Fan, Qing Li, “Causal Behavior Pattern Inference for News Recommendation through Multi-interest Matching”, in *International Conference on Web Information Systems Engineering (WISE)* (2024).

3. Xingming Chen, Qing Li, “From Interests to Behavior: Hierarchical Causal-Inspired Structure Modeling for Personalized News Recommendation”, in *World Wide Web (WWW)* (2025).

Acknowledgments

I would like to express my heartfelt thanks to everyone who has supported and helped me throughout my studies.

First and foremost, I am deeply grateful to Prof. Qing Li for his unwavering support of my research. He has fostered an open and encouraging academic environment, allowing us to freely pursue topics that interest us. I also appreciate his provision of ample hardware resources, which has greatly facilitated my work. Throughout my studies, his detailed feedback on my writing has taught me how to revise and improve my papers. Even when my work was rejected by conferences or journals, his encouragement helped me persevere, teaching me how to approach reviewers' comments and decisions constructively.

I am also sincerely thankful to Prof. Wenqi Fan for his thoughtful guidance both academically and personally. Whenever I encountered difficulties, his valuable advice and encouragement motivated me to continue refining my research. He has also been instrumental in helping me recognize and overcome my weaknesses.

Finally, I would like to thank all my group members and the department staff for their constant support and assistance. Their help has been invaluable to me during my studies.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background & Motivation

The rapid expansion of digital content on platforms such as e-commerce websites, social networks, and news portals has fundamentally changed how people access and engage with information. Faced with an overwhelming array of choices, users increasingly depend on recommender systems (RSs) to help filter and personalize the content they see [73, 40]. By leveraging patterns in users' historical behaviors and item attributes, these systems play a critical role in enhancing user satisfaction, increasing engagement, and supporting the sustained growth of digital platforms.

One of the most important advances in this domain is sequential recommendation (SR). Unlike traditional methods that treat each user-item interaction as an independent event, SR explicitly considers the order and temporal dependencies within user behavior [135, 110, 93]. By analyzing the sequence of interactions, such as browsing products, listening to songs, or reading articles, SR models can capture how user preferences evolve over time [123, 81, 9]. For instance, as shown in Figure 1.1, a user's recent interaction history can be used to predict the next item of interest. This temporal modeling is key to understanding shifting preferences, adapting to trends, and providing recommendations that are sensitive to context [109, 63, 108]. Consequently, sequential recommendation has been widely adopted by industry leaders such as Amazon, Netflix, and YouTube [25].

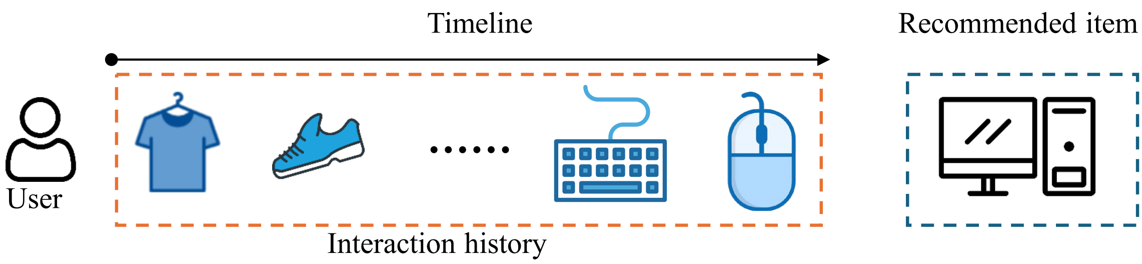


Figure 1.1: An example of a sequential recommendation scenario.

Despite these advances, a central limitation remains: most sequential recommendation systems still rely heavily on statistical correlations in historical data, without considering the underlying causal mechanisms that truly drive user choices. As a result, these models are prone to capturing spurious or coincidental associations, which can introduce or amplify bias [99, 125]. This issue is illustrated in Figure 1.2. For example, a user purchasing bread may later buy milk and butter, creating a pattern of frequent co-occurrence. However, not all items that appear together are causally linked. Milk and butter, for instance, often appear in the same basket but are not directly related in terms of purchase motivation. Similarly, a toothbrush may occasionally be bought alongside these groceries, yet it is not causally connected. Such examples highlight that frequent item combinations do not necessarily reflect true causal relationships. Methods that rely on statistical attention mechanisms may therefore assign undue importance to causally unrelated item pairs, reinforcing existing biases, limiting diversity, and potentially misaligning recommendations with genuine user interests [141, 28].

These challenges become even more pronounced in news recommendation (NR) [136, 118, 114]. Unlike product recommendation, where preferences often shift gradually, news consumption is characterized by rapid changes in user interest, a constant pursuit of novelty, and a need for topical diversity [137, 117]. Users may click on a wide variety of topics in quick succession, and their preferences can change dramatically in response to external events. As a result, simplistic sequential models that assume smooth transitions are often inadequate for modeling news consumption. The unique temporal dynamics and diversity requirements of NR amplify the risks associated with spurious correlations and recommendation bias [86, 30], posing significant challenges for building robust and interpretable systems.

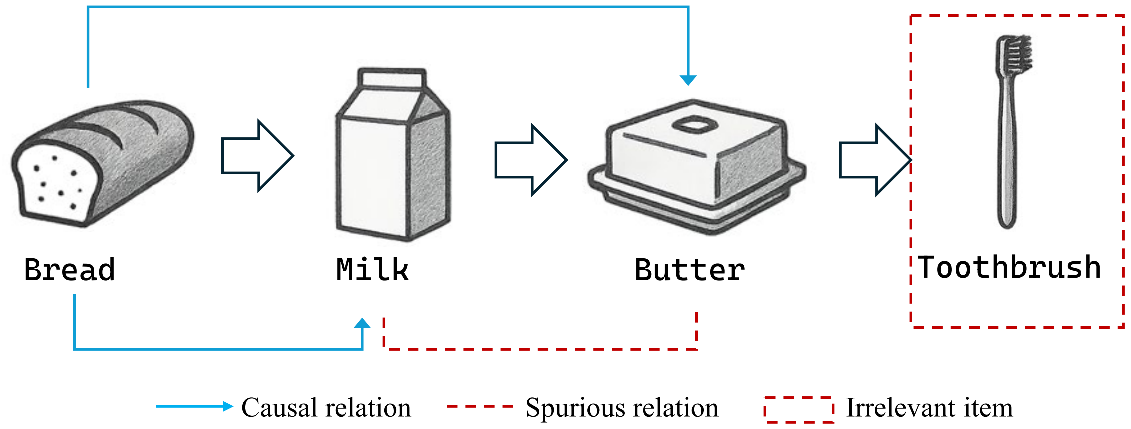


Figure 1.2: Illustration of why causal modeling is important in sequential recommendation.

Understanding and modeling causality is fundamental to addressing these issues. Psychological research suggests that human decision-making is inherently driven by logical, causal reasoning [70, 69], and these causal mechanisms often align with our intuitive understanding of behavior [58]. Recognizing the shortcomings of purely correlation-based approaches, recent work has increasingly focused on causal representation learning in sequential recommendation [138, 47]. By explicitly accounting for confounding factors—both observed and latent—causal methods [4, 87, 66] aim to identify the true drivers of user actions. This transition from correlation to causation is seen as essential for developing recommendation systems that are not only more accurate, but also more robust and interpretable. As this thesis will explore, bridging the gap from generic sequential recommendation to news recommendation through causal representation learning is both a timely and impactful research direction, addressing the pressing needs of modern personalized information services.

1.2 Research Problem & Framework

Although sequential recommendation methods have made significant progress, most existing approaches still rely mainly on statistical correlations found in historical data, rather than identifying true causal relationships [75, 85, 18]. This reliance creates both practical and theoretical challenges, limiting the development of robust, interpretable, and genuinely personalized recommendation systems—especially when moving from general sequential recommendation to the fast-changing field of news recommendation [137, 117].

At the core of this thesis is the challenge of learning causal representations for sequential user behaviors. Our objective is to approach sequential recommendation from a causal perspective, moving beyond superficial co-occurrences in the data to uncover the true factors that drive user actions. This allows us to generate recommendations that are both accurate and resilient to spurious correlations. This issue is especially important in environments where user interests are continuously evolving and subject to external influences or hidden confounders, as is frequently observed in news recommendation scenarios.

To address this overarching problem, three key research challenges are identified:

1. **Debiasing without explicit external signals:** Many current causal approaches in recommendation systems depend on explicit signals [97, 96], such as item popularity or display position, to handle confounding effects. However, these signals are often missing or incomplete in real-world sequential data. This raises the challenge of how to learn unbiased user and item representations using

only behavioral logs, while reducing the impact of misleading correlations.

2. **Adapting to dynamic and complex environments:** In domains like news recommendation, both the set of available items and user interests can change rapidly. The static, pre-defined causal graphs commonly used in generic sequential recommendation are often unable to capture these changes. Therefore, an important challenge is to develop methods that can infer and update causal structures dynamically, adapting to evolving user behaviors and content.

3. **Modeling hierarchical and multi-level causality:** User decision-making is inherently multi-layered, involving a hierarchy of causal factors. For example, a user's underlying interests drive observable behaviors, and these interests can themselves be shaped by external events or long-term preferences. Most existing methods focus on a single layer of variables [59, 105], overlooking the complex, hierarchical nature of real-world user interactions. Overcoming this challenge requires frameworks that can jointly reason about both latent interests and observable actions.

The bottom layer of the framework focuses on causality-driven user modeling. At this stage, the framework processes user-item sequential interaction data and constructs latent causal graphs to capture unbiased patterns in user behavior, even when explicit external confounding signals are missing. By identifying key causal factors within both immediate and longer interaction sequences, this layer lays a strong foundation for reducing bias. As a result, user and item representations become clearer and less affected by misleading associations, which helps improve the accuracy and reliability of recommendations.

The middle layer is centered on causal behavior pattern inference, a component especially important for news recommendation. In this part, the framework is able to refine user interests and build causal graphs at the interest level, while also maintaining logical consistency at the behavior level. This design allows the model to quickly

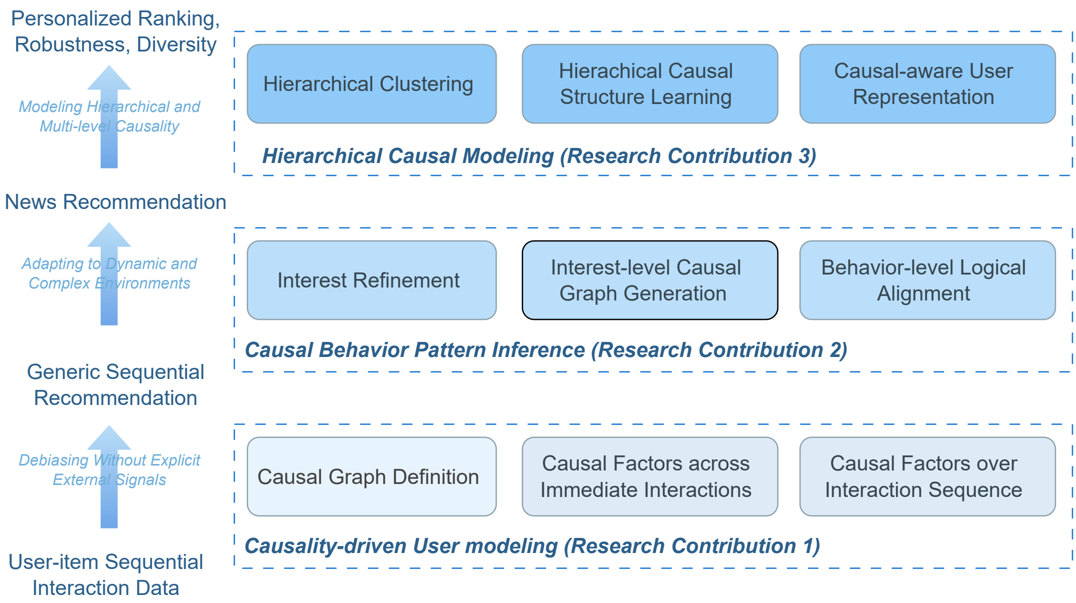


Figure 1.3: Causal representation learning framework sequential recommendation.

adapt to changes in both user preferences and news content, which is essential for keeping up with the fast-paced and ever-changing environment of news platforms.

At the top layer, the framework introduces hierarchical causal-inspired representation to better reflect how user decisions are influenced by multiple, layered factors. This includes using hierarchical clustering, learning causal structures at different levels, and forming user representations that are aware of these causal relationships. By connecting users' underlying motivations with their observable actions, this layer not only makes the recommendations easier to interpret, but also improves their robustness and diversity, supporting personalized ranking that truly matches users' interests.

By systematically covering these three aspects, the proposed framework provides a clear path from traditional correlation-based methods to more advanced causally informed sequential recommendation. This framework forms the basis for the following chapters, where specific methods and experiments will be described for both generic sequential recommendation and the more complex case of news recommendation.

1.3 Contributions

This thesis makes several key contributions to the field of sequential recommendation by integrating causal representation learning into both generic and news recommendation scenarios. The main contributions are as follows:

1. We propose a causality-driven user modeling approach for sequential recommendation, shifting the focus from simple correlations to a causal perspective. This approach leverages causal graphs to identify confounding factors that cause spurious correlations and to isolate conceptual variables that truly represent user preferences. By learning representations of these disentangled causal variables at the conceptual level, our method distinguishes causal from non-causal associations while preserving the sequential nature of user behaviors. This enables more accurate identification of critical factors and helps mitigate unintended biases.

2. To the best of our knowledge, we introduce the first model for Causal Behavior Pattern Inference (CBPI) to discover latent causal graphs for interest matching in personalized news recommendation. CBPI infers multiple latent interests and uncovers their underlying causal relationships, dynamically adapting to evolving user preferences. This work highlights the importance of identifying causal behavior patterns in news recommendation and initiates the pioneering exploration of causal discovery in this domain.

3. We present Hierarchical Causal-Inspired Representation (HCR), a unified framework that explicitly learns and utilizes hierarchical causal structures in user behavior. By jointly modeling both intra-behavior dependencies and the influence

of user interests on behaviors, HCR provides more valid, causality-aware user representations that support robust and accurate recommendations.

Overall, this thesis bridges the gap between traditional correlation-based and causally grounded sequential recommendation. The proposed innovations address both fundamental theoretical issues and practical challenges, culminating in a robust, interpretable, and adaptive solution for personalized information services.

1.4 Thesis Overview

This thesis is structured as follows:

1. Chapter 2 reviews the relevant literature, including foundational concepts and recent developments in sequential recommendation, news recommendation, and causal inference. This chapter also discusses the shortcomings of current approaches and underscores the importance of causal representation learning in dynamic recommendation settings.

2. Chapter 3 presents our proposed Causality-driven User Modeling framework, which seeks to uncover latent causal factors within users' historical interaction sequences. The goal is to generate unbiased user representations and enhance the robustness of recommendations in general sequential recommendation scenarios

3. Chapter 4 introduces Causal Behavior Pattern Inference (CBPI) for news recommendation. This method is designed to dynamically infer and adapt personalized causal structures in response to rapidly evolving user interests and content.

4. Chapter 5 details the Hierarchical Causal-Inspired Representation (HCR) framework, which models multi-level relationships between latent user interests and

observable behaviors. This approach aims to deliver more interpretable and robust recommendations, with a particular focus on the news domain.

5. Chapter 6 concludes the thesis by summarizing the main research problems addressed and the key contributions made. It also discusses potential directions for future research in this field.

Chapter 2

Literature Review

2.1 Sequential Recommendation: Generic Methods and News-Specific Advances

Sequential recommendation systems are developed to capture and anticipate how users interact with items over time, with the goal of delivering personalized suggestions that reflect users' evolving preferences. The effectiveness of these systems relies on several basic aspects: (1) user-item matching, which pertains to the collection of historical user interactions with items—such as impressions, feedback scores, transactions and clicks; (2) sequential user modeling, which encompasses techniques for learning temporal dependencies within user behaviors by taking into account the order and timing of interactions to enhance the predictive power of recommendations; and (3) the differentiation between short-term and long-term interests, where robust models are able to distinguish immediate preferences expressed within a single session from broader, more persistent interests that develop over multiple sessions or longer periods of activity.

Traditional solutions predominantly relied on matrix factorization (MF) [148, 13] and Markov chain (MC) models [27]. MF-based methods decompose the user–item interaction matrix into latent factors, uncovering hidden relationships that explain observed behaviors [33]. For example, FISM [37] focuses on item–item similarities to model low-order dependencies. MC models, in contrast, are adept at capturing short-term dynamics by assuming recent actions are strong predictors of subsequent behavior, with higher-order MCs incorporating longer interaction histories [24, 25]. However, both paradigms struggle with data sparsity and are limited in modeling complex, long-range and high-order dependencies.

The advent of deep neural networks (DNNs) has significantly transformed sequential recommendation. Recurrent neural networks (RNNs) and their variants (LSTM, GRU) [64, 94, 8, 144] are widely adopted for modeling temporal dynamics but may introduce spurious correlations and overlook collaborative effects. Convolutional neural networks (CNNs) [84, 132] emphasize local sequential patterns but are constrained in capturing long-term dependencies due to their fixed receptive fields. More recently, graph neural networks (GNNs) [121] have emerged, structuring user-item sequences as graphs to capture intricate, high-order dependencies and richer contextual signals. For instance, PUP [146] leverages graph convolutional networks (GCNs) to jointly model user price sensitivity and category influence, while TiDA-GCN [22] integrates temporal and cross-domain information for multi-platform recommendation. To overcome the limitations of single-model architectures, hybrid models further advance the field by combining multiple neural architectures, enabling the capture of a broader range of user behavior and preferences. For instance, some frameworks incorporate memory networks into RNN-based models, such as adding a key-value memory module to a GRU network [34], to preserve and update detailed knowledge of user interactions and uncover nuanced attribute-level interests. Attention mechanisms are also commonly fused with recurrent structures [130], enabling the model to selectively focus on the most relevant parts of an interaction sequence and to balance both short- and long-term dependencies. Other advanced hybrid models leverage diverse encoder types for specialized purposes: some infer user intentions at each step, while others identify a variety of items that match different user needs [95, 12]. In cross-domain scenarios, models like SGCross employ multi-view learning and hierarchical gating to integrate personal, temporal, and collaborative information from auxiliary domains, enhancing the quality of cross-domain recommendations. Some recent works incorporate external knowledge and multimodal features. For example, KPHAN [67] uses knowledge graphs and hierarchical attention to jointly model users' dynamic and static preferences, while OCBSR [50] unifies multimodal item features through convolutional autoencoders and models user dynamics with transformers and GRU networks, improving recommendation performance across diverse content types and time scales.

Within this landscape, personalized news recommendation stands out due to the unique characteristics of news consumption: extreme freshness, rapidly shifting user interests, high item turnover, and the need to process rich content features. Deep content-based models [49, 2] exploit powerful encoders and attention mechanisms to represent both news articles and user reading sequences, effectively capturing short-term preferences but often treating user histories as flat temporal lists. To capture more nuanced and structural dependencies, graph-based methods [36, 126] have been proposed, modeling higher-order relationships among news and user actions. More recent work explores modeling user interest diversity and latent intent: MINER [43] uses category-aware routing for multi-vector user representation, HDNR [92] reweights interactions to mitigate popularity bias, and GLoCIM [127] extracts long-term interests via global click graphs.

Despite these advances, most methods, including both generic and news-specific—primarily, learn statistical associations between user behaviors and content, often missing the underlying causal mechanisms that drive user clicks. This gap limits interpretability, robustness, and the system's ability to adapt to content or environment shifts. As a result, uncovering the causal structure behind user interests and observable actions has become an emerging research direction, aiming to provide more reliable, explainable, and robust sequential recommendation systems, especially in dynamic domains like news.

2.2 Causal Inference

Causal inference seeks to determine how specific variables, often referred to as treatments, directly impact outcomes, rather than merely identifying statistical associations. For example, in medicine, we may want to know whether a new drug actually improves patient health, not just whether drug usage and recovery tend to occur together. Similarly, when examining rising mental health issues, one might hypothesize that increased online gaming is a contributing cause and aim to quantify its effect.

A key point in causal analysis is that causation and correlation are not the same. As illustrated in Figure 2.1, an ice cream shop may observe that both electricity bills and ice cream sales increase on hot days, resulting in a statistical association between the two. However, leaving the lights on overnight will not boost sales, because the real cause behind both increased sales and higher electricity bills is the temperature. This example highlights the importance of identifying and accounting for confounding factors, which can create misleading associations between variables.

In practice, causal relationships are often described using causal graphs. These graphs represent variables (i.e., treatments, outcomes, confounders, and instrumental variables) and the causal links among them. Estimating the true effect of a treatment requires careful consideration of confounders, which are variables that influence both the treatment and the outcome. Unaddressed confounders can introduce bias and lead to incorrect conclusions about causality, as seen in the ice cream example where temperature acts as a confounding variable.

The most reliable way to eliminate confounding bias is through randomized con-

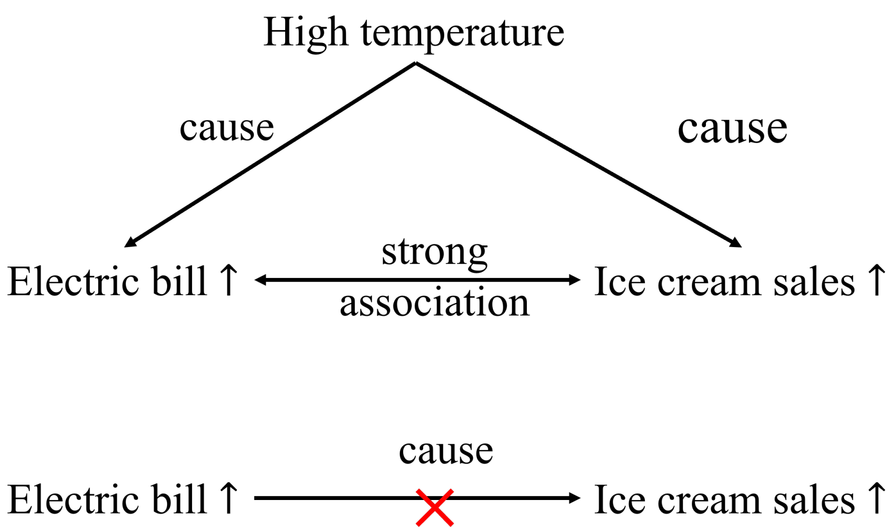


Figure 2.1: An ice cream shop example [129].

trolled trials (RCTs), where treatments are randomly assigned. However, RCTs are frequently impractical or unethical in many real-world scenarios. As an alternative, a range of methods has been developed for observational data. Propensity score methods (PSM) [53] attempt to mimic randomization by grouping instances with similar observed features but different treatments, while inverse probability of treatment weighting (IPTW) [32] achieves covariate balance through reweighting. In recent years, deep learning models have been adapted for causal inference, enabling the estimation of causal effects in settings with high-dimensional data. For instance, some approaches minimize both prediction error and distributional differences between treatment groups using multi-task neural networks [16, 87], or incorporate propensity scores into topic models to balance predictive performance and causal adjustment [66].

When confounders cannot be fully observed or controlled, instrumental variable (IV) methods offer an alternative. An instrumental variable affects the treatment but only influences the outcome through its effect on the treatment. The classic two-stage least squares (2SLS) method [39] is widely used for linear settings, while modern research extends IV analysis to nonlinear and high-dimensional domains. For example, kernel-based approaches [54] reformulate IV regression tasks, and deep learning models [23, 42] combine neural networks with IV or generalized method of moments (GMM) frameworks to improve estimation accuracy.

In summary, distinguishing genuine causal relationships from mere associations is fundamental for reliable decision-making and robust modeling. The methodological foundations introduced here, i.e., confounder adjustment and instrumental variable analysis, provide essential tools for uncovering causal effects in complex data, and they form the theoretical basis for causal representation learning in sequential recommendation, which is the focus of this thesis.

2.3 Causal Representation Learning for Recommendation

Recommender systems driven by large-scale data have made substantial progress in addressing the challenge of information overload [104, 100]. However, these systems often operate under the assumption that data points are independently and identically distributed (IID), which can unintentionally reinforce biases present in the training data [6, 96]. Such biases may contribute to issues like unfair treatment of users [1, 15], the emergence of filter bubbles [20], and a reduced ability to generalize to new or out-of-distribution (OOD) scenarios [76]. As a response, causal representation learning has emerged as a promising direction in recommendation systems, aiming to identify and model the genuine causal factors influencing user actions, rather than relying exclusively on observed associations. This research area is increasingly recognized for its potential to mitigate bias and enhance both the robustness and interpretability of recommender models.

Research on causal modeling in recommendation is primarily grounded in two theoretical frameworks: the potential outcome framework [68] and structural causal models [57]. Within the potential outcome paradigm, widely adopted techniques such as inverse propensity scoring (IPS) [72] and the doubly robust estimator [101] have been applied to correct for bias in explicit feedback [139], thereby supporting more objective recommendations. For example, Schnabel et al. [75] conceptualize recommendations as treatments and utilize IPS within an empirical risk minimization context to counteract bias. The LTD approach [102] leverages a small unbiased sample to learn a propensity model, and then applies IPS-based modeling to the remaining data. The UFAR framework [83] incorporates IPS for fairness-aware top-N recommendation, while IB [122] employs doubly robust methods [17, 3] to construct uplift-based models that are both unbiased and resilient. Although IPS-based methods are straightforward to implement, they require precise estimation of propensities and can suffer from high variance [71].

On the other hand, research based on structural causal models often investigates the fundamental relationships among variables by constructing and analyzing causal graphs. This methodology leverages interventions to address bias, enhance fairness, and provide interpretability in recommendation outcomes [96]. By moving beyond mere correlations, this approach offers a deeper insight into the mechanisms that shape user choices. Much of the existing literature centers on mitigating confounding bias in recommendation systems. For instance, Christakopoulou et al. identify exposure rates as confounding factors that affect both a user's likelihood of being shown an item and their eventual satisfaction, and correct for this by modeling propensity scores [11]. The DLCE method [74] considers user and item features as potential confounders, reweighting training samples to obtain a more accurate estimate of the true causal impact of recommendations. To counteract popularity bias, PDA [141] utilizes causal intervention techniques to decouple item popularity from user representations, thereby enabling the model to capture authentic user interests. Furthermore, some research introduces IVs, such as user query contexts from different domains, to help isolate the causal effects of recommendations by regressing user representations on these external factors [78, 35]. Most of these approaches are developed for conventional recommendation settings and typically rely on predefined objectives and manually specified causal variables. There remains a significant need for more adaptive methods that can automatically uncover implicit causal factors, especially within sequential recommendation contexts where user behaviors and interests evolve over time

Another recent direction is counterfactual learning for recommendation, which aims to generate and utilize counterfactual data to address challenges such as data sparsity. For instance, Mehrotra et al. have developed frameworks to estimate the causal impact of new content, such as newly released music tracks, by generating counterfactual samples using quasi-experimental Bayesian models [52]. In the context of sequential recommendation, Wang et al. [106] focuses on generating plausible counterfactual sequences near decision boundaries, employing data-oriented sampling strategies and model-aware matching techniques to maximize the utility and informativeness of these synthetic sequences. Causerec [138] construct counterfactual worlds at both the item and concept levels, using contrastive learning between the actual and counterfactual scenarios to augment training data and improve model generalization. Additionally, the COR approach [98] employs a variational autoencoder to learn causal representations of user preferences. In this method, counterfactual inference is performed by forcibly resetting specific historical interactions to zero, which helps reduce the influence of outdated behaviors and thus achieves stronger OOD generalization and rapid adaptation.

Overall, causal representation learning provides a promising foundation for building recommendation systems that are not only more accurate but also more robust and interpretable. As research in this area progresses, developing methods that can automatically identify and model causal structures in complex, sequential environments will be crucial for advancing the next generation of personalized recommendation.

Chapter 3

Causality-driven User Modeling for Sequential Recommendation over Time

3.1 Introduction

Sequential recommendation systems play a crucial role in filtering and personalizing content for users on modern digital platforms. These systems are designed to predict future user interactions by analyzing sequences of historical behaviors, which typically include user actions, item properties, and various contextual signals such as user demographics, item attributes, or social information [95, 26]. The predictive power of such systems relies heavily on accurately representing these multifaceted information sources to infer user preferences over time.

Recent progress in this field has produced a variety of methods that leverage the statistical patterns embedded in users' interaction histories. Techniques such as attention mechanisms [142] and RNNs [31] are widely used to capture temporal dependencies and model the evolution of user interests. Additionally, GNNs [124] and hybrid architectures [128] have been introduced to extract complex co-occurrence patterns and high-order structural relationships within sequential data.

However, despite these advances, most existing methods share a critical limitation: they typically assume that user profiles and item exposures are independent in the observed data. This simplifying assumption does not hold in real-world scenarios, where user behaviors are shaped by a multitude of contextual and temporal factors. In practice, user interactions are often influenced by confounding variables, such as item popularity, trending topics, or exposure mechanisms, which can introduce spurious correlations and obscure the true drivers of user preferences [29]. As a result, sequential data may contain noisy or irrelevant feedback, reflecting not only genuine interests but also external influences that distort recommendation quality.

Moreover, many state-of-the-art models employ similarity-based attention mechanisms [94, 7] that focus on the pairwise relationships between items within a user's interaction sequence. While such mechanisms are effective at capturing local item correlations, they can inadvertently emphasize what are known as "shortcut paths." A shortcut path refers to a direct connection between two items in the interaction graph, where the model ignores the sequential dependencies and intermediate actions that genuinely reflect the user's evolving interests. As illustrated in Figure 3.1, consider a sequence where, from a global causal perspective, the user's interactions are clustered around major U.S. holidays, particularly those occurring later in the year, such as Thanksgiving and Christmas. In this context, an appropriate recommendation would be items related to upcoming holidays, such as a New Year's Eve party kit. However, if the model focuses only on the last two items in the sequence, it may interpret the user's intent as a strong preference for Christmas decorations, i.e., a shortcut view, without recognizing the broader pattern of holiday-related interests. This tendency to rely on shortcut paths can obscure the underlying causal relationships and user motivations, ultimately diminishing both the accuracy and interpretability of

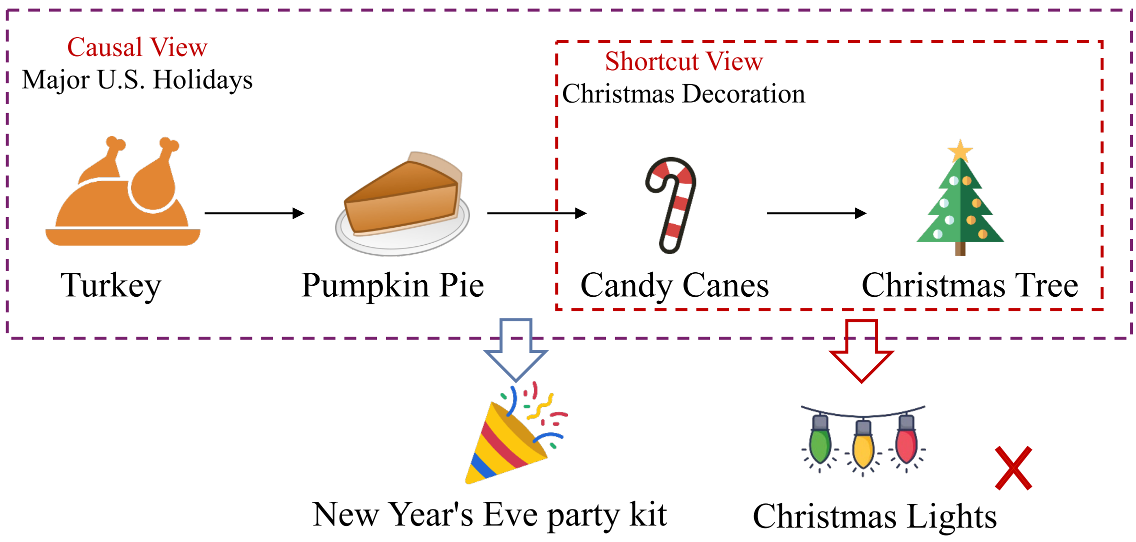


Figure 3.1: Depiction of shortcut path effects on sequential recommendation outcomes.

the recommendations.

To address these challenges, we propose a novel approach that incorporates causal analysis into generic sequential recommendation. Our method constructs a causal graph to explicitly model the dependencies among user behaviors, item exposures, and exogenous variables that may introduce bias. By leveraging this causal framework, we are able to identify and disentangle the true factors that drive user interests and item interactions over time. Specifically, we develop a causality-driven user modeling technique that treats user interaction histories as observational data. This approach autonomously discovers meaningful causal factors at a conceptual level, reducing reliance on manually defined features and effectively mitigating the influence of confounding variables.

Our framework further distinguishes between latent causal concepts that shape long-term user preferences and immediate contextual factors that affect short-term interactions. The identification of latent causal factors helps to filter out noise from shortcut features and ensures that the model captures relevant information across the entire sequence. At the same time, preserving immediate context enables the model to accurately characterize the causal influence of recent interactions, without amplifying real-time confounding effects.

By integrating these causally-driven representations into the base matching model, our approach delivers more precise predictions of user interactions and enhances the interpretability of the recommendation process.

The principal contributions of this work are as follows:

- We implement a causal graph to frame the task of generic sequential recommendation, enhancing the understanding of user behavior and improving prediction accuracy.
- We develop a causality-driven technique for user modeling over time, which can autonomously identify causal factors from user interaction data, mitigating biases without relying on predefined features.
- Our learned causal concept representations can be utilized to mitigate confounding bias and preserve critical context, leading to more refined predictions of user interactions.

3.2 Methodology

3.2.1 Problem Formulation

In generic sequential recommendation systems, let $U = u_1, u_2, \ldots, u_{|U|}$ denote the set of users and $I = i_1, i_2, \ldots, i_{|I|}$ denote the set of items, where $u_j \in U$ is an individual user and $i_k \in I$ is an individual item. For each user u, their interaction sequence is represented as $S_u = (i_1^u, i_2^u, \ldots, i_n^u)$, where each $i_t^u \in I$ indicates the item that user u interacted with at time t, and the sequence is ordered chronologically. Both users and items are typically initialized as ID embeddings in a d-dimensional space. The central task of sequential recommendation is to predict the next item i_{t+1} that user u is likely to interact with, given their historical sequence. Formally, this involves estimating the conditional probability $Y_{t+1} = P(i^u_{t+1} \mid S_u)$ for all candidate items. The system then recommends items to the user based on the ranking of these predicted probabilities.

3.2.2 Viewing from Causal Graph

To predict the next item a user will interact with, it is crucial to understand the reasons behind a user's interaction at each timestep, thereby learning accurate and robust representations of user behavior. Existing approaches often rely on user-item correlation matching, operating under the assumption that the co-occurrence of users and exposed items are independent within the observational data. For instance, at a given timestep t, conventional thinking suggests that the user's behavior of interacting with item i^u_t is exclusively driven by their preferences, implying the absence of a confounding back-door path between user u and item i_t. However, this assumption often fails to hold in complex, real-world scenarios.

By framing the sequential recommendation task from a causal perspective, we can address these discrepancies. We propose the use of a causal graph, as depicted in Figure 3.2, for modeling user preferences. This graph helps to identify exogenous variables that create undesired correlations between user preferences and the items they interact with instantaneously. Specifically, the path $i^u_t \leftarrow Z_i(t) \rightarrow u$ can be seen as conveying bias information that misguides the user to interact with item i^u_t at timestamp t, such as popularity bias, while $i^u_t \leftarrow Z_u \rightarrow u$ may represent a shortcut bias that overlooks sequential dependencies and intermediate interactions. Our aim is to discern the true causal factors, Z_u and Z_i, which directly influence user interests and the real-time exposure of items.

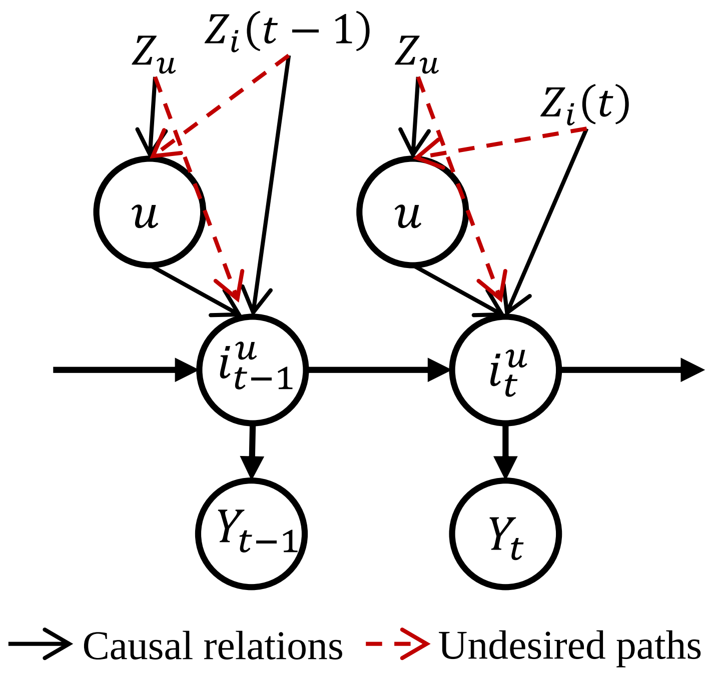


Figure 3.2: Causal graph of generic sequential recommendation.

3.2.3 The Base Matching Model

Our proposed method is model-agnostic in a way, and it can be integrated into multiple existing matching models for sequential recommendation. Since many models for sequential recommendation share similar architecture, we regard it as the base model and implement our method over it. Generally speaking, a matching model includes a user encoder $f_S(u, S_u) \in \mathbb{R}^d$, which takes the user profile u and the user interaction sequence $S_u = (i_1^u, i_2^u, \ldots, i_n^u)$ as input and outputs one d-dimensional vector to represent the user interaction sequence, and an item encoder $f_Y(i) \in \mathbb{R}^d$ representing the item in the same vector space as $f_S(u, S_u)$. The matching score is generally calculated by maximizing the likelihood of the next interacted item given the interaction sequence:

$$L_{base} = -\frac{1}{|U|} \sum_{u \in U} \log P_\theta(i_{n+1}^u | f_s(u, S_u)), \tag{3.1}$$

3.2.4 Causality-Driven User Modeling

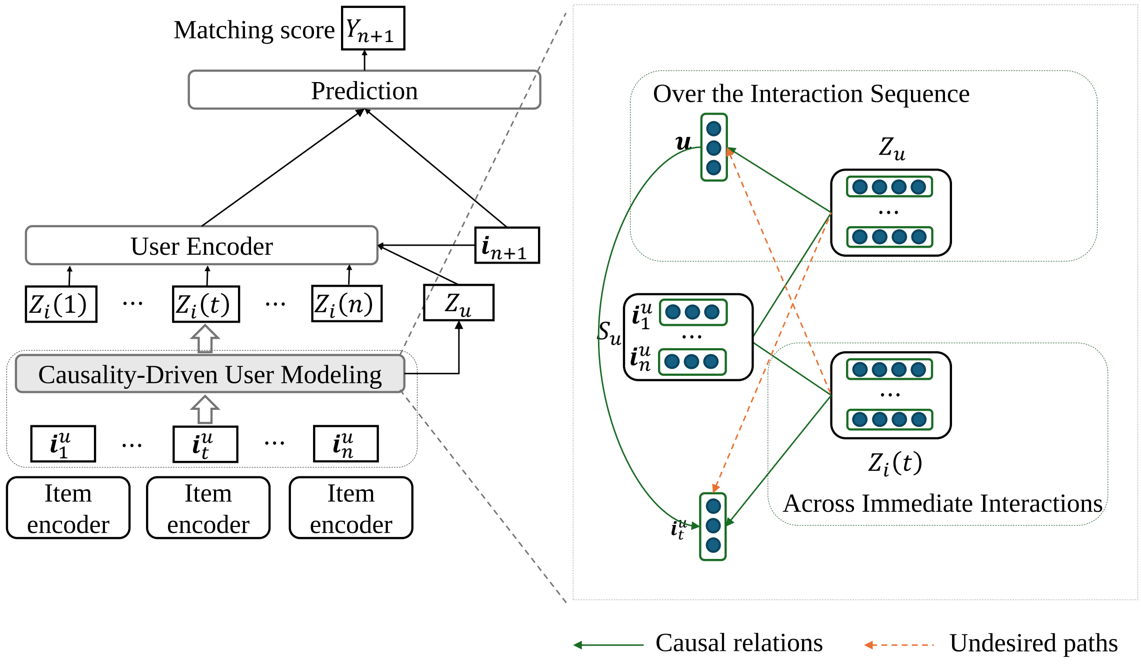


Figure 3.3: Framework architecture. Left: Integration of Causality-Driven User Modeling within the standard sequential recommendation matching Scheme. Right: Detailed implementation of Causality-Driven User Modeling.

Existing de-biasing techniques in the realm of sequential recommendation commonly rely on precisely identified confounding factors or supplementary contexts [55, 141, 78], such as user search queries, to serve as proxy variables. These proxies are instrumental in uncovering the genuine causal factors that reflect user interests. In this section, we treat users' sequential interaction histories as observational data. Our objective is to autonomously discern valid causal factor across both the interaction sequence and the immediate interaction, at the conceptual level. Additionally, we aim to mitigate the influence of potential back-door paths present at each timestep, and we strive to achieve this without depending on explicitly predefined biases. Our causality-driven user modeling method can be depicted as Figure 3.3.

Initially, we leverage Graph Attention Networks (GATs) [79] to encode user and item interaction representations, aiming to extract sophisticated features from all observed interaction sequences:

$$\mathrm{E}_{\mathcal{G}} = \mathrm{f}_{\mathrm{readout}}(\mathrm{Attn}(\mathrm{WI}, \mathrm{A})), \tag{3.2}$$

where W is a trainable weight matrix and A denotes the adjacency matrix that encapsulates the structure of interaction sequences. Attn(.) signifies a multi-head attention mechanism [88] which serves to assess the significance of neighboring behaviors and enhance the representation of item interactions at each timestep. And $f_{readout}$ refers to a readout function that propagates item-level features $\boldsymbol{i}_t^u$ to generate a graph-level representation $\mathrm{E}_{\mathcal{G}}$, and the representation $\boldsymbol{u}$ for encapsulating user preference can be obtained by inputting the corresponding interaction sequence from the graph.

Causal Concept Representation over the Interaction Sequence

To effectively represent the latent variables Z_u that causally impact user interests at a conceptual level, our primary approach is to identify aspects of the historical interaction sequence that meet two criteria: first, they must be strongly associated with the user representation $\boldsymbol{u}$; second, they should affect the real-time item interaction $\boldsymbol{i}_t^u$ only via the user, not directly.

Assuming there are K latent conceptual prototypes $Z_u \in \mathbb{R}^{K \times d}$, we first address the relevance of these prototypes to user interests. Inspired by the causal inference literature [133], we seek to isolate those aspects of the sequence S_u that are correlated with the user representation $\boldsymbol{u}$, so that they can be captured in the prototype matrix Z_u. To this end, we employ a variational distribution $q_{\theta_{Z_u}}(\boldsymbol{u}|Z_u)$ to approximate the conditional distribution of $\boldsymbol{u}$ given Z_u, using a two-layer multi-layer perceptron (MLP) for implementation. The objective is to maximize the likelihood of correctly reconstructing the user representation from the causal concepts, which is formulated as:

$$\mathcal{L}_{Z_u}^r = -\frac{1}{|U|} \sum_{u \in U} \log q_{\theta_{Z_u}}(\boldsymbol{u}|W_{ru} Z_u), \tag{3.3}$$

where W_{ru} denotes a linear transformation of the causal concepts Z_u. Subsequently, to reinforce the connection between the conceptual prototypes and user interests, we maximize the mutual information between $\boldsymbol{u}$ and Z_u. This is achieved by encouraging the model to distinguish the true user representation from randomly selected negative samples:

$$\mathcal{L}^{r'}_{Z_u} = -\frac{1}{|U|}\sum_{u\in U}\frac{1}{L}\sum_{j=1}^{L}(\log q_{\theta_{Z_u}}(\boldsymbol{u}|W_{ru}Z_u) - \log q_{\theta_{Z_u}}(\boldsymbol{u}'|W_{ru}Z_u)), \tag{3.4}$$

where u denotes the positive sample, whereas u'_j signifies a negative sample that has been randomly selected, and L is the number of negative samples selected. By maximizing the gap in log-likelihood between positive and negative samples, we ensure the learned causal concepts are truly relevant to user interests, thereby enabling a more precise depiction of user preferences for items.

As discussed earlier, a user's historical sequence of interactions over time can create shortcut paths. For example, a user might directly jump from an initial item i^u_1 to a distant item i^u_t, bypassing intermediate items. If the model relies solely on similarity-based statistical correlations, it may overstate the direct connection between two items and discard sequential dependencies. This shortcut perspective could prevent the model from capturing a true representation of user preference through the paths $u \leftarrow Z_u \rightarrow i^u_1$ and $u \leftarrow Z_u \rightarrow i^u_t$, making the reason behind a user's item choice more obscure. Therefore, it is crucial to block the back-door path $u \leftarrow Z_u \rightarrow i^u_t$ over time to mitigate the influence of these shortcut views. This can be accomplished by excluding the concepts Z_u associated with the interaction at a specific timestamp, thereby extracting valid insights that accurately reflect real user preferences based on sequential dependencies, rather than an immediate interaction's shortcut view. To satisfy the exclusion condition, we can apply methods similar to those used for maintaining the relevance condition. Initially, the variational distribution $q_{\phi_{Z_u}}(\boldsymbol{i}^u_t|Z_u)$ is used to approximate the conditional distribution of each i^u_t over time given Z_u. The corresponding loss function for negative maximum log-likelihood estimation is:

$$\mathcal{L}^{e}_{Z_u} = -\frac{1}{|U|}\sum_{u\in U}\sum_{t=1}^{n}\log q_{\phi_{Z_u}}(\boldsymbol{i}^u_t|W_{eu}Z_u), \tag{3.5}$$

where W_{eu} denotes the trainable weights. Then, to ensure that the causal concept representations Z_u from the user's perspective are not influenced by the shortcut views of items at one or two time steps, it is imperative to minimize the mutual information between i^u_t and Z_u. This can be optimized by minimizing the difference in mutual information between the positive sample (u, i^u_t) and the negative sample (u', i'^u_t):

$$\begin{aligned}\mathcal{L}^{e'}_{Z_u} = \frac{1}{|U|}\sum_{u\in U}\sum_{j=1}^{L}\sum_{t=1}^{n}(&sim(\boldsymbol{u},\boldsymbol{u}')\times\\ &(\log q_{\phi_{Z_u}}(\boldsymbol{i}^u_t|W_{eu}Z_u)-\\ &\log q_{\phi_{Z_u}}(\boldsymbol{i}'^u_t|W_{eu}Z_u))).\end{aligned} \tag{3.6}$$

Unlike that in keeping relevance condition, the causal concept representations cannot be absolutely irrelevant with the immediate interaction, for the reason that they are actually indirectly causally connected through the path $Z_u \rightarrow u \rightarrow i^u_t$. The exclusion requirement should be fulfilled only when conditioning on the user preference. Therefore, $sim(u, u')$ is employed to describe the similarity between the positive and negative ones, i.e., u and u'. When they are similar to each other, the weight of the discrepancy about log-likelihood between the positive and negative samples should be large. Otherwise, the weight should be small. In other words, users with similar interaction sequences should place greater emphasis on minimizing the mutual information. To obtain a normalized similarity-aware weighting among multiple negative samples, we use a softmax over the negative Euclidean distances:

$$sim(\boldsymbol{u},\boldsymbol{u}') = \mathrm{softmax}(e^{-\|\boldsymbol{u}-\boldsymbol{u}'\|}), \tag{3.7}$$

which ensures that more similar closer samples receive higher weights, which helps emphasize their influence on the mutual information loss. While the softmax function is not a similarity metric, it is used here to convert similarity scores into a probability distribution over samples for weighting purposes.

Through this design, the model is encouraged to focus on sequential dependencies that genuinely reflect user interests, while reducing the impact of shortcut connections that may arise from coincidental or spurious correlations.

Causal Concept Representation across Immediate Interactions

While mitigating the impact of shortcut features within the causal concepts throughout the interaction sequence is crucial for reducing bias, it's important to recognize that these features are still valuable. They encompass contextual information that has a direct causal effect on item interactions at each timestep, which is vital for enhancing the performance of recommendations. For this reason, beyond modeling causal concepts at the global sequence level, it is important to identify the causal concepts Z_i associated with each immediate interaction as well. Since the causal concepts underlying each item interaction can change dynamically over time, we introduce a set of K concept prototype embeddings, denoted as $Z_i(t) \in \mathbb{R}^{K\times d}$, to represent the causal concepts relevant to each specific timestep.

In a manner analogous to the method employed for learning Z_u, we utilize the variational distribution $q_{\phi_{Z_i}}(\boldsymbol{i}_t^u|Z_i(t))$ to uncover the latent information that causally influences the immediate item interactions, which can be achieved by minimizing the negative log-likelihood function:

$$\mathcal{L}_{Z_i}^r = -\frac{1}{|U|}\sum_{u\in U}\frac{1}{n}\sum_{t=1}^{n}\log q_{\theta_{Z_i}}(\boldsymbol{i}_t^u|W_{ri}Z_i(t)), \tag{3.8}$$

where W_{ri} is a trainable weight matrix. To further enhance the learning of causal relevance, we ensure that the log-likelihood for the observed (positive) item i_t^u is separated from that of a randomly sampled negative item i'_t. This is achieved by minimizing:

$$\mathcal{L}_{Z_i}^{r'} = -\frac{1}{|U|}\sum_{u\in U}\frac{1}{n}\sum_{t=1}^{n}(\log q_{\theta_{Z_i}}(\boldsymbol{i}_t^u|W_{ri}Z_i(t)) - \log q_{\theta_{Z_i}}(\boldsymbol{i}'_t|W_{ri}Z_i(t))), \tag{3.9}$$

Through this optimization, the immediate causal concept representation $Z_i(t)$ is encouraged to preserve critical shortcut features that are genuinely relevant to the user's current decision.

However, we must also guard against potential confounding introduced by back-door paths such as $i_t^u \leftarrow Z_i(t) \rightarrow u$ in the causal graph. Such bias might mislead users into clicking on items that do not align with their genuine interests. For instance, the popularity of an item might increase its likelihood of being exposed to users, prompting clicks on the item due to its popularity rather than its relevance to the users' preferences. To address this, it is necessary to remove any spurious correlation between the immediate causal concept $Z_i(t)$ and the user embedding $\boldsymbol{u}$, so that Z_i serves as a valid instrumental variable. To be specific, the relevance between $\boldsymbol{u}$ and $Z_i(t)$ can also be captured with the variational approximation via estimating the maximum likelihood of the distribution:

$$\mathcal{L}_{Z_i}^{e} = -\frac{1}{|U|}\sum_{u\in U}\frac{1}{n}\sum_{t=1}^{n}\log q_{\theta_{Z_i}}(\boldsymbol{u}|W_{ei}Z_i(t)), \tag{3.10}$$

where W_{ei} represents a linear transformation. To prevent the immediate causal concepts $Z_i(t)$ from capturing spurious associations, we regularize them so that they are only related to the user representation $\boldsymbol{u}$ in contexts where such a correlation is justified by the actual item interaction i_t^u:

$$\begin{aligned}\mathcal{L}_{Z_i}^{e'} = \frac{1}{|U|}\sum_{u\in U}\sum_{t=1}^{n}\big(& sim(\boldsymbol{i}_t^u, \boldsymbol{i}_t'^u)\times \\ & \big(\log q_{\theta_{Z_i}}(\boldsymbol{u}|W_{ei}Z_i(t)) - \\ & \log q_{\theta_{Z_i}}(\boldsymbol{u}'|W_{ei}Z_i(t))\big)\big),\end{aligned} \tag{3.11}$$

where $\boldsymbol{i}_t'^u$ and $\boldsymbol{u}'$ denote the representations of randomly sampled negative items and users, respectively, and $sim(\boldsymbol{i}_t^u, \boldsymbol{i}_t'^u)$ is computed using the softmax function as in Equation 3.7.

At this stage, we have identified the latent causal concepts Z_u and Z_i that correspond to the overall interaction sequence and to the real-time item interactions, respectively. Representations derived from these causal concepts provide a more faithful account of user preferences, where biases arising from shortcut paths and item popularity are effectively controlled. This causality-driven modeling approach is instrumental in achieving more accurate and reliable recommendations.

3.2.5 Model Optimization

The latent causal concept representations Z_u and Z_i are integrated into the base matching model to forecast the user's subsequent interactions. These predictions take the form $f_S(Z_u, [Z_i(1), \ldots, Z_i(t)])$ for the sequence. The comprehensive optimization objective that we strive to minimize is expressed as follows:

$$L_{overall} = L_{base} + \alpha(\mathcal{L}^{r}_{Z_u} + \mathcal{L}^{r'}_{Z_u} + \mathcal{L}^{e}_{Z_u} + \mathcal{L}^{e'}_{Z_u}) + \beta(\mathcal{L}^{r}_{Z_i} + \mathcal{L}^{r'}_{Z_i} + \mathcal{L}^{e}_{Z_i} + \mathcal{L}^{e'}_{Z_i}), \quad (3.12)$$

where the hyperparameters α and β serve to adjust the relative importance of the variational approximation losses and the mutual information constraint losses. These hyperparameters are selected and tuned based on performance on the validation set.

The primary objective of the optimization remains the loss function of the base model, as the fundamental goal is to accurately predict the user's next item of interest by leveraging both their profile and their interaction history. This modeling framework is designed to be flexible, allowing for integration with most existing mainstream sequential recommendation models that share a similar structural foundation.

3.3 Experiment

3.3.1 Datasets and Experimental Settings

To evaluate the effectiveness of our proposed method for sequential recommendation, we conducted experiments on three publicly available datasets: *Diginetica*[1], *MovieLens*[2], and *Book-Crossing*[3]. The key statistics of these datasets are summarized in Table 3.1.

- **Diginetica.** This dataset is derived from an e-commerce platform and contains user purchase records.
- **MovieLens.** A widely-used benchmark for movie recommendation, featuring user movie rating and viewing histories.
- **Book-Crossing.** A classic dataset that includes book ratings contributed by members of the Book-Crossing community.

Table 3.1: Statistics of the datasets.

Dataset	# Users	# Items	# interactions	Domain
Diginetica	33,364	43,589	223,562	E-commerce
MvoieLens	6,040	3,629	836,478	Movies
Book-Crossing	22,817	319,199	1,028,948	Books

In our experiments, items within each user's interaction sequence are ordered chronologically. To increase the number of training instances per user, we generate multiple sub-sequences from each interaction history. For example, given a sequence

[1]https://competitions.codalab.org/competitions/11161
[2]https://grouplens.org/datasets/movielens/
[3]http://www2.informatik.uni-freiburg.de/ cziegler/BX/

$i_1 \to i_2 \to i_3 \to i_4$, we create three training instances: i_1, $i_1 \to i_2$, and $i_1 \to i_2 \to i_3$, where the label for each instance is the item the user interacted with next.

Following established protocols [121, 147], we reserve each user's last interaction for testing, the penultimate interaction for validation, and use the remaining interactions for training. Model performance is evaluated based on NDCG and F1 metrics, using the top 5 recommended items compared to the actual user interactions.

For hyperparameter tuning, we conduct a grid search on the validation set. The loss weighting coefficients α and β are set to 0.1 and 0.15, respectively. In particular, we search over $\alpha \in \{0.05, 0.1, 0.15, 0.2, 0.25, 0.3\}$ and $\beta \in \{0.05, 0.1, 0.15, 0.2, 0.25.0.3\}$, and observe that these values yield a good balance between the base prediction loss and the causal regularization terms. The number of latent concepts K is set to 12, and the number of negative samples L for learning Z_u is set to 8. Parameters for baseline models are adopted from previously reported optimal settings.

3.3.2 Baselines

Our proposed framework is designed to enhance existing models for sequential recommendation that have a base structure analogous to the one described in Section 3.2.3. To demonstrate the efficacy of our method, we have implemented and integrated it with four established baseline models:

- **STAMP [46]**. A model leverages an attention mechanism to emphasize short-term user preferences while also capturing their long-term interests.
- **SASRec [121].** A model based on self-attention mechanisms that captures the relationships between items in the entire sequence.
- **SRGNN [38].** A GNN-based model that represents each sequence as a graph to effectively model the complex transitions between items.

- **GES-SASRec [147].** A framework that applies graph convolutions in the context of embedding smoothing for sequential recommendation, leveraging semantic item relationships derived from inherent item attributes.

3.3.3 Main Results

Our experimental evaluation involves four baseline models applied to three representative datasets. The detailed results are summarized in Table 3.2. For each baseline, we also report the performance of its causality-driven variant, denoted as CD-STAMP, CD-SASRec, CD-SRGNN, and CD-GES-SASRec, where the prefix CD-" stands for Causality-Driven".

A close examination of the results reveals several notable trends:

1. Consistent Improvements Across All Baselines and Datasets. Across Diginetica, MovieLens, and Book-Crossing, every causality-driven (CD-) model outperforms its original counterpart in both NDCG@5 and F1@5 metrics. For example, on Diginetica, CD-STAMP achieves an NDCG of 15.89 compared to 14.67 for STAMP, and an F1 of 7.82 versus 7.01. Similar improvements are observed with the other models and datasets. These gains demonstrate the effectiveness of incorporating causal concept representations into sequential recommendation frameworks, leading to more accurate modeling of user intent.

2. Robust Generalization Across Domains. The three datasets span different application scenarios: e-commerce (Diginetica), movies (MovieLens), and books (Book-Crossing). The observed improvements with causality-driven models on all datasets indicate that our approach is not limited to a specific domain, but rather provides robust generalization ability for sequential recommendation tasks in diverse contexts.

3. Magnitude of Improvement Varies by Baseline. While all causality-driven variants show improvements, the degree of performance gain varies. Notably, CD-GES-SASRec achieves relatively modest improvements compared to its baseline, especially on MovieLens and Book-Crossing. This can be attributed to the unique property of GES-SASRec, which relies on additional item attribute information. Our current framework does not explicitly disentangle causal concepts associated with these item attributes, which may include noisy or confounding factors. As a result, the full potential of causality-driven modeling is not realized in this setting. Extending our method to better handle attribute-based causal reasoning could be an avenue for future work.

3. Strongest Gains in Challenging Domains. On the Book-Crossing dataset, which features a much larger item space and more sparse interactions, the relative improvements of the causality-driven models are particularly marked. For instance, CD-STAMP increases NDCG from 1.65 to 2.60 and F1 from 0.78 to 1.06, representing substantial relative gains. This suggests that the benefits of causal representation learning are especially pronounced in challenging or sparse recommendation environments.

In summary, our causality-driven framework consistently enhances sequential recommendation models, yielding improvements in both ranking quality (NDCG) and recommendation accuracy (F1). Although there remains room for further refinement, particularly in handling complex item attribute signals, our results demonstrate the broad applicability and strong potential of causality-driven representations in generic sequential recommendation.

3.3.4 Confounding Bias Alleviation

To further evaluate whether our method can reduce confounding bias, we focus on the impact of item popularity in sequential recommendation. In many real-world systems, popular items are more likely to be exposed and thus more frequently interacted

Table 3.2: Comparative main results between the baseline models and our causality-driven methods.

Datasets	Diginetica		MovieLens		Books-Crossing	
metric (@5)	NDCG	F1	NDCG	F1	NDCG	F1
STAMP	14.67	7.01	9.07	4.50	1.65	0.78
CD-STAMP	15.89	7.82	9.33	4.82	2.60	1.06
SASRec	16.29	8.03	8.57	4.50	1.69	0.80
CD-SASRec	16.96	8.62	8.80	4.62	1.92	1.12
SRGNN	16.55	8.23	8.75	4.72	1.74	0.91
CD-SRGNN	16.98	8.71	8.99	4.83	1.96	1.15
GES-SASRec	17.14	8.66	9.02	4.79	1.92	1.02
CD-GES-SASRec	17.35	8.94	9.13	4.86	2.03	1.12

with, sometimes regardless of a user's genuine preferences. To quantify this effect, we calculate the interaction frequency of each item in the Diginetica dataset and split both the training and evaluation sets into two groups: the top 20% of items are labeled as popular, and the remaining 80% as unpopular.

Figures 3.4 and 3.5 display the results for NDCG@5 and F1@5 on both item groups. Here, we compare the original baseline models with their causality-driven counterparts, where our approach introduces causal concepts to better capture the underlying reasons for user choices. GES-SASRec is not included in this analysis since it relies on additional item attributes.

From the results, several clear patterns emerge. First, all baseline models show a significant drop in performance when moving from popular to unpopular items. For

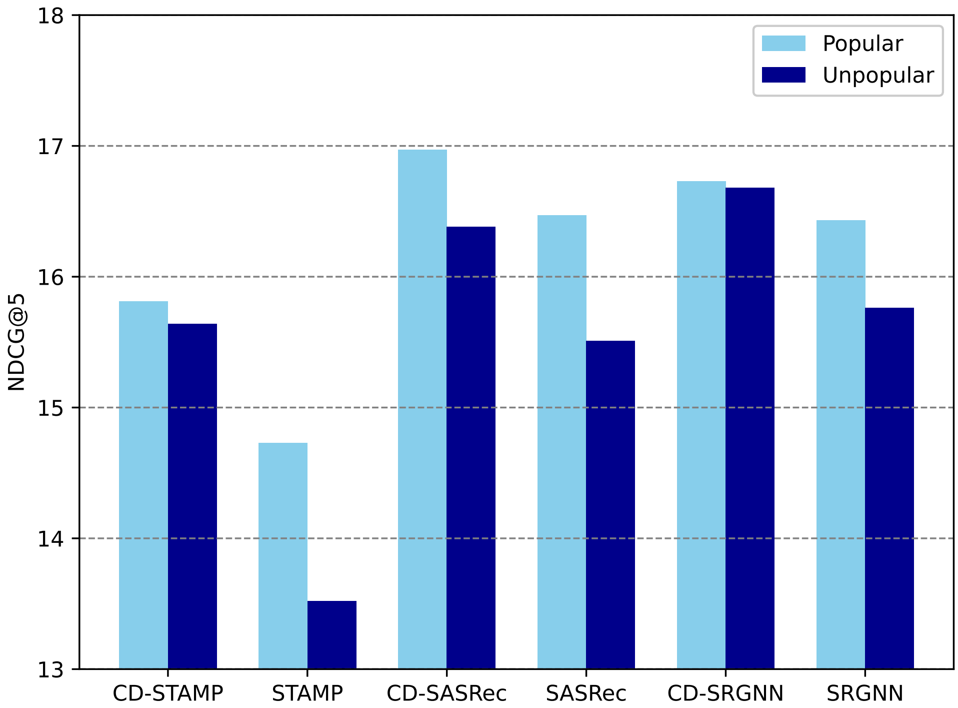


Figure 3.4: Results highlighting popularity bias (NDCG@5).

example, in the STAMP model, both NDCG@5 and F1@5 are notably higher for popular items than for unpopular ones. This indicates that the baselines are influenced by an implicit popularity bias, tending to recommend items that have already received considerable attention.

After integrating our causality-driven approach, the performance gap between popular and unpopular items is substantially reduced. For instance, CD-SASRec achieves a much smaller difference in NDCG@5 and F1@5 between the two groups compared to SASRec. Similar trends are observed in CD-STAMP and CD-SRGNN. This improvement demonstrates that the introduction of causal concepts helps the model focus more on the true drivers of user behavior, rather than simply relying on item popularity.

Another important observation is that our causality-driven models show especially strong improvements for unpopular items. The scores for these items increase notably,

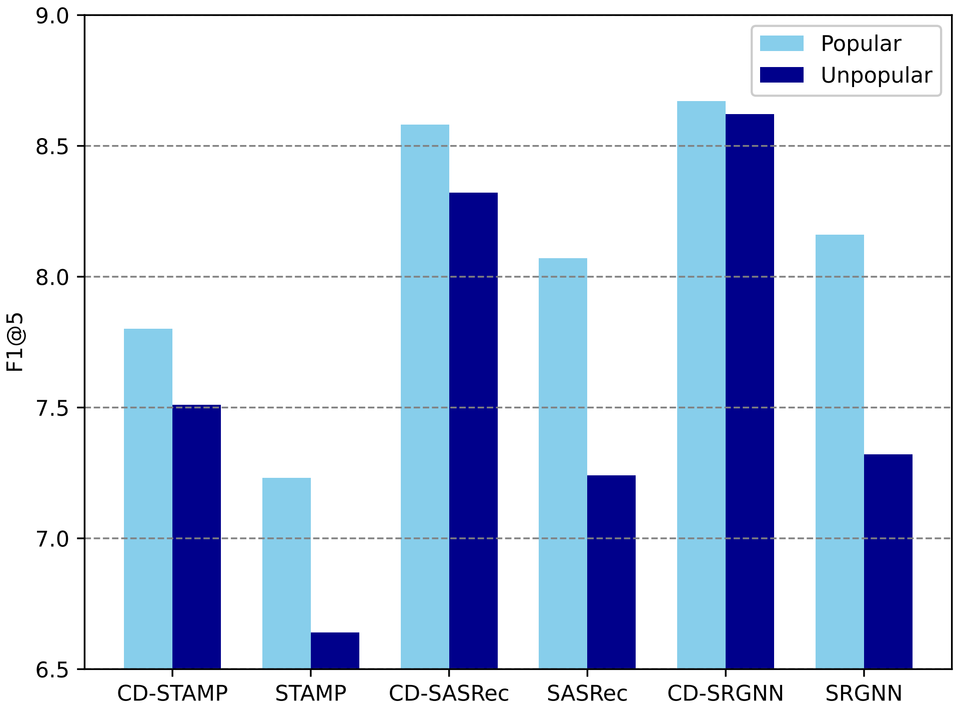


Figure 3.5: Results highlighting popularity bias (F1@5).

making the recommendation results more balanced and fair. This suggests that by modeling the latent causal relationships within user interaction sequences, the system can uncover more authentic user interests, which benefits less popular items that might otherwise be overlooked.

The experimental results and visualizations show that our approach can not only improve overall recommendation accuracy but also effectively mitigate popularity bias by introducing causal concepts. This allows the model to provide fairer and more accurate recommendations that are better aligned with users' actual preferences.

3.4 Summary

In this work, we propose a causality-driven user modeling approach for generic sequential recommendation, which significantly enhances prediction accuracy and mitigates bias. Unlike conventional methods that rely primarily on statistical correlations, our approach redefines the recommendation process through the use of a causal graph. By interpreting user interaction histories as observational data, we are able to independently identify latent causal factors that reflect user preferences at a conceptual level, avoiding the need for manually predefined features. Furthermore, our method leverages these latent causal factors to block back-door paths that can introduce spurious associations, thereby preserving the most relevant information throughout the user's interaction sequence. We also extend this framework to capture causal concepts at the level of immediate item interactions, which helps address confounding biases that arise in real time. By integrating these causal concept representations into user preference modeling, our approach improves the performance of various mainstream matching models and leads to more accurate predictions of user behavior.

It is important to note, however, that our current approach still depends on a predefined, static causal graph to guide the identification of causal relationships. In practical scenarios where user interests and item content evolve rapidly, such as in news recommendation, relying on a fixed causal structure can limit adaptability and responsiveness to change. In these dynamic environments, the ability to automatically learn and update the underlying personalized causal graph directly from the observed interaction data becomes highly desirable. Motivated by these limitations, the next section will introduce a new method that enables the automatic discovery and adaptation of latent personalized causal structures, aiming to further improve recommendation effectiveness in rapidly changing contexts.

Chapter 4

Causal Behavior Pattern Inference for News Recommendation through Multi-interest Matching

4.1 Introduction

With the rapid growth of online news platforms such as MSN News[1] and Google News[2], users now have access to an overwhelming volume of information at any moment. While this brings great convenience, it also poses the challenge of information overload, making it difficult for users to efficiently discover news that truly matches their interests. As a result, personalized news recommendation systems have become indispensable for tailoring content to individual preferences [61, 45, 56, 19].

In our first work, we leveraged a predefined causal graph to disentangle and utilize latent causal concepts in user behavior sequences. However, in highly dynamic domains like news recommendation, where both user interests and item content can change

[1] https://microsoftnews.msn.com/
[2] https://news.google.com/

rapidly, such static causal graphs often struggle to capture the evolving nature of interactions. This restricts the model's adaptability and responsiveness, making it difficult to provide truly personalized and timely recommendations. As discussed above, extending from generic sequential user modeling to the news recommendation scenario requires a more flexible and adaptive framework. Specifically, there is a pressing need for methods that can automatically discover and update the underlying personalized causal graph directly from observed user interaction data, rather than relying on predefined structures. This dynamic capability is especially crucial in news scenarios, where trends, topics, and user preferences can shift dramatically within short periods.

Most existing news recommendation methods still treat the problem as a generic sequential modeling task, constructing unified user representations based on static features from prior interactions. Approaches utilizing GRUs [2], GNNs [90], attention mechanisms, and BERT [113, 112, 116] primarily emphasize recent browsing behaviors, often resulting in recommendations that closely follow a user's latest actions. While some works have attempted to address preference diversity by inferring multiple interest vectors using explicit signals such as news categories [43], keywords [91], or popularity [92], they still mainly learn statistical correlations from content, neglecting the causal relations that truly drive user engagement. This may result in amplified selection bias and recommendations with limited diversity, as genuine underlying behavior patterns are overlooked. For example, after reading serious news about politics, it is very likely for a user to read some joyful news to relax whose content is completely different from the recent browsed news. Regardless of such causal patterns in news recommendation, when the recent interacted news is dominated by a specific kind of category, it would amplify the selection bias and ultimately recommend news all with highly coincident contents intensively.

In real-world news consumption scenarios, user browsing behaviors often exhibit causal patterns that offer meaningful explanations for recommending specific news

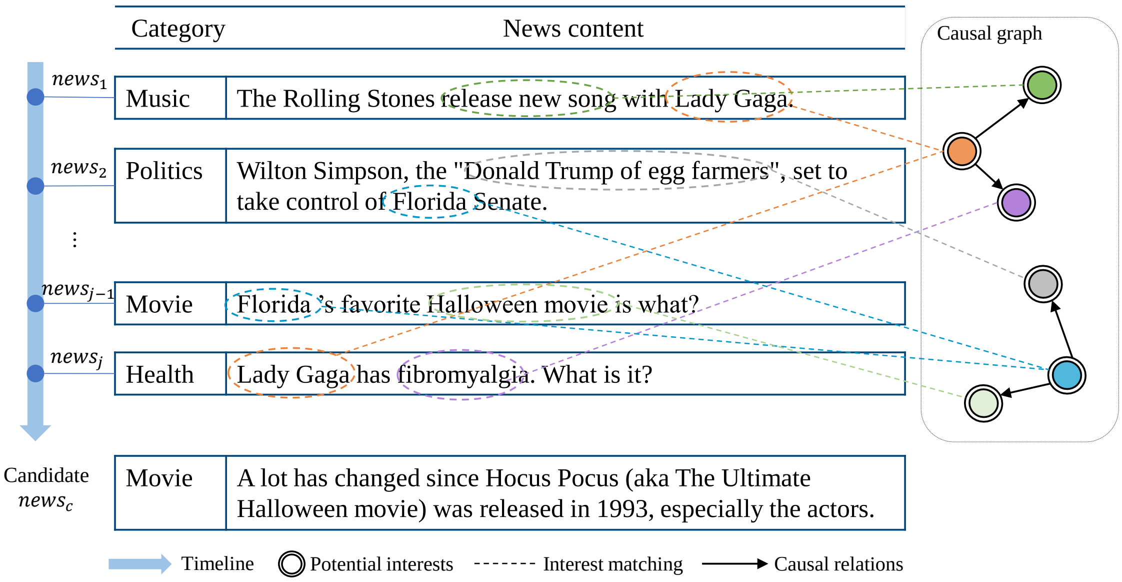


Figure 4.1: An illustrative example of how understanding causality among potential interests enhances user behavior pattern analysis

articles. From a causal perspective, these patterns can be inferred by uncovering the intrinsic causal relationships among key latent interests that drive user behaviors. As depicted in Figure 4.1, a user's engagement with $news_j$ may be a result of a previous encounter with $news_1$, driven by an interest in "Lady Gaga", despite the dissimilarity in content and the significant time gap. This fan status implies a higher probability for the user to access $news_j$ connected to "Lady Gaga", even they are not interested in movies or health. Similarly, the behavior of browsing $news_{j-1}$ can be inferred from $news_2$ due to the user residing in Florida. The user's location in Florida sparks an interest in movies or political events. Without the premise of living in Florida, movies or politics may not be significant factors for inferring the user preferences. Therefore, the user would not browse a news article like $news_c$, despite its content being similar to $news_1$ and $news_{j-1}$, as it is unrelated to "Lady Gaga" or Florida. Existing methods based on statistical associations among news contents would mistakenly emphasize the importance of movie or actor-related news, leading to erroneous recommendations such as $news_c$.

To tackle the above challenges, we propose a causal behavior pattern inference framework named CBPI for news recommendation through multi-interest matching. To be specific, we first disentangle the sequence of historical browsed news into multiple interest concepts. Motivated by the recent research in causal discovery [145, 41], we integrate an encoder-decoder framework to jointly learn the causal pattern among those concepts, and model the temporal interest transition process. For the encoder, we aim to infer a directed acyclic graph (DAG) with the whole time-series concepts as nodes to reveal their causal relations, which are insensitive across time. In contrast, the decoder is designed to model the dynamics from current concepts to predict the next time-index concepts by aggregating the information from other causally associated ones. Moreover, based on the consideration that causal patterns at the interest level generally align with those at the behavior level, by mapping the interest-level causalities to generate a news-level causal graph, whose edges represent inherent logic between two pieces of news. With the news-level causal graph, when learning the user representation, we can filter the effect of causally irrelevant news at each time index and thus each piece of news is only relative to its causal parent nodes. Therefore, spurious correlations will be alleviated and the diverse user interests and preferences can be captured more accurately and robustly.

The principal contributions of this work are as follows:

- We propose a novel framework for modeling behavior patterns from a causal perspective in the news recommendation context. To our knowledge, it is the first work to explore causal discovery in multi-interest matching within this domain.

- We develop an encoder-decoder structure to jointly learn time-invariant causal relationships among latent interests and their dynamic transitions. The learned causality at the interest level is further mapped to the news-level, uncovering intrinsic causal patterns of historical user behaviors.

- By overcoming the limitations of fixed, predefined causal structures, our method enables the automatic discovery and adaptation of latent personalized causal graphs from the observed user behaviors, making it especially suitable for dynamic and fast-evolving domains as news recommendation.

4.2 Background

Differentiable causal discovery has recently emerged as a powerful approach for uncovering the underlying generative mechanisms of observed data. By parameterizing the adjacency matrix of a latent DAG and imposing acyclicity constraints, these methods formulate causal discovery as a differentiable optimization problem that can be efficiently solved via gradient-based training [145, 131, 48, 82]. Recent advances have also explored the integration of causal discovery with supervised learning tasks, demonstrating that identifying and leveraging key causal variables can enhance model robustness, especially under distributional shifts [59, 41].

Within the recommender systems domain, several works have introduced causal discovery techniques to improve user modeling. For example, CausPref [29] employs a DAG-based regularizer to stabilize user preference learning across varying environments, while CAUSER [105] extends these ideas to uncover item-level causal dependencies in sequential recommendation scenarios. CGSR [120] further constructs dual cause-effect graphs to reduce spurious correlations, though its focus remains largely on mitigating confounding at the item level.

Despite these advances, existing methods predominantly operate on item IDs or handcrafted features, which limits their applicability in highly dynamic domains such as news recommendation. In these settings, news content evolves rapidly and user intent is often highly contextual. For instance, TCCM [10] seeks to mitigate popularity bias in news recommendation, but relies on a fixed, pre-defined causal graph rather than learning causal structures directly from user behavior data. Consequently, while these methods have made important progress, they are largely restricted to static, ID-based environments and struggle to adapt to the challenges posed by dynamically updated news content and shifting user interests.

To address these limitations, we propose to advance causal discovery in the news recommendation setting by modeling the causal relationships among intrinsic user interests, rather than relying solely on item-level or ID-based correlations. This approach enables us to capture deeper behavioral patterns and adapt more effectively to the rapidly changing landscape of news content and user preferences.

4.3 Preliminaries

4.3.1 News Recommendation

In personalized news recommendation, each user u is represented by a time-ordered sequence of previously clicked news articles, denoted as $\mathcal{D}^u = [D_1, D_2, \ldots, D_N]$. The objective is to predict the probability that the user will browse a candidate news article D_c in a future impression. A typical matching-based news recommendation framework comprises a news encoder $\mathcal{F}_n$ and a user encoder $\mathcal{F}_u$. The news encoder $\mathcal{F}_n$ encodes the content of each historical news article, generating a sequence of news representations $E_u = [\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_N]$ corresponding to $\mathcal{D}^u$, as well as an embedding $\boldsymbol{e}_c$ for the candidate news D_c. The user encoder $\mathcal{F}_u$ then aggregates the sequence E_u into a single user representation $\boldsymbol{u} = \mathcal{F}_u(E_u)$, which resides in the same vector space as $\boldsymbol{e}_c$ and captures the user's current preferences. To compute the matching score y between the user and the candidate news, we adopt the dot product between their embeddings, i.e., $y_u = \boldsymbol{e}_c^\top \cdot \boldsymbol{u}$, following common practice in prior work [113, 116, 43].

4.3.2 Causal Discovery

Understanding the underlying causal relationships among variables is fundamental for robust modeling and reasoning in many real-world applications. In the context of sequential user behavior modeling, causal discovery aims to uncover the true generative mechanisms that drive observed data, rather than relying solely on statistical associations.

At the core of causal discovery lies the concept of a Structural Causal Model (SCM), which provides a formal framework for representing and reasoning about causal relationships in a system. An SCM is typically represented by a Directed Acyclic Graph (DAG), where nodes correspond to variables and edges denote direct causal effects.

Definition 1 (Structural Causal Model (SCM))**.** *A Structural Causal Model (SCM) is defined over a set of random variables* $\mathcal{X} = X_1, X_2, \ldots, X_D$ *and consists of a directed causal graph* $\mathcal{G} = (\mathcal{X}, \mathcal{E})$ *along with a set of structural equations. Each edge* $(i, j) \in \mathcal{E}$ *represents a direct causal relationship from* X_i *to* X_j*. For each variable* X_j*, the corresponding observed data is denoted as* $\mathbf{x}_j$*. The probability measure* P *satisfies the Markov property with respect to* $\mathcal{G}$*, allowing the joint distribution to be factorized as* $P(\mathbf{X}) = \prod_{j=1}^{D} P(\mathbf{x}_j \mid \mathbf{x}_{pa(j)})$*, where* $pa(j)$ *denotes the set of parent variables of* X_j *in* $\mathcal{G}$*. We further assume that* P *follows an Additive Noise Model (ANM) [60], i.e.,*

$$\mathbf{x}_j = f_j(\mathbf{x}_{pa(j)}) + \epsilon_j, \tag{4.1}$$

where $f_j(\cdot)$ *is a (possibly nonlinear) function representing the data-generating mechanism for* V_j*, and* ϵ_j *is an independent noise term.*

Definition 2 (Directed Acyclic Graph (DAG))**.** *A Directed Acyclic Graph (DAG) is a graph consisting of nodes representing variables and directed edges between them. Each edge* $X_i \to X_j$ *signifies a direct causal effect from* X_i *to* X_j*. A path between* X_i *and* X_j *is a sequence of directed edges where the end node of each edge is the start node of the next. The graph is acyclic if there is no sequence of directed edges that starts and ends at the same node, i.e., no variable can be its own ancestor.*

Recently, differentiable causal discovery methods have gained popularity for learning such causal graphs directly from observational data [5, 48]. Given an input dataset $\boldsymbol{X} \in \mathbb{R}^{N \times D}$, where each sample $\boldsymbol{x}_i = [x_i^1, \ldots, x_i^D]$ records observations of random variables $X^1, \ldots, X^D$, the goal is to infer a causal graph that captures the underlying relationships among the variables. In causal discovery, the causal graph is typically modeled as a DAG, encoded by an adjacency matrix $A \in \mathbb{R}^{D \times D}$, where $A_{i,j} > 0$ implies a directed causal influence from X^i to X^j, and $A_{i,j} = 0$ implies no causal influence.

A common approach [145, 131, 41] formulates causal discovery as minimizing the following objective:

$$\min \mathcal{L}_{dag} = \mathcal{L}_{rec}(A, \boldsymbol{X}) + \lambda_1 \mathcal{L}_{acyc}(A) + \lambda_2 \mathcal{L}_{spa}(A), \tag{4.2}$$

where $\mathcal{L}_{rec}$ measures the reconstruction error of the observed data $\boldsymbol{X}$, $\mathcal{L}_{acyc}$ imposes acyclicity to ensure a valid DAG structure, and $\mathcal{L}_{spa}$ encourages sparsity in the adjacency matrix A. The reconstruction loss is commonly defined as:

$$\mathcal{L}_{rec}(A, \boldsymbol{X}) = \sum_{i=1}^{N} \sum_{j=1}^{D} (x_i^j - \boldsymbol{x}_i^\top A_{.j})^2, \tag{4.3}$$

where $A_{.j}$ denotes the j-th column of A. This formulation encourages each variable to be linearly reconstructed from its causal parents, while variables with no causal influence are effectively pruned by zero entries in A.

To guarantee that A encodes a directed acyclic structure, the acyclicity constraint is typically formulated as $\mathcal{L}_{acyc} = \mathrm{trace}(e^{A \odot A}) - D$, where $\odot$ denotes the element-wise product. This constraint leverages the property that, according to the Taylor expansion, the trace operation sums over all possible cycles in the graph. Thus, enforcing $\mathcal{L}_{acyc} = 0$ rules out cycles of any length, ensuring that no variable can be both a cause and an effect in a directed path. Meanwhile, sparsity is encouraged via an ℓ_1-norm regularization term $\mathcal{L}_{spa}(A)$, promoting an interpretable and parsimonious graph structure.

Prior research [41, 59] has shown that optimizing the objectives in Equation 4.2 can reveal the causal structure embedded in the data and support the learning of representations that invariant across different environments. Building on this insight, similar ideas can be incorporated into news recommendation to discover personalized causal structures within users' historical behaviors, thereby enabling more robust and invariant modeling of user interests.

4.4 Methodology

In this section, we provide a comprehensive overview of our proposed CBPI framework, as illustrated in Figure 4.2. The framework is designed to be flexible and compatible with various types of news and user encoders, making it adaptable to different real-world scenarios.

CBPI consists of three core modules: (1) *Interest Activation*, which disentangles the historical behavior sequence into multiple latent time-series interests; (2) *Interest-level Causality*, which leverages an encoder to construct a DAG to uncover the causal relations among interests, and a decoder to model the structural transition information of interests along with time; (3) *News-level Causality*, which aims to generate the causal graph at the behavior level and reconstruct the news representations to be aware of causal behavior patterns. Finally, the causality-aware news representations are integrated within the user encoder to enhance prediction accuracy. This design ensures that the recommendation process is guided not only by observed behavioral data but also by the underlying causal mechanisms that drive user behavior.

4.4.1 Interest Activation

Users typically exhibit a wide range of interests and interact with news content that aligns with these varied preferences over time. Such interests are often too complex

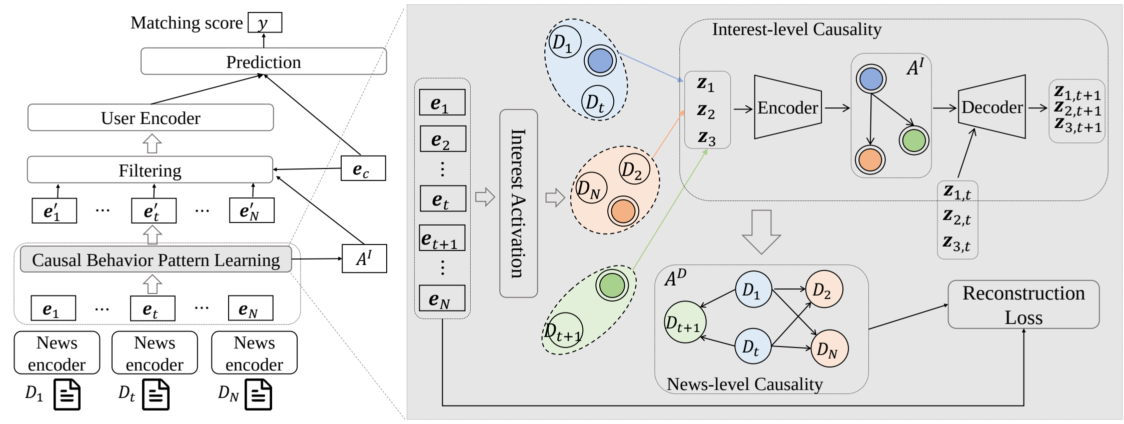


Figure 4.2: The CBPI framework architecture. Left: Integration of CBPI within the Standard matching Scheme. Right: Detailed CBPI module implementation.

to be captured by explicit features like news categories. Instead, it is necessary to infer high-level, latent interest concepts from users' behavioral histories. Given K latent concepts and a sequence of browsed news embeddings E_u produced by a news encoder, our objective is to disentangle each news item into these latent concepts and reveal the underlying interests activated by each interaction.

For each timestep t, we compute the concept assignment distribution $\boldsymbol{s}_t = [s_{t,1}, s_{t,2}, \ldots, s_{t,K}]$ by minimizing the interest activation loss $\mathcal{L}_{IA}$:

$$\min \sum_{\boldsymbol{e}_t} \|\boldsymbol{e}_t - \sum_{k=1}^{K} s_{t,k}\boldsymbol{c}_k\|_2^2, \quad \text{s.t.} \sum_{k=1}^{K} s_{t,k} = 1, s_{t,k} \geq 0, \tag{4.4}$$

where $\boldsymbol{c}_k$ denotes the prototype center of the k-th latent concept. By optimizing this objective, we simultaneously refine each news item's semantic representation $\boldsymbol{e}_t$ and determine its concept assignment vector $\boldsymbol{s}_t$, thereby aligning semantic embeddings with the underlying interest space. Importantly, at each timestep, only the relevant concepts are activated, while others are zeroed out to prevent interference from unrelated concepts.

4.4.2 Interest-level Causality

User interests are inherently complex and interrelated, which makes it insufficient to represent them using isolated concept embeddings. There often exist intrinsic causal relationships among different interest concepts, where the emergence of one may trigger another. Motivated by advances in causal discovery [41, 48], our goal is to capture these causal dependencies by constructing a user interest graph based on browsing history E_u. To this end, we define an adjacency matrix $A^I \in \mathbb{R}^{K\times K}$ to represent the causal links among interest concepts within a directed acyclic graph (DAG). In this matrix, $A^I_{kk'} > 0$ indicates that the k-th concept exerts a causal effect on the k'-th concept, while a zero entry denotes no causal relationship.

Traditional causal discovery approaches typically treat user interests as static when generating A^I, representing only immediate concepts as nodes and using a decoder to reconstruct the original embeddings. However, such methods are limited in the context of news recommendation, as they focus too narrowly on immediate causal relations and lack adaptability to dynamic user interests. Since causal patterns should not be overly sensitive to temporal fluctuations, we propose an encoder $q_E(A^I_{kk'}|Z_u)$ that operates over the entire time series of interests, $Z_u = [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N]^\top$. Here, each k-th element corresponds to a node in A^I and represents the k-th time-series concept, denoted as $\boldsymbol{z}_k = [\boldsymbol{z}_{k,1}, \ldots, \boldsymbol{z}_{k,N}]$:

$$q_E(A^I_{kk'}|Z_u) = \text{softmax}(\mathcal{F}_E(Z_u)_{kk'} + \boldsymbol{g}), \tag{4.5}$$

where $\boldsymbol{g}$ is the Gumbel distributed noise [77] used to relax the distribution during training. The graph edges represent causal relations among concepts. The encoding function $\mathcal{F}_E(Z_u)$ can be implemented with any graph-based network. Here, we implement two stages of edge aggregation across the causal graph:

$$\boldsymbol{e}^{(1)}_{kk'} = f^{(1)}_{edge}(W_E\boldsymbol{z}_k \| W_E\boldsymbol{z}_{k'}), \quad \boldsymbol{e}^{(2)}_{kk'} = f^{(2)}_{edge}(\boldsymbol{h}_k \| \boldsymbol{h}_{k'}), \tag{4.6}$$

$$\boldsymbol{h}_k = f_{node}(\sum_{i \neq k}^{K} \boldsymbol{e}_{ik}), \quad \boldsymbol{h}_{k'} = f_{node}(\sum_{i \neq k'}^{K} \boldsymbol{e}_{ik'}), \tag{4.7}$$

where message passing functions $f_{edge}^{(1)}$, f_{node}, $f_{edge}^{(2)}$ are implemented with MLPs, and $\boldsymbol{e}_{kk'}^{(2)}$ is the result of $\mathcal{F}_E(Z_u)_{kk'}$. This allows us to predict the time-insensitive causal pattern of user interests by learning A^I over time-series concepts.

Unlike the encoder, the decoder cannot treat the concepts as static features, as it may overlook the temporal evolution characteristic of news recommendation. At time t, the behavior $\boldsymbol{e}_t$ only relates to limited concepts, and the activated concepts vary over time. To dynamically model this evolving process, we develop the decoder $p_D(\boldsymbol{z}_{k,t+1}|A^I_{kk'}, \boldsymbol{z}_{k,t})$ through a single stage of aggregation:

$$\boldsymbol{\phi}_{kk'}^t = f_{edge}(W_D \boldsymbol{z}_{k,t} \| W_D \boldsymbol{z}_{k',t}), \quad \boldsymbol{\mu}_{k,t+1} = f_{node}(\sum_{i \neq k}^{K} \boldsymbol{\phi}_{ik}^t \| W_D \boldsymbol{z}_{k,t}) + \boldsymbol{z}_{k,t}, \tag{4.8}$$

$$p_D(\boldsymbol{z}_{k,t+1}|\boldsymbol{z}_t, A^I) = \mathcal{N}(\boldsymbol{\mu}_{k,t+1}, \sigma^2 \mathbb{I}), \tag{4.9}$$

where aggregation functions f_{edge} and f_{node} use MLP decoding similar to the encoder. $\mathbb{I}$ is the identity matrix. The decoder $p_D(\boldsymbol{z}_{k,t+1}|\boldsymbol{z}_t, A^I)$ captures evolving user interests by transforming the current time-series interest embedding $\boldsymbol{z}_{t,k}$ into subsequent $\boldsymbol{z}_{t+1,k}$, incorporating information from causally related concepts. To parameterize this generative process, we adopt an isotropic Gaussian distribution, as specified in Equation 4.9. Compared to the standard normal distribution, this choice offers greater generality and flexibility, allowing the model to learn arbitrary mean vectors and a shared variance across dimensions while maintaining a spherical covariance structure. This formulation effectively balances expressiveness and regularization, and has been widely applied in related causal discovery and variational modeling literature [41, 48].

In essence, our framework comprises an encoder $q_E(A^I_{kk'}|E_u)$ that identifies causal relationships among user interests, and a decoder $p_D(\boldsymbol{z}_{t+1,k}|A^I_{kk'}, \boldsymbol{z}_{t,k})$ that tracks the temporal evolution of these interests. The interest-level causal graph A^I is optimized using the following objective function:

$$\mathcal{L}_{IC} = \sum_{u}^{|\mathcal{D}|} \mathrm{KL}\left[q_E(A^I|Z_u)||p(A^I)\right] - \mathbb{E}[\log p_D(\boldsymbol{z}_{k,t+1}|\boldsymbol{z}_t, A^I)] + \lambda||A^I||, \\ \text{s.t.} \quad \mathrm{trace}(e^{A^I \odot A^I}) - K = 0 \tag{4.10}$$

where the first term forms a variational lower bound in encoder-decoder framework, minimizing KL-Divergence to a prior distribution over the encoder, and a negative log-likelihood for the decoder. $||A^I||$ promotes graph sparsity with coefficient $\lambda > 0$. The constraint $\mathrm{trace}(e^{A^I \odot A^I})$ ensures A^I is directed acyclic.

4.4.3 News-level Causality

Recognizing that causal patterns at the interest level mirror user behaviors, we project these findings to the news level to uncover the logic behind each action.

Intuitively, if concept k causally influences k' and they are activated when browsing news D_t and D_t' respectively, we can infer a causal link between these behaviors. For example, a user's location may cause interest in local trends; browsing location-based news can predict subsequent interest in local culture news. By examining all activated concepts and their causal links, we determine logical associations between browsing behaviors. With the learned adjacency matrix A^I, we construct the news-level causal graph $A^D = Z_u^\top A^I Z_u \in \mathbb{R}^{N\times N}$, where $A^D_{t,t'}$ represents the probability that browsing D_t leads to browsing $D_{t'}$. In recommendations, a high $A^D_{t,t'}$ indicates a likelihood of interest in $D_{t'}$ subsequently after browsing D_t. This operation refines user preference predictions and tailors user representations. We reconstruct news embeddings using causative news with the loss function $\mathcal{L}_{rec} = \min \sum_u^{|\mathcal{D}|} \|E_u - A^D E_u\|_2^2$, integrating user action causality to enhance behavior pattern insights and preference encoding.

We have constructed a news-level causal graph to uncover the causal pattern in user behaviors. For a candidate news article D_c, we evaluate its causal impact from a user's history as $A^D_c = Z_u^\top A^I \boldsymbol{s}_c$, where $\boldsymbol{s}_c \in \mathbb{R}^K$ represents the concept assignment distribution of D_c. This distribution is obtained by applying the softmax function to the normalized cosine similarities between the candidate news embedding $\boldsymbol{e}_c$ and the center concept vectors $\{\boldsymbol{c}_k\}_{k=1}^K$. Additionally, given extensive browsing histories, many articles may lack causal relevance to D_c. To prioritize those with stronger causal connections, we filter out irrelevant articles before processing them with the user encoder $\mathcal{F}_u$ as $E'_u = \text{diag}(A_c^D > \tau)E_u$, where τ is a threshold parameter to select articles with higher causal relevance.

4.4.4 Model Implementation

For the news encoder $\mathcal{F}_n$ and user encoder $\mathcal{F}_u$, we adopt the implementation from [116]. Specifically, the news encoder leverages a pre-trained BERT model [14] to generate the news embeddings E_u, while the user encoder utilizes a multi-head self-attention mechanism to aggregate user behaviors.

Following established practices [43, 91], we employ negative sampling during training, where, for each user, the model predicts scores for one positive and L negative news samples. The primary objective for the news recommendation task is formulated as a log-likelihood loss:

$$\mathcal{L}_{NR} = -\sum_{u}^{|\mathcal{D}|} \log \frac{\exp(y_u^+)}{\exp(y_u^+) + \sum_{j=1}^{L} \exp(y_u^j)}, \tag{4.11}$$

where y_u^+ and y_u^j denote the predicted scores for the positive and the j-th negative sample for user u, respectively.

Combining the above objective with the auxiliary losses introduced in previous sections, the final training objective is defined as:

$$\mathcal{L} = \mathcal{L}_{NR} + \alpha \cdot \mathcal{L}_{rec} + \beta \cdot \mathcal{L}_{IC} + \gamma \cdot \mathcal{L}_{IA}, \tag{4.12}$$

where α, β, and γ are hyperparameters that balance the contributions of the reconstruction loss $\mathcal{L}_{rec}$, the interest-level causality loss $\mathcal{L}_{IC}$, and the interest activation loss $\mathcal{L}_{IA}$, respectively.

4.4.5 Computational Complexity and Scalability

To evaluate the practicality of CBPI, we analyze its per-user computational complexity.

- **Interest Activation** computes weighted combinations of K concept vectors for each of the N historical news embeddings. This step has a complexity of $O(NKd)$.

- **Interest-level Causal Discovery** models invariant causal relationships among interest concepts. The encoder captures pairwise interactions among K concepts with complexity $O(K^2d)$, while the decoder aggregates temporal transitions over N steps, recurring an additional complexity of $O(NK^2d)$.

- **News-level Causal Projection** constructs a news-level causal graph via matrix multiplication $A_D = Z_u^\top A_I Z_u$, where $Z_u \in \mathbb{R}^{N\times K}$ is the concept distribution matrix and $A_I \in \mathbb{R}^{K\times K}$ is the interest-level adjacency matrix. This operation requires $O(KN^2)$ time. The subsequent reconstruction of news representations through causal aggregation involves $O(N^2d)$ complexity.

The total computational complexity per user can be summarized as:

$$O(NKd + K^2d + NK^2d + KN^2 + N^2d). \tag{4.13}$$

Although the complexity includes several quadratic terms with respect to N, it remains manageable in practice due to the typically limited length of user histories. Furthermore, the framework supports several optimization strategies, including sparse graph pruning, low-rank approximations, and incremental updates, which can further improve scalability in large-scale deployments.

4.5 Experiment

4.5.1 Dataset

We assess the effectiveness of our proposed approach on the real-world news recommendation dataset MIND [119], which includes two versions: MIND-small and MIND-large. The MIND-large version comprises approximately 161,000 news articles and over 15 million impression logs collected from 1 million users. By comparison, MIND-small is a sampled subset containing interaction data from 50,000 users randomly selected from MIND-large. Due to licensing restrictions, the ground-truth labels of the official test set in MIND-large are inaccessible. Following previous work [51, 91], we randomly divide the original validation set into two equal parts to construct a new validation set and a separate test set. Throughout our experiments, we utilize only the user behavior logs and news metadata provided in the MIND dataset. A summary of dataset statistics is provided in Table 4.1.

Table 4.1: Statistics of the two datasets

	MIND-small	**MIND-large**
# News	65,238	161,013
# Categories	18	20
# Impressions	230,117	15,777,377
# Users	50,000	1,000,000
# Clicks	347,727	24,155,470

4.5.2 Experimental Settings

In our experiments, we utilized the most recent 50 news articles browsed by each user to learn user representations. The news embedding dimensionality was set to 300, with a batch size of 64. Negative sampling was applied at a ratio of 1:4 (one positive sample and four negative samples per user). The hyperparameters λ, α, β, and γ were carefully tuned based on validation set performance and set to 0.1, 0.15, 0.15, and 0.2, respectively. For evaluation, we adopted standard metrics including AUC, MRR, nDCG@5, and nDCG@10, consistent with previous studies [43, 51]. To focus on modeling core features without relying on external information, we used only the title text as input to the news encoder. Accordingly, we did not compare our method with approaches that utilize heterogeneous or hybrid auxiliary features [112, 36].

4.5.3 Baseline Methods

To measure the performance of our proposed CBPI on news recommendation, we compare it with the following state-of-the-art methods: **Neural recommendation models:** (1) DKN [89] employs CNNs for news representations and a target-aware attention network for user representations. (2) NPA [111] integrates a personalized attention mechanism to enhance word and news interaction recognition for user and news profiling. (3) LSTUR [2] leverages GRU to detect short-term user interests, while viewing user ID embeddings as long-term preferences. (4) NRMS [113] utilizes multihead self-attention to learn news representations and understand content relationships to model user interests. (5) NNR [51] combines biLSTM with cross-attention for news encoding and GCNs for structural user encoding.

Transformer-enhanced models: (1) IPNR [91] integrates tf-idf features with ConceptNet to model users' browsing intentions, employing GCNs for temporal dynamics and FastFormer for preference capture. (2) NRMS+BERT and LSTUR+BERT [116] enhance NRMS and LSTUR models with BERT for superior news encoding. (3) MINER [43] introduces category-aware attention to refine user representation learning and accommodate varied browsing interests.

Graph learning methods: (1) GAT [88] enhances graph representation by atten-

Table 4.2: Main results of different methods (values in % without the symbol).

Methods	MIND-small				MIND-large			
	AUC	MRR	nDCG@5	nDCG@10	AUC	MRR	nDCG@5	nDCG@10
DKN	62.75	28.94	31.46	38.02	64.51	30.02	33.04	39.47
NPA	64.51	30.02	33.05	39.46	66.24	31.94	34.52	40.59
LSTUR	65.77	30.85	34.07	40.51	67.01	32.17	35.34	40.78
NRMS	66.20	31.12	34.65	41.17	67.53	31.94	35.45	41.85
NNR	66.96	31.50	35.01	41.41	67.89	32.36	35.93	42.24
IPNR	68.16	32.43	35.98	42.20	69.63	34.42	37.82	44.02
NRMS+BERT	68.33	32.76	36.49	42.71	69.45	34.67	37.72	43.69
LSTUR+BERT	68.12	32.39	36.08	42.39	69.40	34.39	37.50	43.65
MINER	69.58	33.81	37.61	43.88	71.48	36.15	39.70	45.36
CBPI	**69.73**	**33.95**	**37.93**	**44.07**	**72.32**	**36.64**	**40.27**	**46.01**
CBPI+GAT	68.42	32.94	36.54	42.78	69.46	34.73	37.86	43.70
CBPT+CASTLE	68.96	33.17	36.94	43.05	70.38	35.13	38.39	44.28
CBPT+CAUSER	69.02	33.23	36.99	43.22	70.44	35.14	38.36	44.32

tively weighting neighboring nodes. (2) CASTLE [41] employs a DAG regularizer as Equation 4.2 for improved supervised prediction. (3) CAUSER [105] integrates causal discovery to capture causalities among user sequential behaviors.

4.5.4 Main Results

The model performances of our proposed CBPI and the baselines on both dataset versions are summarized in Table 4.2. Our analysis yields the following insights.

Transformer-based methods (Rows 6-9) outperform traditional neural networks (Rows 1-5). NRMS+BERT and LSTUR+BERT excel over shallow models, emphasizing the value of deep models in capturing rich semantics. However, they overlook the time-evolving nature of news browsing behaviors. IPNR and MINER address this by using external knowledge graphs and category-driven interest vectors, respectively, achieving better results. These findings validate the need to understand latent user interests driving news engagement.

CBPI tops performance charts across all metrics on both datasets, underscoring its effectiveness. Unlike NRMS+BERT, which lacks high-level user interest modeling, CBPI incorporates these factors without explicit signals, leading to superior performance. It also surpasses multi-interest methods enhanced with explicit signals like IPNR and MINER, highlighting the importance of modeling causal relationships in user interests. By aligning interest-level and news-level causality, CBPI uncovers causal patterns from underlying user interests and reduces spurious correlations from content similarity, fostering robust and accurate user representations. The improvement on MIND-small is smaller than that on MIND-large, likely because the smaller dataset has less diversity in user interests, making the learned causal patterns less applicable to a broader user base.

Constructing an interest-level causal graph is crucial for uncovering causal behavior patterns. We evaluated its effectiveness using various graph learning methods (Rows 11-13). All alternatives showed obvious performance declines compared to our causal discovery approach. GAT significantly dropped due to its reliance on statistical correlations and content similarity. While Castle and Causer consider causality and perform better, they treat graph nodes as static features, missing dynamic transitions in news content. In contrast, our approach captures time-invariant causal relations and dynamically models evolving interest concepts, making it more adaptive and suitable for news recommendation.

4.5.5 Ablation Study

To evaluate the impact of each core component within our framework, we conduct an ablation study by systematically removing four key modules: the interest activation loss (*CBPI* w/o *IA*), the interest-level causal graph (*CBPI* w/o *IC*), the news representation reconstruction (*CBPI* w/o *NC*), and the filtering operation (*CBPI* w/o *F*). The results are presented in Table 4.3.

The complete CBPI model achieves the best overall performance, with an AUC of 72.32, an MRR of 36.64, nDCG@5 of 40.27, and nDCG@10 of 46.01. When the interest activation loss is removed, AUC drops to 71.56 and MRR to 36.23, indicating that the interest activation mechanism substantially contributes to the model's ability to match user interests with conceptually similar news. This module narrows the search space for causal discovery by grouping related news items, which in turn leads to more accurate identification of underlying causal relationships. Although the performance of *CBPI* w/o *IA* remains relatively strong, the observed decline highlights the added value of the interest activation loss in refining user interest modeling.

The removal of the interest-level causal discovery module (*CBPI* w/o *IC*) results in a further decrease, with AUC at 69.51 and MRR at 34.73. This suggests that capturing the dynamic and evolving nature of user interests through causal structures is important for effective user representation. However, even without this component, the model still outperforms most baseline methods, which underscores the robustness of the remaining modules. Nevertheless, in the absence of interest-level causality modeling, the framework's performance lags behind state-of-the-art models such as IPNR and MINER, particularly because it does not utilize explicit semantic features or news categories.

The news representation reconstruction module also plays a critical role. When this component is omitted (*CBPI* w/o *NC*), the AUC drops to 71.96 and MRR to 36.39. This decline is more significant than the drop observed when the filtering module is removed (*CBPI* w/o *F*), which achieves an AUC of 72.08 and MRR of 36.51. This comparison indicates that reconstructing news representations to explicitly encode causal behavior patterns across the user's full historical context is more impactful than merely filtering out irrelevant news for each candidate. While the filtering operation helps to reduce noise in the recommendation process, the reconstruction of news

Table 4.3: Effects of different CBPI components.

Methods	MIND-large			
	AUC	MRR	nDCG@5	nDCG@10
CBPI ω/o IA	71.56	36.23	39.77	45.45
CBPI ω/o IC	69.51	34.73	37.71	43.90
CBPI ω/o NC	71.96	36.39	39.89	45.56
CBPI ω/o F	72.08	36.51	40.01	45.81
CBPI	72.32	36.64	40.27	46.01

embeddings enables the model to develop a deeper understanding of causal user-news interactions.

In conclusion, the ablation results demonstrate that each module in the CBPI framework contributes to the overall performance. Among them, the interest activation and news reconstruction modules are particularly crucial for precise modeling of user interests and causal relationships, while the interest-level causal graph and filtering operation provide further enhancements in capturing dynamic user behaviors and reducing irrelevant information.

4.5.6 Hyperparameter Analysis

We further investigate the influence of two key hyperparameters in the CBPI framework: the number of latent concepts K and the filtering threshold τ. All other parameters are fixed at their optimal values during this analysis to ensure fair comparison. The results are illustrated in Figure 4.3.

The left panel of Figure 4.3 demonstrates how model performance varies with different values of K. Both AUC and nDCG@10 improve as K increases from 4 to 16,

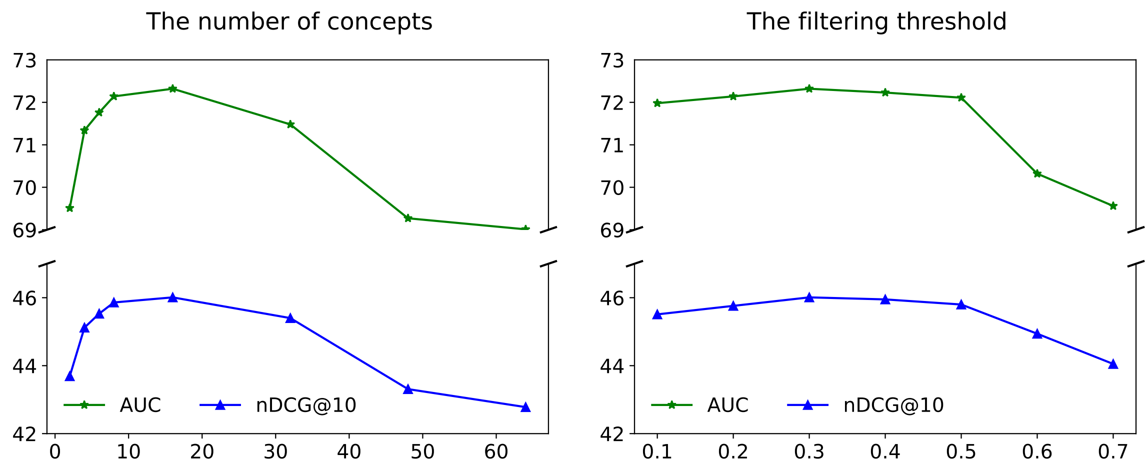


Figure 4.3: The influence of hyperparameters on MIND-large.

reaching their highest values at $K = 16$. When K is set to 16, the model achieves an AUC of approximately 72.3 and an nDCG@10 of about 46.0, indicating a well-balanced representation of user interests. Increasing K beyond this point leads to a gradual decline in both metrics. This trend suggests that a moderate number of concepts allows the model to sufficiently capture the diversity in user interests without introducing unnecessary complexity. When K becomes too large, the model may overfit or fragment user interests, which reduces its ability to generalize. Conversely, when K is too small, the model lacks the expressiveness to describe the full range of user preferences, resulting in suboptimal performance.

The right panel of Figure 4.3 presents the effect of varying the filtering threshold τ. The model achieves optimal results when τ is set to 0.3, with AUC and nDCG@10 again reaching peak values. Lowering τ below 0.3 causes a decline in performance, likely because more irrelevant news items are retained, introducing noise and weakening the model's ability to focus on causally relevant behaviors. Raising τ above 0.3 also leads to reduced performance, as overly strict filtering may eliminate useful historical context and disrupt the modeling of relevant news-level causal relationships. Notably, the model maintains relatively stable performance for τ values in the range from 0.2 to 0.4, indicating that the identification of causal patterns among browsing behaviors is robust to moderate changes in the filtering threshold.

These results highlight the importance of careful hyperparameter selection for maximizing model effectiveness. Choosing an appropriate number of latent concepts and a suitable filtering threshold ensures that CBPI can accurately capture user interest diversity and causal behavior patterns, leading to strong recommendation performance.

4.6 Summary

In this work, we addressed news recommendation as a specialized instance of sequential recommendation, focusing on the complex and dynamic nature of user interests. We proposed the Causal Behavior Pattern Inference (CBPI) framework, which, to our knowledge, is the first approach to introduce causal discovery techniques into multi-interest modeling for news recommendation. Unlike previous methods that rely on static or predefined structures, CBPI automatically learns personalized, latent causal graphs from users' behavioral histories.

CBPI first disentangles each user's browsing sequence into multiple latent interest concepts. Building on recent advances in causal discovery, we designed an encoder-decoder framework to jointly infer time-invariant causal relationships among these interests and to model their temporal evolution. The encoder learns a DAG among the interests, capturing the stable causal dependencies across the user's interest space. The decoder then models how current interests dynamically transition over time, informed by these causal relations. A further innovation of CBPI is the mapping of causal relations from the interest level to the news level. By projecting the learned interest-level DAG onto news items, we construct a news-level causal graph that reveals the intrinsic logical connections between individual pieces of news. This graph enables the model to filter out causally irrelevant news during user representation learning, ensuring that each news item is only influenced by its causal parent nodes in the sequence. As a result, CBPI effectively mitigates spurious correlations and enhances the robustness and rationality of user interest modeling. Extensive experiments on real-world datasets demonstrate that CBPI achieves significant improvements over state-of-the-art baselines. These results validate the importance of uncovering and leveraging causal behavior patterns in sequential user modeling for news recommendation.

Despite these strengths, CBPI also has certain limitations. While it models causal relationships among latent interests and successfully projects them to the news level, it does not explicitly capture the hierarchical interactions between different levels of user intent and observable behaviors. Additionally, the news-level causal graph is derived from interest-level causality and may not strictly satisfy the acyclicity constraints required for formal causal discovery at all levels. This limitation restricts the framework's ability to represent the full complexity and interdependence inherent in user behavior.

These observations highlight the necessity of capturing multi-level causality in user modeling, motivating the development of more comprehensive approaches that can jointly represent hierarchical and cross-level causal dependencies in sequential user behaviors. This direction will be further explored in the following chapter.

Chapter 5

From Interests to Behavior: Hierarchical Causal-Inspired Structure Modeling for Personalized News Recommendation

5.1 Introduction

News recommendation, as a specialized form of sequential recommendation, presents unique challenges due to the dynamic evolution of user interests and the intricate dependencies among user behaviors. Recent advances in causal discovery have demonstrated promising potential for revealing the underlying structures that drive user decision-making processes [145, 41, 44]. However, most existing causal discovery approaches in recommendation systems [105, 120] are limited to learning a flat causal graph over isolated variables, typically assuming homogeneous node semantics. This oversimplification constrains their ability to represent the hierarchical and interdependent nature of user interests and behaviors that are prevalent in news recommendation scenarios. These limitations underscore the importance of hierarchical causal modeling, which can disentangle diverse user interests and explicitly trace their influence on observable news interactions. Building such a principled framework is crucial for capturing the complex, multi-level causal relationships that underlie user browsing sequences. The work presented in this chapter is dedicated to addressing this gap by introducing a unified approach for discovering and leveraging multi-level causal structures in news recommendation.

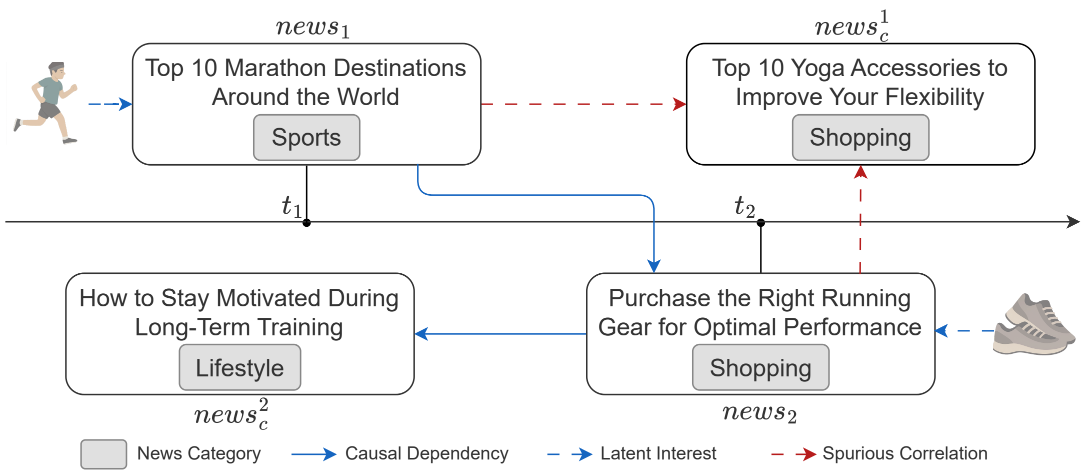


Figure 5.1: Illustration of causal and spurious dependencies in news recommendation.

Our approach is motivated by two key observations regarding user news consumption behavior. First, user behaviors are typically guided by evolving high-level interests or intentions. For example, as illustrated in Figure 5.1, a user may browse $news_1$ due to an interest in marathon events at time t_1, and subsequently read $news_2$ with the intention of purchasing related equipment. These actions reflect semantically distinct but causally connected interests. Second, historical user behaviors often exhibit inherent causal dependencies. In this case, the decision to browse $news_2$ is causally influenced by the prior exposure to marathon-related content in $news_1$. Without this preceding browsing behavior, surface-level attributes such as ”sports” or

”shopping” may not be sufficient to accurately infer the user preference. Consequently, conventional approaches based on statistical associations among news content may misinterpret the user intent. They may overemphasize correlations between keywords and mistakenly recommend irrelevant news (e.g., $news_c^1$), which share only superficial content similarities with $news_1$ (as a sports-related article) and $news_2$ (as shopping for gear), but lack the underlying causal context that drives user behavior. In contrast, a causally-informed perspective can recognize $news_2^c$ as a plausible candidate, as it aligns with the user's underlying decision process shaped by prior interactions—even without obvious topical or categorical similarity.

These observations highlight the need for a principled framework that accounts for two complementary forms of causality: one capturing how latent user interests give rise to observable actions, and the other modeling how user behaviors exhibit structural causal dependencies beyond temporal order. In response, we propose Hierarchical Causal-Inspired Representation (HCR), a unified framework designed to discover and leverage multi-level causal structures within user behavior sequences for news recommendation. To be specific, HCR consists of three key components. First, Latent Interest Discovery disentangles users' browsing history into multiple semantically coherent latent interests via hierarchical clustering and cluster-specific attention, providing an interpretable abstraction of user preferences. Second, Hierarchical Causal Graph Learning constructs a two-level causal graph through differentiable causal discovery. By combining latent interest representations and historical behavior embeddings, we employ a set of behavior-specific neural networks to reconstruct each behavior embedding from these entangled inputs. The reconstruction weights are interpreted as causal influence scores, enabling the discovery of both interest-to-behavior and intra-behavior causal dependencies. The differentiable DAG learner is constrained to reflect the nature of user preferences in news recommendation, encouraging both temporal consistency and semantically meaningful causal structures. This hierarchical causal graph provides a principled foundation for understanding and modeling the dynamics of user behaviors. Finally, Causal-Aware User Representation aggregates user behaviors under the guidance of the learned hierarchical causal graph, effectively alleviating spurious correlations and producing robust, interpretable user representations for accurate recommendation.

The key contributions of this work are outlined as follows:

- We identify and formally model two complementary types of causal relationships in user behaviors: one reflecting how latent user interests give rise to observable actions, and the other capturing the underlying dependencies among historical behaviors.
- We develop a hierarchical causal modeling framework that jointly incorporates user interests and historical behaviors through differentiable causal discovery, with architectural constraints tailored to the characteristics of news recommendation scenarios.
- We theoretically analyze how discovering the causal structure can help capture robust multi-interests for news recommendation. Extensive experiments on the real-world datasets demonstrate that our method consistently outperforms state-of-the-art baselines in terms of recommendation accuracy and diversity.

5.2 Problem Statement

5.2.1 Task Definition

Let u denote a user with a chronologically ordered history of N browsed news articles, represented as $\mathcal{D}^u = [D_1, D_2, \ldots, D_N]$. As illustrated in Section 4.3.1, the news recommendation task is to estimate the likelihood that user u will interact with a candidate news article D_c. Each news article D_i is mapped to an embedding vector $\boldsymbol{e}_i$

via a news encoder, resulting in historical news embeddings $E_u = [\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_N]$. The candidate news D_c is similarly encoded as $\boldsymbol{e}_c$. The final matching score is computed as the dot product $y_u = \boldsymbol{e}_c^\top \cdot \boldsymbol{u}$, where $\boldsymbol{u}$ represents the user's current preferences.

5.2.2 Reflection on Interest Matching

In this section, we explore the necessity of discovering causality to better capture and preserve diverse user interests in news recommendation.

As previously discussed, user browsing behaviors are typically driven by multiple latent interests. To enhance recommendation performance, it is crucial not only to accurately model these interests but also to ensure that the distribution of user interests reflected in the recommended list aligns with that in the user's historical behaviors. In recommendation systems, *interest calibration*, i.e., matching the distribution of recommended content to the user's underlying interest distribution, has been widely recognized as an effective proxy for measuring personalized diversity [80, 62, 103].

Formally, the calibration of user interest can be measured by the following divergence:

$$Q(p, q) = \sum_{z_k \in Z^u} p(z_k|u) \log \frac{p(z_k|u)}{q(z_k|u)}, \tag{5.1}$$

where Z^u denotes the set of latent interests for user u, and z_k represents the k-th interest factor. Here, $p(z_k|u)$ is the distribution of user interests estimated from historical interactions $\mathcal{D}^u$, while $q(z_k|u)$ is the distribution inferred from the recommended list $\mathcal{S}^u$. Specifically, they are computed as:

$$p(z_k|u) = \frac{1}{|\mathcal{D}^u|} \sum_{D_i \in \mathcal{D}^u} P(z_k|D_i, u), \quad q(z_k|u) = \frac{1}{|\mathcal{S}^u|} \sum_{D_i \in \mathcal{S}^u} P(z_k|D_i, u), \tag{5.2}$$

where $P(z_k|D_i, u)$ measures the likelihood that news D_i is associated with interest z_k for user u.

We adopt KL divergence in Equation (5.1) due to its asymmetry, which places greater emphasis on missing interest modes in the recommendation distribution. This property aligns with the goal of avoiding under-representation of user interests. Moreover, KL divergence provides a principled way to quantify the semantic discrepancy between historical behaviors and recommendations, serving as a meaningful indicator of interest mismatch [80, 62].

To achieve robust interest calibration, it is desirable that $P(z_k|D_i, u)$ remains invariant across different environments, i.e., $P(z_k|D_i, u) = P'(z_k|D_i, u)$ even when the joint distributions $P(D_i, u)$ and $P'(D_i, u)$ differ. Such invariance ensures that the model's understanding of user interests remains stable under distributional shifts between users' historical behaviors and candidate news items.

According to Bayes' theorem, $P(z_k|D_i, u)$ can be expanded as:

$$P(z_k|D_i, u) = \frac{P(D_i|u, z_k)P(u|z_k)P(z_k)}{\sum_{z_{k'} \in Z^u} P(D_i|u, z_{k'})P(u|z_{k'})P(z_{k'})}. \tag{5.3}$$

To better understand the factors influencing the stability of $P(z_k|D_i, u)$, we further analyze Equation 5.3. The denominator is a summation over all possible interests, weighted by $P(D_i|u, z_{k'})P(u|z_{k'})P(z_{k'})$. In practice, the interacted news at a specific time is often dominated by limited latent interests, leading to a sparse activation pattern over Z^u. Under this assumption, the contribution from irrelevant interests $z_{k'}$ becomes negligible, and the denominator is primarily determined by the terms associated with the most relevant interests. Thus, maintaining the stability of the numerator, $P(D_i|u, z_k)P(u|z_k)P(z_k)$, becomes critical for ensuring the invariance of $P(z_k|D_i, u)$. Since the prior $P(z_k)$ is typically assumed to be static or minimally impacted by environmental shifts, the key factors that determine the robustness of $P(z_k|D_i, u)$ are $P(D_i|u, z_k)$ and $P(u|z_k)$. Alternatively, from a causal perspective, ensuring that $D_i \perp\!\!\!\perp \mathrm{Env} \,|\, (u, z_k)$ and $u \perp\!\!\!\perp \mathrm{Env} \,|\, z_k$ would imply that the conditional distributions $P(D_i|u, z_k)$ and $P(u|z_k)$ are stable across different environments. These insights motivate the need to explicitly model the underlying causal processes that govern user behaviors and their interactions with news items.

Inspired by recent advances in causal discovery for recommendation [29, 107], we argue that achieving robust interest matching requires simultaneously modeling two types of inherent causal relationships. The first concerns behavior-level causality, which describes how historical user behaviors causally influence subsequent news interactions and thus preserves the invariance of $P(D_i|u, z_k)$. The second concerns interest-to-behavior causality, which models how latent user interests generate observable user behaviors and thereby maintains the invariance of $P(u|z_k)$. Accordingly, we propose constructing a hierarchical causal graph for each user, where one layer captures the causal dependencies among interacted news articles, and the other models the generative process from latent interests to interactions. Based on the learned causal structure, we reconstruct behavior representations that better reflect the underlying drivers of user actions, thereby enabling more accurate, robust, and interest-calibrated recommendations.

5.3 Methodology

As illustrated in Figure 5.2, the proposed model consists of three main components: (1) Latent Interest Discovery, which identifies underlying user interests from behavioral sequences; (2) Hierarchical Causal Graph Learning, which constructs a hierarchical causal graph to model both the relationships between user interests and behaviors as well as dependencies among behaviors themselves; and (3) Causal-Aware User Representation, which aggregates user behaviors under causal guidance to generate the final prediction.

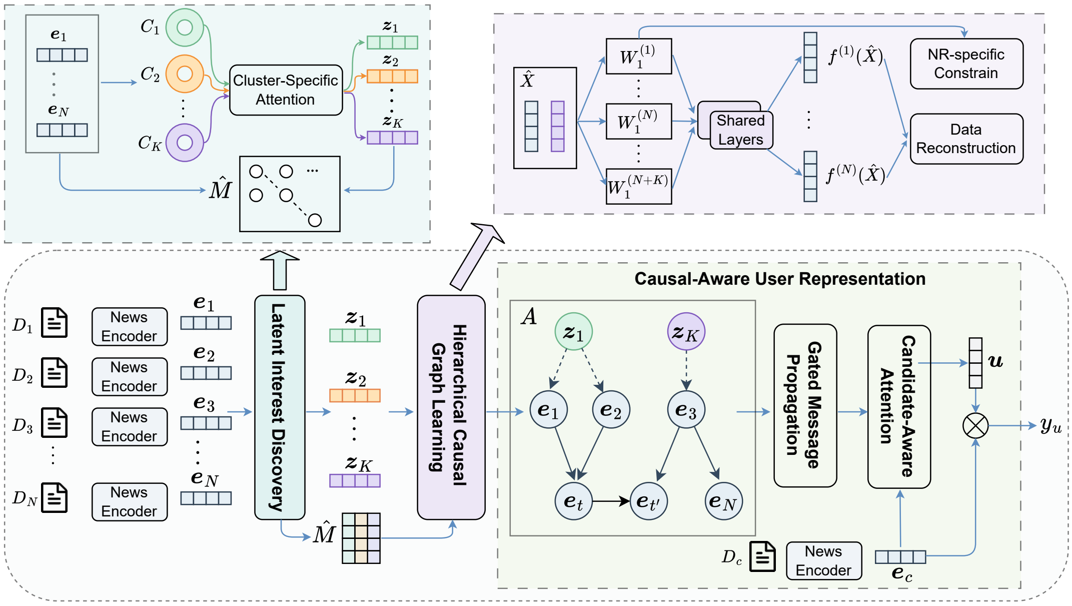


Figure 5.2: Framework of HCR, which consists of three key components: latent interest discovery, hierarchical causal graph learning, and causal-aware user representation.

5.3.1 Latent Interest Discovery

Users often exhibit multiple latent interests that evolve over time and jointly influence their news browsing behavior. These interests are typically implicit and cannot be fully captured by observable features such as news categories or metadata. To enable hierarchical structure modeling over user preferences in subsequent modules, we first derive a set of disentangled latent interest representations that serve as a compact and interpretable abstraction of user intents.

Unlike existing methods of interest matching that apply multi-head attention over the full interaction sequence, which can lead to interest entanglement and cross-topic interference, we explicitly separate distinct user interests during representation learning. The objective is to generate high-level, semantically coherent interest representations that are both informative and independent.

Traditional approaches often apply multi-head attention over the entire interaction sequence to infer user interests [43, 140]. However, such methods tend to entangle unrelated topics, leading to noisy and overlapping interest representations. In contrast, we explicitly separate distinct interests during representation learning to ensure semantic purity and mutual independence.

Given a user's historical interaction sequence $E_u = [\boldsymbol{e}_1, \boldsymbol{e}_2, \dots, \boldsymbol{e}_N] \in \mathbb{R}^{N\times d}$, where each $\boldsymbol{e}_i$ is encoded by a news encoder, we first apply the average linkage clustering to group the news items into K clusters. This method provides a simple yet effective way to group semantically similar news articles, without relying on strong assumptions about cluster shape or number of points. Each cluster $\mathcal{C}_k = [\boldsymbol{e}_1, \dots, \boldsymbol{e}_{|C_k|}]$ contains news pieces that are semantically similar and reflects a coherent aspect of the user's interests.

To summarize each cluster into a latent interest vector $\boldsymbol{z}_k$, we employ a self-attention mechanism that captures contextual relevance within the cluster:

$$\alpha_t^z = \mathrm{softmax}(\boldsymbol{v}_k^\top \tanh(W\boldsymbol{e}_t + \boldsymbol{b})), \tag{5.4}$$

$$\boldsymbol{z}_k = \sum_{t=1}^{|C_k|} \alpha_t^z \boldsymbol{e}_t, \tag{5.5}$$

where $\boldsymbol{v}_k$, W, $\boldsymbol{b}$ are learnable parameters, and α_t^z denotes the importance of the t-th news item within cluster C_k. By applying attention independently within each cluster, the model ensures that the resulting interest vectors are disentangled, each of which summarizes a distinct, semantically grounded component of user preference. We denote the set of interest representations as $Z_u = [\boldsymbol{z}_1, \dots, \boldsymbol{z}_K] \in \mathbb{R}^{K\times d}$.

To further support downstream causal graph learning, we construct an interest assignment matrix $\hat{M} \in \mathbb{R}^{N\times K}$, where each entry $\hat{M}_{i,k}$ measures the probability that the interaction $\boldsymbol{e}_i$ is driven by the k-th latent interest. Specifically, we compute:

$$\hat{M}_{i,k} = \frac{\exp(\mathrm{sim}(\boldsymbol{e}_i, \boldsymbol{z}_k))}{\sum_{j=1}^{K} \exp(\mathrm{sim}(\boldsymbol{e}_i, \boldsymbol{z}_j))}, \tag{5.6}$$

where $\mathrm{sim}(\cdot, \cdot)$ denotes the cosine similarity. This matrix provides a soft and informative mapping between user interactions and latent interests, enabling structure-aware modeling in the causal graph learning module.

5.3.2 Hierarchical Causal Graph Learning

To uncover the underlying causal structure of user preferences, we perform causal discovery over both historical interactions and latent interests. Specifically, we aim to learn a hierarchical causal graph that captures two types of dependencies: (1) intra-behavior causality, describing how different user behaviors are causally related, and (2) interest-to-behavior causality, modeling how latent user interests drive observable actions of browsing news articles.

We formulate this causal discovery process as a representation reconstruction problem. Given the historical interaction embeddings E_u and latent interest vectors Z_u, we concatenate them into a unified input $\hat{X} = [E_u \| Z_u]$ and feed it into a set of behavior-specific neural networks $f_\theta = \{f^{(n)}\}_{n=1}^{N}$, where each $f^{(n)}$ is responsible for reconstructing the n-th interaction embedding:

$$f^{(n)} = W_2 \,\mathrm{ReLU}(W_1^{(n)} \hat{X}), \tag{5.7}$$

where $W_1^{(n)} \in \mathbb{R}^{H \times (N+K)}$ is the input projection layer for the n-th interaction, and $W_2 \in \mathbb{R}^{N \times d}$ is a shared output layer. To prevent trivial self-loops, the n-th row of $W_1^{(n)}$ is masked out. The data reconstruction loss is defined as:

$$\mathcal{L}_{rec} = \sum_{u \in \mathcal{D}} \|E_u - f_\theta(\hat{X})\|^2. \tag{5.8}$$

This reconstruction process implicitly defines a causal adjacency matrix $A \in \mathbb{R}^{(N+K) \times N}$, where the upper submatrix $A^E \in \mathbb{R}^{N \times N}$ encodes causal dependencies among user interactions, and the lower submatrix $A^Z \in \mathbb{R}^{K \times N}$ captures how each latent interest contributes to the decision of interacted news. Compared to conventional causal discovery methods that operate on homogeneous variables (e.g., all nodes representing the same type of observed behavior), our framework explicitly models heterogeneous node types (i.e., user behaviors and latent interests) within a unified causal graph. This enables our model to reason not only about sequential influence among behaviors but also about how abstract user interests manifest in concrete actions, which is particularly important in the news recommendation system, where intent shifts are frequent and subtle.

To estimate the causal strength between nodes, we extract the effective influence weights from the first layer of each reconstruction network. For intra-behavior causality, we define:

$$A^E_{i,j} = \|(W_1^{(j)})_{i,:} + \hat{M}_{i,:}(W_1^{(j)})_{N:,:}\|_2, \tag{5.9}$$

where the first term measures the direct causal effect of $\boldsymbol{e}_i$ on $\boldsymbol{e}_j$, while the second term accounts for the indirect influence mediated via latent interests, using the interest assignment matrix $\hat{M}$. The ℓ_2-norm aggregates the influence across all dimensions into a scalar that reflects the overall causal strength. In the context of news recommendation, a higher value of $A^E_{i,j}$ suggests that browsing news item D_i increases the likelihood of subsequently browsing D_j, i.e., the behavior at time i is a potential prerequisite for the interaction at time j. In contrast to mere correlation, this formulation reflects directional influence, allowing the model to better capture behavioral dependencies that would be disrupted if earlier interactions were absent.

Likewise, the strength of interest-to-behavior causality is computed as $A^Z_{k,j} = \|(W_1^{(j)})_{N+k,:}\|_2$, which quantifies how the k-th latent interest indicates the user's decision to browse news D_j.

To enforce acyclicity in the resulting graph without resorting to matrix exponentials [44], we first compute $D = A \circ A$, where $\circ$ denotes element-wise multiplication. We then perform Schur decomposition on D to obtain an upper triangular matrix R. The graph is acyclic when the squared sum of the diagonal elements of R equals zero, and we formalize this requirement with the constraint $\mathcal{L}_{acyc}(A) = \|\text{diag}(R)\|_2^2$.

At the same time, we encourage sparsity in the learned structure by incorporating an ℓ_1-norm regularization term, $\mathcal{L}_{spa}(A)$, as detailed in Section 4.3.2.

In addition, we introduce two architectural regularization terms tailored to the news recommendation scenario. First, since user behaviors are typically driven by some latent intent, we encourage all behavior nodes to have non-zero in-degree:

$$\mathcal{L}_{nzd} = \sum_{i=1}^{N} \log \|A_{:,i}\|. \tag{5.10}$$

Second, although we do not impose strict temporal ordering, we discourage implausible chronological violations by penalizing extreme edge directions at the sequence boundaries:

$$\mathcal{L}_{chr} = \sum_{i=1}^{N} |A(i,1)| + \sum_{j=1}^{N} |A(N,j)|, \tag{5.11}$$

which ensures the first news browsing behavior has no parents while the last one has no causal descendants, preserving basic temporal coherence.

The overall objective for hierarchical causal graph learning specific to news recommendation is:

$$\mathcal{L}_{dag}^{NR} = \mathcal{L}_{rec} + \lambda_1 \mathcal{L}_{acyc} + \lambda_2 \mathcal{L}_{spa} + \lambda_3 \mathcal{L}_{nzd} + \lambda_4 \mathcal{L}_{chr}, \tag{5.12}$$

By integrating multi-level causal signals and NR-specific architectural constraints into a unified structure, our model goes beyond traditional single-layer causal discovery approaches.

5.3.3 Causal-Aware User Representation

To enable interpretable preference modeling, we propose a causal-aware user representation based on a causal propagation mechanism over the learned hierarchical structure. Unlike prior methods that aggregate user interactions E_u using simple pooling or attention [116, 112], we leverage the learned hierarchical causal graph

$A \in \mathbb{R}^{(N+K)\times N}$ as a structural prior to incorporate multi-hop dependencies among user behaviors and latent interests. Furthermore, to enhance the sparsity and reliability of the learned causal graph, we apply a thresholding operation to the adjacency matrix A prior to propagation, retaining only edges with causal scores exceeding a predefined threshold τ.

Let $H^{(0)} \in \mathbb{R}^{(N+K)\times d}$ represent the initial set of node embeddings, which include both the user's historical behavior sequence E_u and latent interest vectors Z_u. Notably, E_u has already been refined through the causal graph learning objective defined in Equation 5.8, allowing the model to capture more informative and causally grounded behavioral semantics. To incorporate the structural dependencies encoded in the adjacency matrix A, we apply a two-step gated message propagation to update the behavior node representations accordingly:

$$\boldsymbol{m}_j^{(l)} = \sum_{i=1}^{N+K} A_{i,j} \cdot W_m \cdot \boldsymbol{h}_i^{(l-1)}, \qquad \boldsymbol{h}_j^{(l)} = \mathrm{GRU}(\boldsymbol{m}_j^{(l)}, \boldsymbol{h}_j^{(l-1)}), \tag{5.13}$$

where W_m is a shared transformation matrix and $\mathrm{GRU}(\cdot)$ denotes a gated recurrent unit that regulates information integration. Representations of interest nodes Z_u remain unchanged during propagation. After two iterations of causal propagation, we obtain updated interaction representations $\boldsymbol{h}_t^{(2)}$ at each time step, which encode both direct and high-level causal signals.

To construct a candidate-aware user representation, we employ an attention mechanism to estimate the relevance between the candidate news representation $\boldsymbol{e}_c$ and each updated interaction $\boldsymbol{h}_t^{(2)}$:

$$\alpha_t = \mathrm{softmax}\left(\boldsymbol{e}_c^\top W_a \boldsymbol{h}_t^{(2)}\right), \qquad \boldsymbol{u} = \sum_{t=1}^{N} \alpha_t \cdot \boldsymbol{h}_t^{(2)}, \tag{5.14}$$

where the attention weights α_t can be considered as a form of counterfactual reasoning, allowing the model to emphasize behaviors that are most likely to causally support the engagement with the candidate news.

The final recommendation score is computed via the dot product between the candidate news representation and the user representations: $y = \boldsymbol{e}_c^\top \cdot \boldsymbol{u}$. Following standard practice [116, 140], we adopt negative sampling to optimize the model by distinguishing the clicked item y_u^+ from L negative samples $\{y_u^j\}_{j=1}^L$ using a softmax-based log-likelihood:

$$\mathcal{L}_{\text{NR}} = -\sum_{u \in \mathcal{D}} \log \frac{\exp(y_u^+)}{\exp(y_u^+) + \sum_{j=1}^{L} \exp(y_u^j)}. \tag{5.15}$$

To ensure a valid causal structure, we further incorporate a DAG regularization term tailored for personalized news recommendation into the final objective:

$$\mathcal{L} = \mathcal{L}_{\text{NR}} + \gamma \cdot \mathcal{L}_{\text{dag}}^{\text{NR}}, \tag{5.16}$$

where γ is a hyperparameter balancing the recommendation loss and the NR-specific DAG constraint.

5.3.4 Computational Complexity and Scalability

To support the practical deployment of the proposed HCR framework, we present a detailed analysis of its computational complexity and discuss its scalability in large-scale recommendation scenarios.

The overall computation for each user comprises three primary components: latent interest discovery, hierarchical causal graph learning, and causal-aware user representation.

Latent Interest Discovery involves performing hierarchical clustering on historical news embeddings, which incurs a worst-case complexity of $\mathcal{O}(N^2)$. Given that N is typically fixed to a relatively small constant in practice (e.g., 50), this cost is limited and manageable. The subsequent cluster-level self-attention and the construction of the interest-to-behavior assignment matrix require an additional complexity of

$\mathcal{O}(NKd)$.

Hierarchical Causal Graph Learning constitutes the most computationally intensive part of the model. Reconstructing N behavioral embeddings from $N + K$ inputs using a set of behavior-specific neural networks results in a complexity of $\mathcal{O}(N(N + K)H)$. In addition, enforcing the differentiable DAG constraint on the causal graph introduces a complexity of $\mathcal{O}((N + K)^3)$. These computations are performed independently for each user and can be efficiently parallelized using GPU-based acceleration and shared parameterization.

Causal-Aware User Representation incorporates gated message propagation over the learned hierarchical causal graph. With T propagation steps, the total complexity is $\mathcal{O}(T(N + K)Nd)$, assuming the graph remains sparsely connected after pruning. This is followed by a candidate-aware attention aggregation step with complexity $\mathcal{O}(Nd)$.

By aggregating the computational costs of all modules, the total complexity during training becomes:

$$\mathcal{O}\left(N^2 + NKd + N(N + K)H + (N + K)^3 + T(N + K)Nd\right). \tag{5.17}$$

In applied settings, the parameters N, K, and T are typically fixed to small values. As a result, the overall complexity scales linearly with respect to batch size and embedding dimensions, which ensures computational efficiency. Furthermore, a threshold-based pruning strategy is applied to the learned causal graph, effectively reducing the number of edges and thereby accelerating both the message-passing process and inference.

Regarding scalability, the HCR framework is designed to be compatible with large-scale deployment. Most operations within each module, such as attention computation, neural reconstruction, and graph-based propagation, involve matrix and tensor operations that are highly parallelizable and well-suited for GPU acceleration. Although the overall framework is executed in a sequential, end-to-end manner, its internal components can be efficiently implemented on modern GPU hardware to ensure scalability in large-scale settings. In addition, the use of fixed-length user histories and shared network components contributes to bounded memory usage and stable runtime across varying user interaction patterns. These properties collectively enable HCR to operate efficiently on large-scale datasets such as MIND-large and support real-time inference within practical news recommendation systems.

Compared with CBPI proposed in Chapter 4, which involves quadratic operations over the full user history during concept-level and news-level reconstruction, HCR avoids repeated dense matrix multiplications and significantly reduces costly $\mathcal{O}(N^2 \cdot d)$ terms through efficient hierarchical abstraction and sparse causal propagation.

5.4 Experiment

5.4.1 Experimental Settings

At the beginning of our experimental setup, we employ the same real-world news recommendation dataset as described in Section 4.5.1, namely MIND [119], which includes both the MIND-large and MIND-small versions. All experimental protocols, including the use of MIND-large and its sampled subset MIND-small, as well as data preprocessing and split strategies, are kept consistent with the previous setting. Following [113, 112], we construct each user's historical interaction sequence using their 50 most recent clicked news items. Training is performed with mini-batches of size 64, and each positive instance is paired with 4 negative samples.

We set the number of latent interests K to 8 and the filtering threshold τ to 0.4, based on preliminary validation performance and insights from prior studies. These values strike a balance between representation expressiveness and graph sparsity, as further analyzed in Section 5.4.7. We also observe that the model remains relatively stable within a moderate range of K and τ, suggesting that the learned causal structure is robust to small variations in these settings. The hyperparameter γ, which controls the strength of the overall DAG regularization, is tuned to 0.2 based on validation performance. For the NR-specific DAG loss in Equation 5.12, we set the weighting coefficients to $\lambda_1 = 0.3$, $\lambda_2 = 0.1$, $\lambda_3 = 0.2$, and $\lambda_4 = 0.2$, with values selected through grid search on the validation set, after fixing K and τ.

For evaluation, we adopt standard metrics widely used in news recommendation, including AUC, MRR, nDCG@5, and nDCG@10, following previous studies [126, 143]. To isolate and assess the model's core reasoning capabilities without external knowledge, we only use the news titles as input to the encoder. The news encoder is implemented using a pre-trained BERT model, with the output embedding dimension set to 128.

5.4.2 Baseline Methods

To ensure fair and focused comparison, we exclude two categories of methods: (1) those that rely on explicit knowledge graph reasoning or entity-level alignment [36, 49], which introduce additional semantic supervision beyond our scope; and (2) those that assume static item IDs or explicit treatment assignments [120, 29], which are incompatible with the dynamic, content-driven nature of news recommendation. We compare HCR against representative baselines that are compatible with the dynamic and content-based nature of news recommendation. These baselines are grouped according to their primary user modeling paradigms. For each baseline, we adopt the best-performing hyperparameter settings reported in the original papers when their experimental setups were consistent with ours. If the original configurations were not directly applicable or key hyperparameters were not specified, we conducted additional tuning on a validation set. All models were trained and evaluated using the same data splits, negative sampling strategy, and evaluation metrics to ensure a fair and consistent comparison.

Feature-based Modeling. These methods rely on manually designed features and shallow collaborative filtering techniques to model user-news interactions.

- **LibFM** [65]: Performs matrix factorization based on user/item IDs and sparse textual features obtained from news titles, such as TF-IDF embeddings.
- **DeepFM** [21]: Integrates factorization machines with deep neural networks to capture both first-order and complex high-order feature interactions in a unified framework.

Sequential User Modeling. These methods model user interests as a single evolving sequence of historical interactions.

- **DKN** [89]: Encodes news content using CNNs and incorporates a target-aware attention mechanism to dynamically fuse user representations based on browsing history.
- **NPA** [111]: Introduces a hierarchical attention mechanism over both word-level and article-level features to enhance user modeling.
- **LSTUR** [2]: Captures users' short-term browsing preferences through a GRU-based sequence encoder, while incorporating user ID embeddings to represent long-term interests.
- **NRMS** [113]: leverages multi-head self-attention to model content associations among browsed news and derive context-aware user embeddings.

Graph-based Structural Modeling. These approaches enhance user and news representations by leveraging structural information from global click graphs or user-news interaction networks.

- **NNR** [51]: Integrates bi-directional LSTMs with cross-attention to collaboratively encode news content, and employs graph convolutional networks (GCNs) to capture user structural information from interaction graphs.

- **GLORY** [126]: Incorporates a global click graph into a gated graph neural network to enrich news representations with global behavioral signals.

- **GLoCIMT** [127]: Captures both local and long-range user interests by extracting and fusing long-chain dependencies from a global click graph.

Multi-interest User Modeling. These approaches explicitly model multiple latent user interests to better capture diverse and fine-grained preferences.

- **MIRec** [143]: Introduces a multi-interest modeling framework that explicitly accounts for news popularity, aiming to reduce the bias caused by overly popular items.

- **IPNR** [91]: Identifies latent user interests by matching TF-IDF features with external concepts from ConceptNet and models interest transitions via GCNs. A FastFormer-based encoder is used to produce final user representations.

- **MINER** [43]: Leverages category-aware attention to extract fine-grained user interests from browsing behavior.

Causal-aware Modeling. These approaches leverage causal modeling techniques to improve the accuracy and robustness of learning from sequential user behavior in recommendation systems.

- **CauTailReS** [134]: Models user-side popularity bias using a predefined causal graph and applies do-calculus to disentangle the beneficial and detrimental effects of item popularity. It promotes long-tail exposure through consistency-aware re-ranking and causal embedding. Note that CauTailReS is originally designed for session-based recommendation and not specifically tailored for news recommendation scenarios.

- **TCCM** [10]: Constructs a predefined content-level causal graph that captures the influence of news popularity, timeliness, and content on user interactions. It applies causal intervention during inference to reduce popularity bias in news recommendation.
- **CBPI**: As introduced in Chapter 4, it develops an encoder-decoder framework to infer users' latent multi-interest patterns and their causal dependencies, using an interest-level graph that is projected to the news level.

5.4.3 Main Comparison Results

Table 5.1 presents the performance of various baseline methods and our proposed HCR on the MIND-small and MIND-large datasets. Overall, HCR consistently achieves the best results across all evaluation metrics, demonstrating its superior capability in modeling user preferences for personalized news recommendation.

Traditional feature-based methods (Rows 1-2), such as LibFM and DeepFM, show relatively weak performance. These models rely heavily on shallow ID features and handcrafted interactions, failing to capture rich contextual semantics and the behavioral patterns underlying user clicks. In contrast, sequential user modeling approaches (Rows 3-6), including DKN, LSTUR, and NRMS, improve performance by leveraging the temporal order of user behaviors to extract high-level semantic features. However, they typically model user interests as a single evolving vector, which limits their ability to capture diverse and disentangled user preferences, especially when user behaviors are driven by multiple concurrent interests.

Table 5.1: Performance comparison across various methods (percentages shown without the % sign).

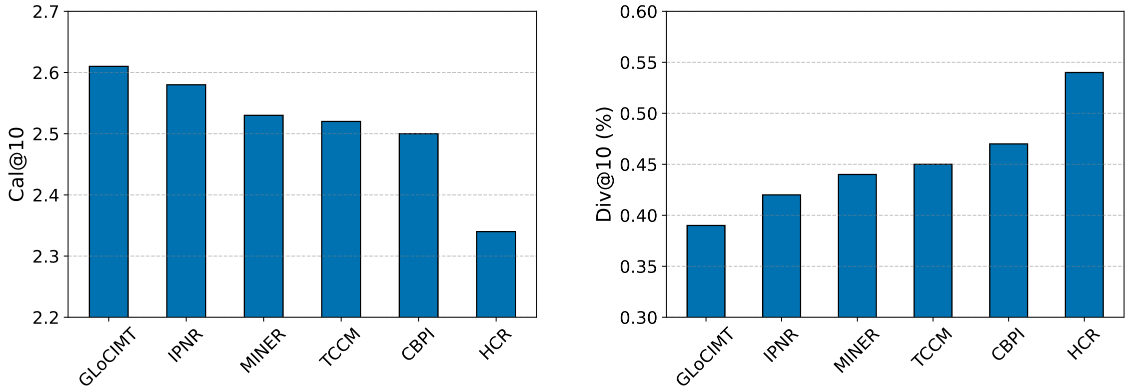
Methods	MIND-small				MIND-large			
	AUC	MRR	nDCG@5	nDCG@10	AUC	MRR	nDCG@5	nDCG@10
LibFM	59.66	26.03	27.90	34.31	60.97	28.16	30.02	36.17
DeepFM	60.42	26.91	29.26	35.89	61.41	28.29	31.21	36.65
DKN	62.75	28.94	31.46	38.02	64.51	30.02	33.04	39.47
NPA	64.51	30.02	33.05	39.46	66.24	31.94	34.52	40.59
LSTUR	65.77	30.85	34.07	40.51	67.01	32.17	35.34	40.78
NRMS	66.20	31.12	34.65	41.17	67.53	31.94	35.45	41.85
NNR	66.96	31.50	35.01	41.41	67.89	32.36	35.93	42.24
GLORY	67.63	32.35	35.74	42.06	68.96	33.80	37.27	43.61
GLoCIMT	68.09	32.91	36.47	42.49	69.38	34.31	37.68	43.95
MIRec	66.34	29.92	33.95	40.46	68.26	33.79	37.72	43.47
IPNR	68.16	32.43	35.98	42.20	69.63	34.42	37.82	44.02
MINER	69.58	33.81	37.61	43.88	71.48	36.15	39.70	45.36
CauTailReS	66.33	31.22	34.62	41.02	67.65	32.08	35.51	41.79
TCCM	69.61	33.65	37.40	43.76	71.63	36.21	39.66	45.29
CBPI	69.73	33.86	37.84	43.97	71.87	36.48	40.10	45.83
HCR	**71.50**	**34.91**	**39.08**	**45.05**	**73.52**	**37.51**	**42.13**	**46.87**

Graph-based structural models (Rows 7-9), achieve further improvements by modeling high-order dependencies among users and news articles through graph neural networks. These methods demonstrate the effectiveness of incorporating global structural signals beyond local interaction histories. For instance, GLoCIMT performs notably well on MIND-large, emphasizing the value of long-chain structural modeling. However, these approaches primarily capture statistical correlations within the graph and lack mechanisms to infer the underlying causal relationships that drive user behavior.

Multi-interest modeling approaches (Rows 10-12) aim to represent users with multiple latent interest vectors, leading to significant gains in accuracy. For example, IPNR incorporates external commonsense knowledge from ConceptNet to infer users' reading intents, while MINER applies category supervision to disentangle news representations into multiple interest-aligned vectors. These methods underscore the importance of modeling diverse latent user interests as key drivers of behavior. However, they generally rely on external annotations or heuristic signals, and their modeling remains rooted in statistical correlation, without uncovering the causal mechanisms that govern how interests influence user decisions.

Causal-aware modeling approaches (Row 13-15) incorporate causal reasoning to improve recommendation robustness. CauTailReS mitigates popularity bias via a predefined user-side causal graph and path-level intervention using do-calculus. It heuristically treats the latest click as short-term interest, without modeling deeper intents or behavioral dependencies. Originally designed for session-based tasks, its assumptions misalign with the complexity of news reading patterns, leading to suboptimal results among causal-aware methods. TCCM mitigates popularity bias by modeling a fixed causal graph over news content, timeliness, and popularity, and applying causal intervention at inference time. While effective in debiasing, it focuses on content-level factors and does not model user intent or behavioral causality, limiting its ability to capture complex user decision patterns. CBPI takes a step further by introducing a causality-aware framework that models dependencies among latent interests, achieving strong empirical results. Nonetheless, it assumes that such interest-level causality is sufficient, without explicitly modeling intra-behavior causality, i.e., how past behaviors causally influence future ones, limiting its capacity to fully capture the dynamics of user decision-making. It is also worth noting that, due to differences in experimental settings, the performance of CBPI reported here may differ from that in Section 4.5.4.

In comparison, HCR outperforms all baselines across both datasets, achieving the highest scores in AUC, MRR, and nDCG metrics. For instance, on MIND-large,



(a) Cal@10 (b) Div@10

Figure 5.3: Performance in terms of calibration and diversity.

HCR attains 73.52% AUC and 46.87% nDCG@10, surpassing CBPI by +2.30% and +2.27%, respectively. These improvements highlight the effectiveness of HCR's hierarchical causal modeling framework, which unifies user interaction history and latent preference signals into a structured representation. Unlike prior approaches that rely on surface-level correlations or require external supervision, HCR discovers causal dependencies directly from behavioral data, enabling more robust and interpretable modeling. The consistent performance gains across all evaluation metrics confirm the advantage of incorporating multi-level causal structure for modeling complex user behavior patterns for personalized news recommendation.

5.4.4 Interest Diversity

In addition to raising recommendation accuracy, prior studies [117, 115, 43] have emphasized the importance of promoting recommendation diversity to alleviate content redundancy and enhance users' browsing experiences. Motivated by the observation that user interests are inherently diverse and often latent, we evaluate whether the generated recommended lists effectively reflect this diversity. Since explicit interest annotations are unavailable in MIND, we follow common practice and use news categories as a proxy for user interests.

For a focused and meaningful comparison, we report interest diversity results for HCR and the top five baselines based on overall recommendation performance (as reported in Table 5.1). These methods represent the strongest competitors and provide a representative benchmark for evaluating diversity.

To quantify interest diversity, we adopt two metrics. First, the calibration score Cal@S measures the alignment between the distribution of recommended news categories and the user's historical category distribution (i.e., we replace the interest distribution in Equation 5.1 with the category distribution). This metric is grounded in KL divergence, which allows us to capture the asymmetry between the two distributions and penalize the under-representation of user interests in recommendations. Its use is consistent with our causal modeling framework, where interest calibration serves as an essential signal for validating the stability of inferred user preferences.

Second, we use the category diversity Div@S, defined as:

$$\text{Div@}S = \frac{\sum_{i=1}^{S}\sum_{j=i+1}^{S}\sigma(\text{CATE}(D_i) \neq \text{CATE}(D_j))}{S \times (S-1)/2}, \tag{5.18}$$

where $\text{CATE}(D_i)$ denotes the category of the i-th news item D_i in the recommended list $\mathcal{S}^u$, and $\sigma(\cdot)$ is the indicator function. This metric calculates the proportion of news item pairs in the top-S list that belong to different categories, thereby reflecting intra-list diversity.

To ensure reliable evaluation, we filter out users in MIND-large whose historical browsing records contain fewer than five unique news categories, as such users lack sufficient signals for assessing personalized interest diversity. We set $S = 10$ and $N = 100$, and report average metric values over all qualified test users in Figure 5.3.

As shown in the results, HCR substantially outperforms all baselines in both Cal@10 and Div@10, despite not relying on any auxiliary category supervision or external knowledge sources. In contrast, the baseline methods exhibit weaker performance, as interest diversity is not explicitly modeled in their training objectives. The superior performance of HCR highlights the effectiveness of its hierarchical causal modeling framework, which not only captures the structure of user preferences but also uncovers the underlying causal relationships between latent interests and user behaviors. By disentangling multiple user intents and modeling how they give rise to observable interactions, HCR is able to generate recommendation lists that are both accurate and well-calibrated to the full spectrum of a user's interests.

5.4.5 Ablation Study

To investigate the contribution of each fundamental part of the HCR framework, we undertake a component-wise ablation study on the MIND-large dataset. Specifically, we construct several degraded variants of HCR by systematically removing or modifying key modules and compare their performance with the full model.

HCR w/o Interest Discovery (IntDisc) This variant removes the latent interest discovery module and represents the user solely through the sequence of historical interaction embeddings. As a result, high-level user intents are not explicitly modeled, and the interest-to-behavior causal edges are omitted from the causal graph.

HCR w/o Causal Graph Learning (CausGl) We disable the hierarchical causal graph learning module by removing the DAG regularization term from Equation 5.16. Message propagation is thus performed based only on the native interest assignment matrix and historical behaviors, without structural guidance from a learned causal graph.

HCR w/o DAG Constraints (DAGCons) We retain the hierarchical structure modeling process but remove the structure-specific regularization terms $\mathcal{L}_{\text{nzd}}$ and $\mathcal{L}_{\text{chr}}$

from Equation 5.12, which are designed to enforce properties of user behavior tailored for personalized news recommendation.

HCR w/o Causal Propagation (CausProp) We replace the gated message propagation mechanism with a standard attention-based aggregation over historical interactions, similar to NRMS [113]. This variant assesses whether causal message passing provides additional benefits beyond conventional attention aggregation.

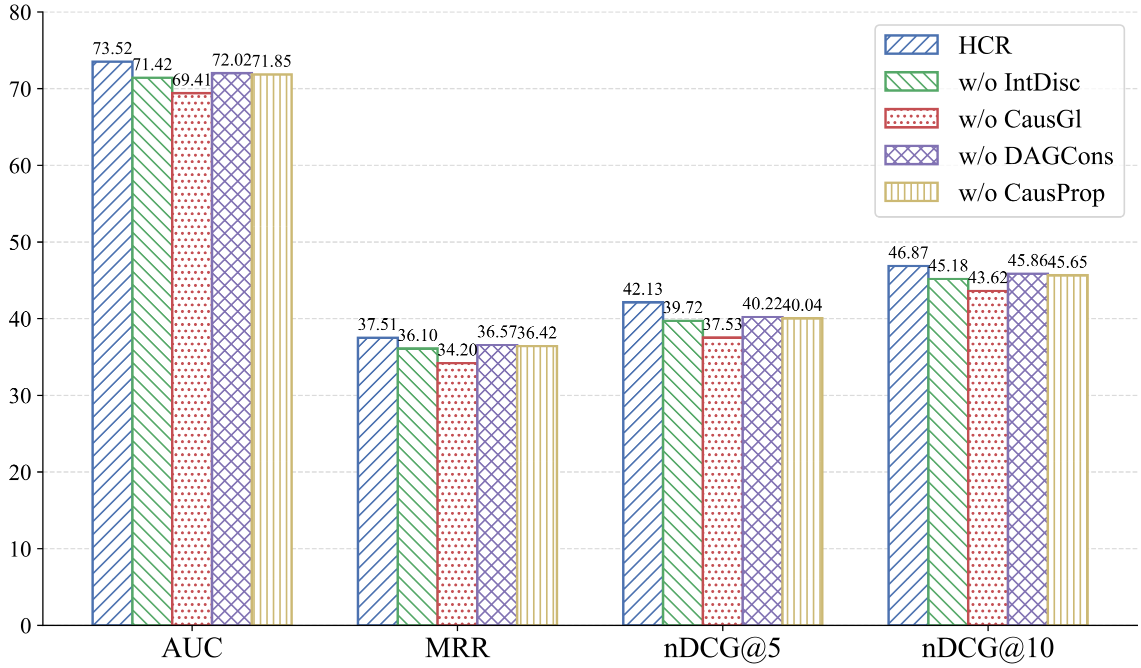


Figure 5.4: Comparison of different degraded variants.

As illustrated in Figure 5.4, the exclusion of any individual component results in a clear decline in performance, validating the necessity of each module. Among all variants, removing the hierarchical causal graph learning module yields the most significant degradation, underscoring the importance of hierarchical causal modeling in capturing behavioral dependencies. Removing intra-behavior causality modeling weakens the model's ability to capture temporal dependencies among user actions, leaving it to rely solely on statistical correlations, which ultimately leads to less robust recommendations.

Eliminating the latent interest discovery module also leads to a substantial decline in performance, highlighting the need to disentangle multi-interest structures for accurate user modeling. Removing DAG constraints results in a moderate performance decrease, suggesting that the structural regularization terms—specific to the news recommendation context—are effective in encouraging meaningful and coherent causal graphs.

The relatively minor performance degradation observed when replacing gated causal propagation with attention-based aggregation suggests that the causal graph already offers substantial guidance in modeling user behaviors. Nonetheless, the gated propagation component remains beneficial, contributing additional refinement to the user representations.

Overall, the findings highlight that each component of HCR is indispensable to its performance, and that their integrated design is essential for capturing the hierarchical causal structure inherent in user preferences.

5.4.6 Graph Construction Variants

To further validate the effectiveness of our proposed hierarchical causal graph learning module, we conduct module-level ablation by replacing it with several representative graph construction methods on the MIND-large dataset. These variants reflect different paradigms of graph-based modeling for user preferences, including attention-based mechanisms, causal regularization, and sequential causal discovery techniques. Specifically, we consider: (1) GAT [88], a graph attention network that captures statistical correlations based on local attention scores computed from interacted news embeddings; (2) CASTLE [41], a supervised causal discovery method that enforces acyclicity through a differentiable DAG constraint (similar to Eq. 4.2); and (3) CAUSER [105], a causal graph learning method tailored for sequential recommendation, which focuses on discovering causal relations among latent user interests.

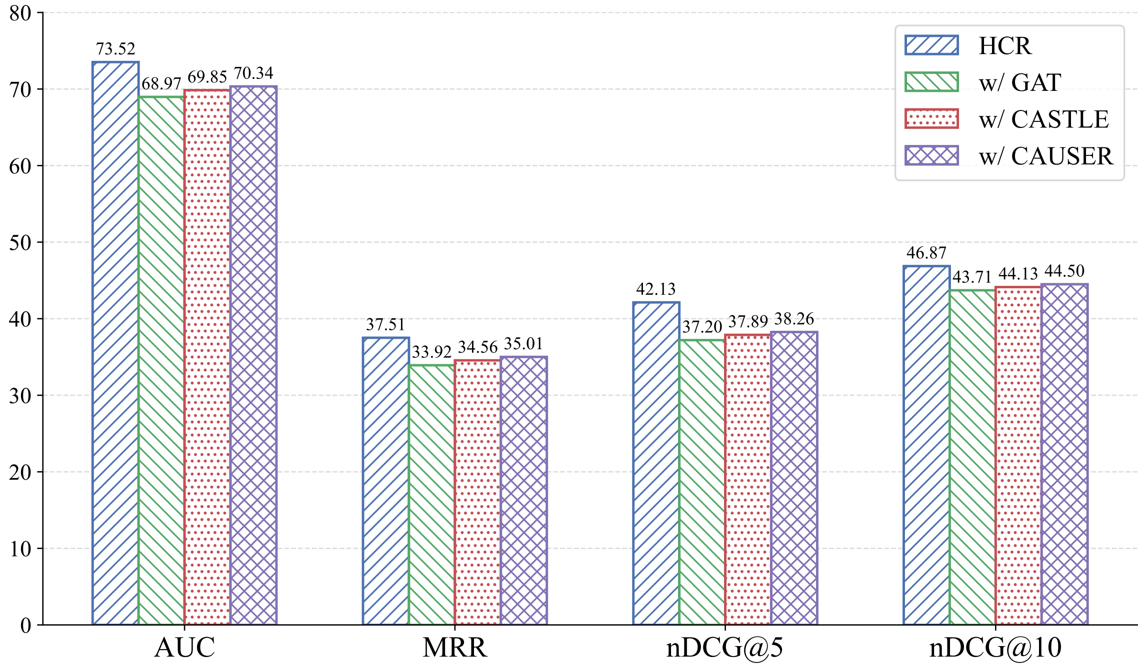


Figure 5.5: Comparison of different graph construction methods.

Experimental results in Figure 5.5 show that replacing our hierarchical causal graph learning module with these alternatives leads to consistent performance degradation. GAT performs the worst due to its reliance on shallow semantic similarity without causal guidance, making it prone to spurious correlations. Although CASTLE and CAUSER incorporate causal modeling and achieve relatively better performance, they are constrained by their flat graph structures that operate on a single level of variables, such as user behaviors or latent interests. Consequently, they are unable to capture the cross-level causal effects between high-level user interests and low-level browsing actions.

These findings underscore a key advantage of our approach: the ability to model causal dependencies across multiple semantic levels. Unlike the baselines, which construct graphs over a homogeneous set of nodes, our method enables hierarchical structure modeling by explicitly separating latent user interests from observable behavioral signals and allowing structured causal propagation between them. This hierarchical design is particularly well-suited to the news recommendation scenario, where user interests are inherently diverse, dynamic, and often sparsely reflected in short-term behaviors.

5.4.7 Hyperparameter Analysis

We investigate the influence of two critical hyperparameters in the HCR framework: the latent interest number K and the threshold τ applied to filter edges in the learned causal graph.

As depicted in Figure 5.6, the model attains optimal performance when $K = 8$. Setting K too low may fail to uncover the diversity of user interests, whereas an excessively large K leads to a fragmented interest space and introduces redundant parameters, making it difficult to form semantically coherent groups. Notably, the performance begins to degrade sharply when K exceeds 12, underscoring the importance of choosing an appropriate granularity for user interest modeling. When too many latent interests are introduced (e.g., $K = 20$), the performance falls even below that of the variant without latent interest discovery. This suggests that excessive interest splitting results in fragmented and noisy representations, which hinder effective information aggregation and may introduce misleading signals during propagation.

Regarding the filtering threshold τ, from Figure 5.7 we observe that the performance peaks around $\tau = 0.4$. A low threshold tends to preserve weak or noisy causal edges, potentially introducing irrelevant information or spurious correlations into the propagation process. Conversely, a higher threshold can over-prune the graph, removing meaningful semantic connections between news items. Interestingly, the model's performance remains relatively stable within a moderate range of τ, indicating that the essential causal structure among user behaviors is robustly captured before the filtering stage. This highlights the effectiveness of our causal discovery mechanism in extracting informative and reliable dependencies.

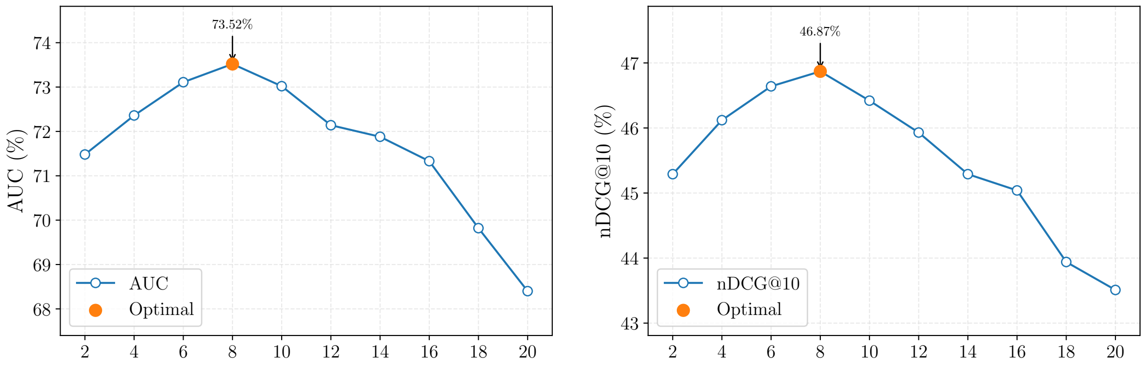


Figure 5.6: The interest number K.

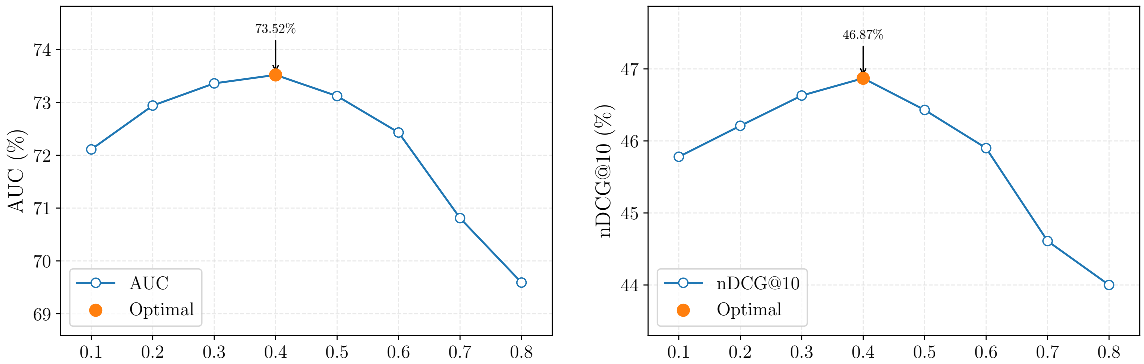
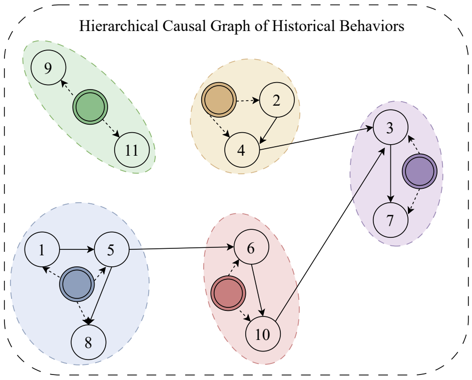


Figure 5.7: The filtering threshold τ.

5.4.8 Case Study

To provide deeper insights into HCR's effectiveness, we conduct a case study based on a sampled user's historical interaction session. As shown in Figure 5.8, we present a hierarchical causal graph inferred from the user's browsing behaviors. In the graph, solid arrows represent intra-behavior causal dependencies (computed via Equation 5.9), while dashed arrows indicate the latent interest most likely responsible for each interaction. The associated interest is identified by selecting the dimension in matrix A^Z with the highest causal effect for the corresponding news item.

ID	Category	History Browsed News
1	TV	Felicity Huffman begins prison sentence for college admissions scam
2	Finance	Tiffany's is selling a holiday advent calendar for $112,000
3	TV	'Wheel Of Fortune' Guest Delivers Hilarious, Off The Rails Introduction
4	Autos	This restored 1968 Winnebago is beyond adorable
5	TV	Lori Loughlin Is 'Absolutely Terrified' After Being Hit With New Charge
6	Movies	Bruce Willis brought Demi Moore to tears after reading her book
7	Entertainment	Celebrity kids then and now: See how they've grown
8	News	Felicity Huffman Smiles as She Begins Community Service Following Prison Release
9	News	Alexandria Ocasio-Cortez apologizes for blocking ex-Brooklyn politician on Twitter
10	Health	The Rock's Gnarly Palm Is a Testament to Life Without Lifting Gloves
11	TV	Donald Trump Jr. reflects on explosive 'View' chat



Recommended by HCR	
Epic 'Jeopardy!' rematch gets off to a hot start	Vedio
Marie Yovanovitch testifies in impeachment hearing before House panel	News
THEN AND NOW: What all your favorite '90s stars are doing today	Entertainment

Recommended by CBPI	Category
Meghan Markle and Hillary Clinton Secretly Spent the Afternoon Together	Lifestyle
Marie Yovanovitch testifies in impeachment hearing before House panel	News
Pete Davidson, Kaia Gerber Are Dating	TV

Recommended by MINER	Category
Kardashians Face Backlash Over 'Insensitive' Food Fight in KUWTK Clip	TV
Pete Davidson, Kaia Gerber Are Dating	TV
THEN AND NOW: What all your favorite '90s stars are doing today	Entertainment

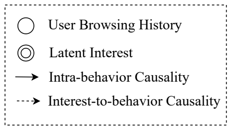


Figure 5.8: Case study of HCR's causal-inspired structure modeling and recommendation outcomes. The top-left part shows the historical browsed news of a sample user, and the top-right displays the corresponding hierarchical causal graph inferred by HCR, in which each colored cluster represents a latent interest. The bottom section compares the top-3 recommendations generated by HCR, CBPI, and MINER, respectively. Ground-truth browsed news is highlighted in blue.

In the lower part of the figure, we compare the top-3 recommendations generated by HCR, CBPI, and MINER, alongside the user's actual clicks (highlighted in blue). We can observe that MINER tends to rely on surface-level content similarity, leading to recommended items from overlapping categories (e.g., TV), which fail to reflect the user's deeper interests. CBPI, while capable of identifying related latent interests, lacks the ability to model intra-behavior causality, resulting in recommendations that are thematically relevant but misaligned with the user's actual browsing trajectory. In contrast, HCR successfully recovers two of the ground-truth browsed items, i.e., an entertainment article and a political news article, by modeling the user's diverse latent interests and causal dependencies among past behaviors. For instance, a multi-hop logical interaction path such as $1 \rightarrow 5 \rightarrow 6 \rightarrow 10 \rightarrow 3$ reflects a gradual behavioral transition from legal drama to emotional retrospection and finally to light-hearted entertainment, enabling HCR to accurately infer the user's final preference for entertainment content (e.g., Jeopardy-related news).

5.5 Summary

In this study, we introduced a hierarchical causal modeling framework for personalized news recommendation, which jointly captures the generative influence of latent user interests and structural dependencies in historical behaviors. By disentangling user interests and discovering causal relationships through a differentiable framework, our method offers interpretable and robust user representations. Experimental results on real-world datasets confirm that our approach substantially improves recommendation performance across accuracy and diversity.

Looking forward, this work presents several limitations that warrant further investigation, along with potential directions for improvement. A central challenge lies in data sparsity: for users with limited interaction histories, particularly in cold-start scenarios, it becomes difficult to infer reliable user-specific causal structures. In such cases, the model may degenerate into capturing mere correlations rather than meaningful causal dependencies. To address this issue, a promising approach is to construct a global, cross-user causal graph that can serve as a structural prior, thereby facilitating more robust and data-efficient inference under sparse conditions.

Beyond addressing sparsity, there are several avenues for extending the proposed framework. First, incorporating multi-modal news features (e.g., images, videos, and metadata) may enrich content representations and improve recommendation quality. Second, fine-grained modeling of temporal dynamics at the user level could better capture evolving interests and behavioral shifts. Third, integrating counterfactual reasoning may allow the system to simulate user interaction under hypothetical scenarios, further enhancing robustness and interpretability in dynamic environments.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

In this thesis, we focus on advancing causal representation learning for sequential user behaviors, with an emphasis on bridging the gap between generic sequential recommendation and the specific requirements of news recommendation. We identify several key challenges in existing methods, particularly their reliance on statistical associations, lack of robustness to hidden biases, and limited capacity to capture the hierarchical causal mechanisms underlying user decision-making.

To address these issues, we propose three novel approaches. First, we propose a causality-driven sequential user modeling framework, which constructs causal graphs to identify and reduce the influence of unobserved confounders in user interaction histories. Second, we develop the Causal Behavior Pattern Inference (CBPI) framework, which learns personalized causal structures directly from user behavior data, enabling flexible adaptation to the evolving nature of user interests and rapidly changing content, as seen in news recommendation scenarios. Third, we present a hierarchical causal modeling framework that simultaneously models the generative influence of latent user interests and the structural dependencies among observed behaviors, resulting in more interpretable and robust recommendations.

We believe our work takes a significant step toward building recommendation systems that are not only accurate and robust, but also transparent and aligned more closely with the true causal processes underlying user decisions. By systematically addressing the fundamental challenges in causal representation learning for sequential user behaviors, this thesis lays a foundation for the next generation of explainable and adaptive personalized recommendation systems.

6.2 Future Work

This thesis has systematically addressed several key challenges in causal representation learning for sequential user behaviors. However, achieving truly human-like causal reasoning in user decision-making remains an ambitious goal. Looking forward, several research directions can further deepen and expand the impact of causal representation learning across sequential recommendation scenarios.

First, integrating multi-modal news content, such as images, videos, and audio, into causal representation frameworks is a promising avenue. Multi-modal signals can provide a richer context for both user behaviors and item attributes, enabling more comprehensive modeling of the causal factors that drive user decisions in dynamic environments.

Second, capturing user-level temporal dynamics with greater granularity remains an important challenge. While current models often assume fixed temporal scales, user interests and the causal patterns that shape them can evolve rapidly or over extended periods. Developing adaptive and temporally-aware causal representation learning methods will be crucial for better understanding and predicting sequential user behaviors in real-world applications.

Third, the incorporation of counterfactual reasoning is a compelling future direction.

By simulating user behavior under hypothetical interventions, such as exposure to new content or changes in recommendation strategies, counterfactual approaches can more effectively disentangle causal effects from spurious correlations, leading to fairer, more robust, and more actionable user representations.

Finally, the rapid advancement of large language models (LLMs) opens up exciting opportunities for causal representation learning in sequential settings. LLMs' powerful language understanding and generation capabilities can be leveraged to enhance both user intent modeling and content enrichment. Future research could explore how LLMs can be integrated with causal frameworks to jointly reason about user interests, behavior sequences, and content semantics, ultimately leading to recommendation systems that are not only accurate and robust, but also interpretable and user-centric.

Overall, these directions reinforce the central theme of this thesis: advancing causal representation learning for sequential user behavior, from generic frameworks to specialized applications like news recommendation, in pursuit of more intelligent, transparent, and adaptive recommender systems.

References

[1] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. The connection between popularity bias, calibration, and fairness in recommendation. In *Proceedings of the 14th ACM Conference on Recommender Systems*, pages 726–731, 2020.

[2] Mingxiao An, Fangzhao Wu, Chuhan Wu, Kun Zhang, Zheng Liu, and Xing Xie. Neural news recommendation with long-and short-term user representations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 336–345, 2019.

[3] Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

[4] Stephen Bonner and Flavian Vasile. Causal embeddings for recommendation. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 104–112, 2018.

[5] Peter Bühlmann. Invariance, causality and robustness. *Statistical Science*, 35(3):404–426, 2020.

[6] Allison JB Chaney, Brandon M Stewart, and Barbara E Engelhardt. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 224–232, 2018.

[7] Jianxin Chang, Chen Gao, Yu Zheng, Yiqun Hui, Yanan Niu, Yang Song, Depeng Jin, and Yong Li. Sequential recommendation with graph neural networks. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 378–387, 2021.

[8] Chen Chen, Jie Guo, and Bin Song. Dual attention transfer in session-based recommendation with multi-dimensional integration. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 869–878, 2021.

[9] Tianwen Chen and Raymond Chi-Wing 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*, pages 1172–1180, 2020.

[10] Yewang Chen, Weiyao Ye, Chen Lin, and Yi Chen. Unbiased news recommendation model combining time and content. *Expert Systems with Applications*, 256:124864, 2024.

[11] Konstantina Christakopoulou, Madeleine Traverse, Trevor Potter, Emma Marriott, Daniel Li, Chris Haulk, Ed H Chi, and Minmin Chen. Deconfounding user satisfaction estimation from response rate bias. In *Proceedings of the 14th ACM Conference on Recommender Systems*, pages 450–455, 2020.

[12] Abhisek Dash, Abhijnan Chakraborty, Saptarshi Ghosh, Animesh Mukherjee, and Krishna P Gummadi. Fairir: mitigating exposure bias from related item recommendations in two-sided platforms. *IEEE Transactions on Computational Social Systems*, 10(3):1301–1313, 2022.

[13] James Davidson, Benjamin Liebald, Junning Liu, Palash Nandy, Taylor Van Vleet, Ullas Gargi, Sujoy Gupta, Yu He, Mike Lambert, Blake Livingston,

et al. The youtube video recommendation system. In *Proceedings of the fourth ACM Conference on Recommender Systems*, pages 293–296, 2010.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, 2019.

[15] Cyrus DiCiccio, Sriram Vasudevan, Kinjal Basu, Krishnaram Kenthapadi, and Deepak Agarwal. Evaluating fairness using permutation tests. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1467–1477, 2020.

[16] Christian Fong and Justin Grimmer. Discovery of treatments from text corpora. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1600–1609, 2016.

[17] Michele Jonsson Funk, Daniel Westreich, Chris Wiesen, Til Stürmer, M Alan Brookhart, and Marie Davidian. Doubly robust estimation of causal effects. *American Journal of Epidemiology*, 173(7):761–767, 2011.

[18] Chen Gao, Yu Zheng, Wenjie Wang, Fuli Feng, Xiangnan He, and Yong Li. Causal inference in recommender systems: A survey and future directions. *ACM Transactions on Information Systems*, 42(4):1–32, 2024.

[19] Shen Gao, Jiabao Fang, Quan Tu, Zhitao Yao, Zhumin Chen, Pengjie Ren, and Zhaochun Ren. Generative news recommendation. In *Proceedings of the ACM Web Conference 2024*, pages 3444–3453, 2024.

[20] Yingqiang Ge, Shuya Zhao, Honglu Zhou, Changhua Pei, Fei Sun, Wenwu Ou, and Yongfeng Zhang. Understanding echo chambers in e-commerce recommender systems. In *Proceedings of the 43rd International ACM SIGIR Confer-*

ence on Research and Development in Information Retrieval, pages 2261–2270, 2020.

[21] Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. Deepfm: a factorization-machine based neural network for ctr prediction. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 1725–1731, 2017.

[22] Lei Guo, Jinyu Zhang, Li Tang, Tong Chen, Lei Zhu, and Hongzhi Yin. Time interval-enhanced graph neural network for shared-account cross-domain sequential recommendation. *IEEE Transactions on Neural Networks and Learning Systems*, 35(3):4002–4016, 2022.

[23] Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. Deep iv: A flexible approach for counterfactual prediction. In *International Conference on Machine Learning*, pages 1414–1423. PMLR, 2017.

[24] Ruining He, Chen Fang, Zhaowen Wang, and Julian McAuley. Vista: A visually, socially, and temporally-aware model for artistic recommendation. In *Proceedings of the 10th ACM Conference on Recommender Systems*, pages 309–316, 2016.

[25] Ruining He, Wang-Cheng Kang, and Julian McAuley. Translation-based recommendation. In *Proceedings of the 11th ACM Conference on Recommender Systems*, pages 161–169, 2017.

[26] Ruining He, Wang-Cheng Kang, Julian J McAuley, et al. Translation-based recommendation: A scalable method for modeling sequential behavior. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 5264–5268, 2018.

[27] Ruining He and Julian McAuley. Fusing similarity models with markov chains for sparse sequential recommendation. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 191–200. IEEE, 2016.

[28] Xiangnan He, Yang Zhang, Fuli Feng, Chonggang Song, Lingling Yi, Guohui Ling, and Yongdong Zhang. Addressing confounding feature issue for causal recommendation. *ACM Transactions on Information Systems*, 41(3):1–23, 2023.

[29] Yue He, Zimu Wang, Peng Cui, Hao Zou, Yafeng Zhang, Qiang Cui, and Yong Jiang. Causpref: Causal preference learning for out-of-distribution recommendation. In *Proceedings of the ACM Web Conference 2022*, pages 410–421, 2022.

[30] Lucien Heitz, Juliane A Lischka, Alena Birrer, Bibek Paudel, Suzanne Tolmeijer, Laura Laugwitz, and Abraham Bernstein. Benefits of diverse news recommendations for democracy: A user study. *Digital Journalism*, 10(10):1710–1730, 2022.

[31] Balázs Hidasi and Alexandros Karatzoglou. Recurrent neural networks with top-k gains for session-based recommendations. In *Proceedings of the 27th ACM international conference on information and knowledge management*, pages 843–852, 2018.

[32] Keisuke Hirano, Guido W Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

[33] Haoji Hu, Xiangnan He, Jinyang Gao, and Zhi-Li Zhang. Modeling personalized item frequency information for next-basket recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1071–1080, 2020.

[34] Jin Huang, Wayne Xin Zhao, Hongjian Dou, Ji-Rong Wen, and Edward Y Chang. Improving sequential recommendation with knowledge-enhanced mem-

ory networks. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 505–514, 2018.

[35] Zhirong Huang, Shichao Zhang, Debo Cheng, Jiuyong Li, Lin Liu, and Guangquan Lu. Data-driven conditional instrumental variables for debiasing recommender systems. *arXiv preprint arXiv:2408.09651*, 2024.

[36] Hao Jiang, Chuanzhen Li, Juanjuan Cai, and Jingling Wang. Rcenr: A reinforced and contrastive heterogeneous network reasoning model for explainable news recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1710–1720, 2023.

[37] Santosh Kabbur, Xia Ning, and George Karypis. Fism: factored item similarity models for top-n recommender systems. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 659–667, 2013.

[38] Wang-Cheng Kang and Julian McAuley. Self-attentive sequential recommendation. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 197–206. IEEE, 2018.

[39] Jan Kmenta. Mostly harmless econometrics: An empiricist's companion, 2010.

[40] Hyeyoung Ko, Suyeon Lee, Yoonseo Park, and Anna Choi. A survey of recommendation systems: recommendation models, techniques, and application fields. *Electronics*, 11(1):141, 2022.

[41] Trent Kyono, Yao Zhang, and Mihaela van der Schaar. Castle: Regularization via auxiliary causal graph discovery. *Advances in Neural Information Processing Systems*, 33:1501–1512, 2020.

[42] Greg Lewis and Vasilis Syrgkanis. Adversarial generalized method of moments. *arXiv preprint arXiv:1803.07164*, 2018.

[43] Jian Li, Jieming Zhu, Qiwei Bi, Guohao Cai, Lifeng Shang, Zhenhua Dong, Xin Jiang, and Qun Liu. Miner: Multi-interest matching network for news recommendation. In *Findings of the Association for Computational Linguistics*, pages 343–352, 2022.

[44] Jiaxuan Liang, Jun Wang, Guoxian Yu, Shuyin Xia, and Guoyin Wang. Multi-granularity causal structure learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 13727–13735, 2024.

[45] Jiahui Liu, Peter Dolan, and Elin Rønby Pedersen. Personalized news recommendation based on click behavior. In *Proceedings of the 15th international conference on Intelligent user interfaces*, pages 31–40, 2010.

[46] Qiao Liu, Yifu Zeng, Refuoe Mokhosi, and Haibin 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*, pages 1831–1839, 2018.

[47] Xiaoyu Liu, Jiaxin Yuan, Yuhang Zhou, Jingling Li, Furong Huang, and Wei Ai. Csrec: Rethinking sequential recommendation from a causal perspective. *arXiv preprint arXiv:2409.05872*, 2024.

[48] Sindy Löwe, David Madras, Richard Zemel, and Max Welling. Amortized causal discovery: Learning to infer causal graphs from time-series data. In *Conference on Causal Learning and Reasoning*, pages 509–525, 2022.

[49] Wenpeng Lu, Rongyao Wang, Shoujin Wang, Xueping Peng, Hao Wu, and Qian Zhang. Aspect-driven user preference and news representation learning for news recommendation. *IEEE Transactions on Intelligent Transportation Systems*, 23(12):25297–25307, 2022.

[50] Yusheng Lu and Yongrui Duan. Online content-based sequential recommendation considering multimodal contrastive representation and dynamic preferences. *Neural Computing and Applications*, 36(13):7085–7103, 2024.

[51] Zhiming Mao, Xingshan Zeng, and Kam-Fai Wong. Neural news recommendation with collaborative news encoding and structural user encoding. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 46–55, 2021.

[52] Rishabh Mehrotra, Prasanta Bhattacharya, and Mounia Lalmas. Inferring the causal impact of new track releases on music recommendation platforms through counterfactual predictions. In *Proceedings of the 14th ACM Conference on Recommender Systems*, pages 687–691, 2020.

[53] Stephen L Morgan and Christopher Winship. *Counterfactuals and causal inference*. Cambridge University Press, 2015.

[54] Krikamol Muandet, Arash Mehrjou, Si Kai Lee, and Anant Raj. Dual instrumental variable regression. *Advances in Neural Information Processing Systems*, 33:2710–2721, 2020.

[55] Guoshun Nan, Rui Qiao, Yao Xiao, Jun Liu, Sicong Leng, Hao Zhang, and Wei Lu. Interventional video grounding with dual contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2765–2775, 2021.

[56] Shumpei Okura, Yukihiro Tagami, Shingo Ono, and Akira Tajima. Embedding-based news recommendation for millions of users. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1933–1942, 2017.

[57] Judea Pearl. *Causality*. Cambridge university press, 2009.

[58] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018.

[59] Ronan Perry, Julius Von Kügelgen, and Bernhard Schölkopf. Causal discovery in heterogeneous environments under the sparse mechanism shift hypothesis. *Advances in Neural Information Processing Systems*, 35:10904–10917, 2022.

[60] Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *The Journal of Machine Learning Research*, 15(1):2009–2053, 2014.

[61] Owen Phelan, Kevin McCarthy, Barry Smyth, Mike Bennett, P Clough, C Foley, C Gurrin, GJF Jones, W Kraaij, H Lee, et al. Terms of a feather: content-based news discovery and recommendation using twitter [j]. *Information Retrieval (ECIR-11)*, 18:21, 2010.

[62] Evaggelia Pitoura, Kostas Stefanidis, and Georgia Koutrika. Fairness in rankings and recommendations: an overview. *The VLDB Journal*, pages 1–28, 2022.

[63] Ruihong Qiu, Jingjing Li, Zi Huang, and Hongzhi Yin. Rethinking the item order in session-based recommendation with graph neural networks. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 579–588, 2019.

[64] Massimo Quadrana, Alexandros Karatzoglou, Balázs Hidasi, and Paolo Cremonesi. Personalizing session-based recommendations with hierarchical recurrent neural networks. In *proceedings of the Eleventh ACM Conference on Recommender Systems*, pages 130–137, 2017.

[65] Steffen Rendle. Factorization machines with libfm. *ACM Transactions on Intelligent Systems and Technology*, 3(3):1–22, 2012.

[66] Margaret E Roberts, Brandon M Stewart, and Richard A Nielsen. Adjusting for confounding with text matching. *American Journal of Political Science*, 64(4):887–903, 2020.

[67] Shuqi Ruan, Chao Yang, and Dongsheng Li. Knowledge-enhanced personalized hierarchical attention network for sequential recommendation. *World Wide Web*, 27(1):2, 2024.

[68] Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American statistical Association*, 100(469):322–331, 2005.

[69] Joseph F Rychlak. Logical learning theory: A teleological alternative in the field of personality. *Journal of personality*, 54(4):734–762, 1986.

[70] Joseph F Rychlak. *Logical learning theory: A human teleology and its empirical support*. U of Nebraska Press, 1994.

[71] Yuta Saito. Asymmetric tri-training for debiasing missing-not-at-random explicit feedback. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 309–318, 2020.

[72] Yuta Saito, Suguru Yaginuma, Yuta Nishino, Hayato Sakata, and Kazuhide Nakata. Unbiased recommender learning from missing-not-at-random implicit feedback. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 501–509, 2020.

[73] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*, pages 285–295, 2001.

[74] Masahiro Sato, Sho Takemori, Janmajay Singh, and Tomoko Ohkuma. Unbiased learning for the causal effect of recommendation. In *Proceedings of the 14th ACM Conference on Recommender Systems*, pages 378–387, 2020.

[75] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. Recommendations as treatments: Debiasing learning and evaluation. In *International Conference on Machine Learning*, pages 1670–1679. PMLR, 2016.

[76] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[77] Jiayi Shen, Xiantong Zhen, Marcel Worring, and Ling Shao. Variational multi-task learning with gumbel-softmax priors. *Advances in Neural Information Processing Systems*, 34:21031–21042, 2021.

[78] Zihua Si, Xueran Han, Xiao Zhang, Jun Xu, Yue Yin, Yang Song, and Ji-Rong Wen. A model-agnostic causal learning framework for recommendation using search data. In *Proceedings of the ACM Web Conference 2022*, pages 224–233, 2022.

[79] Weiping Song, Zhiping Xiao, Yifan Wang, Laurent Charlin, Ming Zhang, and Jian Tang. Session-based social recommendation via dynamic graph attention networks. In *Proceedings of the Twelfth ACM international conference on web search and data mining*, pages 555–563, 2019.

[80] Harald Steck. Calibrated recommendations. In *Proceedings of the 12th ACM conference on recommender systems*, pages 154–162, 2018.

[81] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1441–1450, 2019.

[82] Xiangyu Sun, Oliver Schulte, Guiliang Liu, and Pascal Poupart. Nts-notears: Learning nonparametric dbns with prior knowledge. In *International Conference on Artificial Intelligence and Statistics*, pages 1942–1964, 2023.

[83] Jiakai Tang, Shiqi Shen, Zhipeng Wang, Zhi Gong, Jingsen Zhang, and Xu Chen. When fairness meets bias: a debiased framework for fairness aware top-n recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 200–210, 2023.

[84] Jiaxi Tang and Ke Wang. Personalized top-n sequential recommendation via convolutional sequence embedding. In *Proceedings of the 11th ACM International Conference on Web Search and Data Mining*, pages 565–573, 2018.

[85] Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *International conference on machine learning*, pages 2139–2148. PMLR, 2016.

[86] Celina Treuillier, Sylvain Castagnos, Evan Dufraisse, and Armelle Brun. Being diverse is not enough: Rethinking diversity evaluation to meet challenges of news recommender systems. In *Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, pages 222–233, 2022.

[87] Victor Veitch, Dhanya Sridhar, and David Blei. Adapting text embeddings for causal inference. In *Conference on Uncertainty in Artificial Intelligence*, pages 919–928. PMLR, 2020.

[88] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

[89] Hongwei Wang, Fuzheng Zhang, Xing Xie, and Minyi Guo. Dkn: Deep knowledge-aware network for news recommendation. In *Proceedings of the 2018 World Wide Web Conference*, pages 1835–1844, 2018.

[90] Rongyao Wang, Shoujin Wang, Wenpeng Lu, and Xueping Peng. News recommendation via multi-interest news sequence modelling. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7942–7946. IEEE, 2022.

[91] Rongyao Wang, Shoujin Wang, Wenpeng Lu, Xueping Peng, Weiyu Zhang, Chaoqun Zheng, and Xinxiao Qiao. Intention-aware user modeling for personalized news recommendation. In *International Conference on Database Systems for Advanced Applications*, pages 179–194, 2023.

[92] Shicheng Wang, Shu Guo, Lihong Wang, Tingwen Liu, and Hongbo Xu. Hdnr: A hyperbolic-based debiased approach for personalized news recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 259–268, 2023.

[93] Shoujin Wang, Longbing Cao, Yan Wang, Quan Z Sheng, Mehmet A Orgun, and Defu Lian. A survey on session-based recommender systems. *ACM Computing Surveys (CSUR)*, 54(7):1–38, 2021.

[94] Shoujin Wang, Liang Hu, Yan Wang, Longbing Cao, Quan Z Sheng, and Mehmet Orgun. Sequential recommender systems: challenges, progress and prospects. *arXiv preprint arXiv:2001.04830*, 2019.

[95] Shoujin Wang, Liang Hu, Yan Wang, Quan Z Sheng, Mehmet Orgun, and Longbing Cao. Modeling multi-purpose sessions for next-item recommendations via mixture-channel purpose routing networks. In *International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence, 2019.

[96] Wenjie Wang, Fuli Feng, Xiangnan He, Xiang Wang, and Tat-Seng Chua. Deconfounded recommendation for alleviating bias amplification. In *Proceedings*

of the 27th ACM SIGKDD conference on knowledge discovery & data mining, pages 1717–1725, 2021.

[97] Wenjie Wang, Fuli Feng, Liqiang Nie, and Tat-Seng Chua. User-controllable recommendation against filter bubbles. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1251–1261, 2022.

[98] Wenjie Wang, Xinyu Lin, Fuli Feng, Xiangnan He, Min Lin, and Tat-Seng Chua. Causal representation learning for out-of-distribution recommendation. In *Proceedings of the ACM Web Conference 2022*, pages 3562–3571, 2022.

[99] Wenjie Wang, Yang Zhang, Haoxuan Li, Peng Wu, Fuli Feng, and Xiangnan He. Causal recommendation: Progresses and future directions. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3432–3435, 2023.

[100] Xiang Wang, Hongye Jin, An Zhang, Xiangnan He, Tong Xu, and Tat-Seng Chua. Disentangled graph collaborative filtering. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1001–1010, 2020.

[101] Xiaojie Wang, Rui Zhang, Yu Sun, and Jianzhong Qi. Doubly robust joint learning for recommendation on data missing not at random. In *International Conference on Machine Learning*, pages 6638–6647. PMLR, 2019.

[102] Xiaojie Wang, Rui Zhang, Yu Sun, and Jianzhong Qi. Combating selection biases in recommender systems with a few unbiased ratings. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 427–435, 2021.

[103] Yifan Wang, Weizhi Ma, Min Zhang, Yiqun Liu, and Shaoping Ma. A survey on the fairness of recommender systems. *ACM Transactions on Information Systems*, 41(3):1–43, 2023.

[104] Yifan Wang, Suyao Tang, Yuntong Lei, Weiping Song, Sheng Wang, and Ming Zhang. Disenhan: Disentangled heterogeneous graph attention network for recommendation. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 1605–1614, 2020.

[105] Zhenlei Wang, Xu Chen, Rui Zhou, Quanyu Dai, Zhenhua Dong, and Ji-Rong Wen. Sequential recommendation with user causal behavior discovery. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, pages 28–40. IEEE, 2023.

[106] Zhenlei Wang, Jingsen Zhang, Hongteng Xu, Xu Chen, Yongfeng Zhang, Wayne Xin Zhao, and Ji-Rong Wen. Counterfactual data-augmented sequential recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 347–356, 2021.

[107] Zimu Wang, Yue He, Jiashuo Liu, Wenchao Zou, Philip S Yu, and Peng Cui. Invariant preference learning for general debiasing in recommendation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1969–1978, 2022.

[108] Ziyang Wang, Huoyu Liu, Wei Wei, Yue Hu, Xian-Ling Mao, Shaojian He, Rui Fang, and Dangyang Chen. Multi-level contrastive learning framework for sequential recommendation. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 2098–2107, 2022.

[109] Ziyang Wang, Wei Wei, Gao Cong, Xiao-Li Li, Xian-Ling Mao, and Minghui Qiu. Global context enhanced graph neural networks for session-based recom-

mendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 169–178, 2020.

[110] Peiyang Wei, Hongping Shu, Jianhong Gan, Xun Deng, Yi Liu, Wenying Sun, Tinghui Chen, Can Hu, Zhenzhen Hu, Yonghong Deng, et al. Sequential recommendation system based on deep learning: A survey. *Electronics*, 2025.

[111] Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. Npa: neural news recommendation with personalized attention. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining.*

[112] Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. Neural news recommendation with attentive multi-view learning. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 3863–3869, 2019.

[113] Chuhan Wu, Fangzhao Wu, Suyu Ge, Tao Qi, Yongfeng Huang, and Xing Xie. Neural news recommendation with multi-head self-attention. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 6389–6394, 2019.

[114] Chuhan Wu, Fangzhao Wu, Yongfeng Huang, and Xing Xie. Personalized news recommendation: Methods and challenges. *ACM Transactions on Information Systems*, 41(1):1–50, 2023.

[115] Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. Sentirec: Sentiment diversity-aware neural news recommendation. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 44–53, 2020.

[116] Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. Empowering news recommendation with pre-trained language models. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1652–1656, 2021.

[117] Chuhan Wu, Fangzhao Wu, Tao Qi, Chenliang Li, and Yongfeng Huang. Is news recommendation a sequential recommendation task? In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2382–2386, 2022.

[118] Chuhan Wu, Fangzhao Wu, Xiting Wang, Yongfeng Huang, and Xing Xie. Fairness-aware news recommendation with decomposed adversarial learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 4462–4469, 2021.

[119] Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, et al. Mind: A large-scale dataset for news recommendation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3597–3606, 2020.

[120] Huizi Wu, Cong Geng, and Hui Fang. Causality and correlation graph modeling for effective and explainable session-based recommendation. *ACM Transactions on the Web*, 18(1):1–25, 2023.

[121] Shu Wu, Yuyuan Tang, Yanqiao Zhu, Liang Wang, Xing Xie, and Tieniu Tan. Session-based recommendation with graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 346–353, 2019.

[122] Teng Xiao and Suhang Wang. Towards unbiased and robust causal ranking for recommender systems. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 1158–1167, 2022.

[123] Xu Xie, Fei Sun, Zhaoyang Liu, Shiwen Wu, Jinyang Gao, Jiandong Zhang, Bolin Ding, and Bin Cui. Contrastive learning for sequential recommendation. In *2022 IEEE 38th international Conference on Data Engineering (ICDE)*, pages 1259–1273. IEEE, 2022.

[124] Hao Xu, Bo Yang, Xiangkun Liu, Wenqi Fan, and Qing Li. Category-aware multi-relation heterogeneous graph neural networks for session-based recommendation. *Knowledge-Based Systems*, 251:109246, 2022.

[125] Shuyuan Xu, Jianchao Ji, Yunqi Li, Yingqiang Ge, Juntao Tan, and Yongfeng Zhang. Causal inference for recommendation: Foundations, methods, and applications. *ACM Transactions on Intelligent Systems and Technology*, 16(3):1–51, 2025.

[126] Boming Yang, Dairui Liu, Toyotaro Suzumura, Ruihai Dong, and Irene Li. Going beyond local: Global graph-enhanced personalized news recommendations. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 24–34, 2023.

[127] Zhen Yang, Wenhui Wang, Tao Qi, Peng Zhang, Tianyun Zhang, Ru Zhang, Jianyi Liu, and Yongfeng Huang. Glocim: Global-view long chain interest modeling for news recommendation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6855–6865, 2025.

[128] Zuoxi Yang, Shoubin Dong, and Jinlong Hu. Gfe: General knowledge enhanced framework for explainable sequential recommendation. *Knowledge-Based Systems*, 230:107375, 2021.

[129] Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(5):1–46, 2021.

[130] Haochao Ying, Fuzhen Zhuang, Fuzheng Zhang, Yanchi Liu, Guandong Xu, Xing Xie, Hui Xiong, and Jian Wu. Sequential recommender system based on hierarchical attention network. In *IJCAI International Joint Conference on Artificial Intelligence*, 2018.

[131] Yue Yu, Jie Chen, Tian Gao, and Mo Yu. DAG-GNN: Dag structure learning with graph neural networks. In *International Conference on Machine Learning*, pages 7154–7163, 2019.

[132] Fajie Yuan, Alexandros Karatzoglou, Ioannis Arapakis, Joemon M Jose, and Xiangnan He. A simple convolutional generative network for next item recommendation. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 582–590, 2019.

[133] Junkun Yuan, Anpeng Wu, Kun Kuang, Bo Li, Runze Wu, Fei Wu, and Lanfen Lin. Auto iv: Counterfactual prediction via automatic instrumental variable decomposition. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(4):1–20, 2022.

[134] He Zeyu, Lu Yan, Feng Wendi, Zhang Wei, Fayadh Alenezi, and Prayag Tiwari. Causal embedding of user interest and conformity for long-tail session-based recommendations. *Information Sciences*, 644:119167, 2023.

[135] G Peter Zhang. Time series forecasting using a hybrid arima and neural network model. *Neurocomputing*, 50:159–175, 2003.

[136] Hui Zhang, Xu Chen, and Shuai Ma. Dynamic news recommendation with hierarchical attention network. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 1456–1461. IEEE, 2019.

[137] Lemei Zhang, Peng Liu, and Jon Atle Gulla. Dynamic attention-integrated neural network for session-based news recommendation. *Machine Learning*, 108:1851–1875, 2019.

[138] Shengyu Zhang, Dong Yao, Zhou Zhao, Tat-Seng Chua, and Fei 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*, pages 367–377, 2021.

[139] Wenhao Zhang, Wentian Bao, Xiao-Yang Liu, Keping Yang, Quan Lin, Hong Wen, and Ramin Ramezani. Large-scale causal approaches to debiasing post-click conversion rate estimation with multi-task learning. In *Proceedings of the Web Conference*, pages 2775–2781, 2020.

[140] Xuanyu Zhang, Qing Yang, and Dongliang Xu. Deepvt: Deep view-temporal interaction network for news recommendation. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 2640–2650, 2022.

[141] Yang Zhang, Fuli Feng, Xiangnan He, Tianxin Wei, Chonggang Song, Guohui Ling, and Yongdong 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*, pages 11–20, 2021.

[142] Yihu Zhang, Bo Yang, Haodong Liu, and Dongsheng Li. A time-aware self-attention based neural network model for sequential recommendation. *Applied Soft Computing*, 133:109894, 2023.

[143] Yuxin Zhang, Gaode Chen, Lei Wang, Xiaobo Guo, and Liyue Ren. Mirec: Neural news recommendation with multi-interest and popularity-aware modeling. In *2023 IEEE 47th Annual Computers, Software, and Applications Conference*, pages 458–465, 2023.

[144] Xiaoyao Zheng, Xingwang Li, Zhenghua Chen, Liping Sun, Qingying Yu, Liangmin Guo, and Yonglong Luo. Enhanced self-attention mechanism for long and

short term sequential recommendation models. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2024.

[145] Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31, 2018.

[146] Yu Zheng, Chen Gao, Xiangnan He, Depeng Jin, and Yong Li. Incorporating price into recommendation with graph convolutional networks. *IEEE Transactions on Knowledge and Data Engineering*, 35(2):1609–1623, 2021.

[147] Tianyu Zhu, Leilei Sun, and Guoqing Chen. Graph-based embedding smoothing for sequential recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[148] Tranos Zuva, Sunday O Ojo, Seleman M Ngwira, Keneilwe Zuva, et al. A survey of recommender systems techniques, challenges and evaluation metrics. *International Journal of Emerging Technology and Advanced Engineering*, 2(11):382–386, 2012.