CausalRec: A CausalBoost Attention Model for Sequential Recommendation

Yunbo Hou[*]
School of Software and
Microelectronics
Peking University
Beijing, China
yunboh@stu.pku.edu.cn

Tianle Yang[*]
Alibaba Group
Beijing, China
yangtianle.ytl@alibaba-inc.com

Ruijie Li
School of Software and
Microelectronics
Peking University
Beijing, China
howtolove17@stu.pku.edu.cn

Li He
Alibaba Group
Beijing, China
heli@taobao.com

Liang Wang
Alibaba Group
Beijing, China
wangliang@taobao.com

Weiping Li
School of Software and
Microelectronics
Peking University
Beijing, China
wpli@ss.pku.edu.cn

Bo Zheng
Alibaba Group
Beijing, China
bozheng@alibaba-inc.com

Guojie Song[†]
National Key Laboratory of General
Artificial Intelligence
School of Intelligence Science and
Technology
Peking University
Beijing, China
gjsong@pku.edu.cn

Abstract

Recent advances in correlation–based sequential recommendation systems have demonstrated substantial success. Specifically, the attention-based model outperforms other RNN-based and Markov chains-based models by capturing both short- and long-term dependencies more effectively. However, solely focusing on item co-occurrences overlooks the underlying motivations behind user behaviors, leading to spurious correlations and potentially inaccurate recommendations. To address this limitation, we present a novel framework that integrates causal attention for sequential recommendation, CausalRec. It incorporates a causal discovery block and a CausalBooster. The causal discovery block learns the causal graph in user behavior sequences, and we provide a theory to guarantee the identifiability of the learned causal graph. The CausalBooster utilizes the discovered causal graph to refine the attention mechanism, prioritizing behaviors with causal significance. Experimental evaluations on real-world datasets indicate that CausalRec outperforms several state-of-the-art methods, with average improvements of 7.21% in Hit Rate (HR) and 8.65% in Normalized Discounted Cumulative Gain (NDCG). To the best of our knowledge, this is the first model to incorporate causality through the attention mechanism in sequential recommendation, demonstrating the value of causality in generating more accurate and reliable recommendations. Our code is available at https://anonymous.4open.science/r/CausalRec-202B/.

CCS Concepts

• **Information systems** → **Recommender systems**; • **Mathematics of computing** → **Causal networks**.

Keywords

Recommender Systems, Causal Discovery, Causal Network

[*]Equal contribution.
[†]Corresponding author.

1 Introduction

The goal of sequential recommendation is to uncover hidden patterns in user behavior. Current studies often incorporate correlation to identify these patterns, motivated by the natural intuition that sequential behaviors are typically related. Building on this intuition, various methods have been developed to model correlations

effectively. FPMC [19] models user behavior as Markov chains, assuming the current behavior is influenced only by the most recent one. GRU4Rec [7] employs recurrent neural networks (RNNs) to encapsulate prior actions within a hidden state. Another approach uses attention mechanisms [9, 25] to differentiate the importance of historical actions. As correlation has achieved great performance in

Figure 1: Illustrating the motivation for modeling causal relationships among user behaviors. Consider the example: although the cable and the phone shell frequently co-occur in the same user behavior sequence, it does not imply that they influence each other directly. Their co-occurrence arises from their shared causal relationship with the phone.

recommendations, these approaches typically learn *spurious correlations* [15, 29], which refers to a connection between two or more variables that appear to be causal but are not in fact. For example, in Fig. 1, correlation-based models might identify the correlation between the purchase of the phone shell and cable based on their co-occurrence. However, it doesn't imply that purchasing cables causes the purchase of phone shells. Instead, both purchases are driven by a shared causal factor: the purchase of phones. The example illustrates that by focusing on observed co-occurrence patterns, correlation-based models can be misled by such spurious correlations and find it hard to identify the real causal relationships, leading to wrong recommendations.

To address this limitation, we integrate causality into recommendations to capture the true causal relationships underlying user behavior. However, incorporating it into recommendation systems poses significant challenges. First, ensuring identifiability—whether the learned causal relationship accurately reflects the user behavior pattern—is inherently difficult in sequential recommendation data. Second, the typically large number of items in recommendation systems leads to exponential growth in the search space for the learned causal relationship. Third, effectively integrating mined causal relationships into recommendation models is complex. While prior work [28] employed clustering-based methods to learn stable causal relations and filter out irrelevant items, the absence of an identifiability guarantee can risk learning incorrect user behavior patterns. The clustering-based method alleviates the huge search space issue, but it leads the a loss in user interest, and whether the reconstructed causal relation based on the cluster indeed correctly reflects the real causal relation is questionable. Additionally, the filtering methods can inadvertently discard important items, leading to a decline in recommendation performance. Our experiment in

Section 5.3 has shown that our model with the filtering strategy shows a worse performance in some cases compared with SASRec, which is a correlation-based sequential recommendation model with self-attention mechanisms.

To address these challenges, we propose a CausalBoost attention model for sequential recommendation, CausalRec. We introduce a causal discovery block to infer causal relations (i.e., causal graphs) on user behavior sequences and a CausalBooster to incorporate the learned causality into the attention mechanism. We also provide theoretical guarantees for the identifiability of the learned causal graph, ensuring that it accurately captures user behavior patterns.

The causal discovery block plays a pivotal role in CausalRec. Constructing item-level causal graphs directly from real-world datasets is typically intractable because the search space grows exponentially with the number of items. To address this, we leverage the structure causal model (SCM) framework and estimate causal structures using the covariance matrix, thereby mitigating the large search space. As the attention mechanism can estimate the observed nodes of the SCM, we utilize the final layer representation of the transformer as samples to calculate covariance. By integrating layer normalization and attention mechanism in the transformer, we impose the equal variance and linear SCM assumption, ensuring our identifiability [30]. Therefore, we can employ continuous optimization with acyclicity and sparse constraints within the transformer framework to discern causal structures from input sequences. In the CausalBooster, we address the challenge of integrating causality into recommendation models. Unlike previous methods that rely on a filtering strategy and risk discarding critical historical items, our attention-based fusion enhances the attention weight according to the learned causal graph. This approach preserves valuable user interests, even in cases where the discovered causal graph contains minor inaccuracies, ensuring robust performance.

In summary, the main contributions of this paper are as follows:

(1) We propose CausalRec, a CausalBoost attention model for sequential recommendation that integrates causality into the recommendation. To the best of our knowledge, this is the first model that incorporates causality through the attention mechanism into the sequential recommendation.

(2) We introduce a Causal Discovery Block, which successfully introduces causality into the attention mechanism with identifiability guarantees, enabling attention to focus on causally relevant signals within user interaction sequences.

(3) We present the CausalBooster, an effective procedure to utilize the causal effect matrix, integrate causality into the attention mechanism by amplifying causal effects.

(4) We conduct extensive evaluations on four datasets. CausalRecachieves average improvements of approximately 4.71% in HR and 8.57% in NDCG, validating its effectiveness in improving recommendation performance.

2   Related Work

This work focuses on the intersection of sequential recommendation and causal discovery. In this section, we provide an overview of recent advances in these domains and illustrate the connections between our proposed framework and existing studies.

2.1 Sequential Recommendation

Recommendation systems have increasingly adopted models that account for the temporal and dynamic nature of user-item interactions. Traditional methods like matrix factorization [11] and static content modeling focus on aggregated user preferences, overlooking how these preferences evolve. In contrast, sequential recommendation methods take advantage of the user interaction history to predict future preferences. A common strategy is to employ encoder-like architectures that transform historical interaction sequences into latent representations, which are then used to predict the representation of the next item.

Early sequential models, such as Markov chains [19], treated user behavior as a stochastic process, relying on limited interaction windows. Extensions like Markov decision processes [21] incorporated longer-term decision-making but struggled with data sparsity and complex transition patterns.

Deep learning has significantly advanced this field. Recurrent neural networks (RNNs) [7] captured richer temporal dependencies, while gated variants like GRUs and LSTMs improved session-based recommendations. Attention mechanisms [9] and transformers [25] further enhanced performance by learning contextual embeddings without strict sequential processing, enabling them to handle complex, long-range dependencies.

Our work builds on these research directions by focusing on capturing item relationships. However, we extend them by directly discovering causal relations with identifiability guarantees, marking a significant advancement in sequential recommendation research.

2.2 Causal Discovery

Causal discovery seeks to uncover causal relationships among variables from observational or experimental data, forming a crucial foundation for understanding complex systems across various domains. Classical methods can be categorized into constraint-based, causal function-based, score-based, and gradient-based methods.

Constraint-based approaches rely on systematically testing for conditional independence (CI) among variables to infer causal edges under the assumption of causal sufficiency (i.e., no unmeasured confounders). A prominent example is the PC algorithm [3, 4, 24], which starts with a fully connected graph and iteratively removes edges based on statistical CI tests until all (conditional) independencies are satisfied. While these methods are theoretically sound, they are sensitive to the quality of independence tests and struggle with high-dimensional data.

Score-based methods cast causal discovery as an optimization problem, searching for a directed acyclic graph (DAG) that maximizes—or minimizes—a specific score function. Common scores include the Bayesian Information Criterion (BIC) [2]. While more robust to noise, these methods become computationally expensive as the number of variables increases.

Functional causal models (FCMs) assume each variable is generated by its direct causes through a (potentially nonlinear) function combined with an independent noise term, allowing causal direction to be inferred under certain assumptions. Representative methods include LiNGAM (Linear Non-Gaussian Acyclic Model) [22], which exploits non-Gaussianity to discover causal ordering in

linear settings, as well as additive noise models (ANM) [8] and post-nonlinear models (PNL) [31], which relax linear constraints. These approaches often rely on the independence between noise and cause for identification; however, if such assumptions are violated, performance may decrease. Recent advancements in machine learning have significantly advanced causal discovery. Deep learning models, such as variational autoencoders and generative adversarial networks, have been adapted to infer causal structures in high-dimensional and non-linear datasets [6, 10]. These models exploit the representational power of neural networks to identify causal relationships that traditional methods may overlook. Differentiable causal discovery frameworks, such as NOTEARS [12, 16, 33, 34], reformulate causal graph learning as a continuous optimization problem, enabling the use of gradient-based methods. These approaches have demonstrated scalability and effectiveness in handling large datasets, particularly when combined with sparsity constraints to regularize the learned causal graphs.

Our work builds on the NOTEARS framework, which formulates Bayesian structural learning as a continuous optimization problem. We incorporate this framework into recommendation systems, making an effective attempt to learn user behavior causality.

3 Preliminary

In this section, we first present the formal statement of the structural causal model and the attention mechanism, and then we establish the link between them.

3.1 Structural Causal Models

A Structural Causal Model (SCM) is a framework for representing causal relationships among a set of variables. Formally, an SCM consists of:

(1) A set of variables $x = \{x_1, x_2, \ldots, x_n\}$.
(2) A directed acyclic graph (DAG) $\mathcal{G}$, where nodes represent variables and edges denote functions.
(3) A set of functions f_i connecting the variables, where $x_i = f_i(\text{Pa}(x_i), u_i)$, and $\text{Pa}(x_i)$ represents parents of x_i in $\mathcal{G}$.
(4) u_i: An exogenous term for each variable, assumed to be mutually independent.

As a special case of SCMs, a linear SCM can be represented with:

$$x_i = \sum_{i \in \text{Pa}(j)} \beta_{ij} x_j + \lambda_{jj} u_j. \tag{1}$$

The linear SCM in Equation (1) has the following matrix form:

$$X = BX + \Lambda U, \tag{2}$$

where $X = (x_1, \ldots, x_n)^\top$, $U = (u_1, \ldots, u_n)^\top$: noise matrix, and $B \in \mathbb{R}^{n \times n}$ is an edge weight matrix, an autoregression matrix, or a weighted adjacency matrix with each element $[B]_{ij} = \beta_{ij}$, in which β_{ij} is the linear weight of an edge from x_i to x_j and Λ is a diagonal matrix with coefficient between x_i and u_i, $i = 1, \ldots, n$.

3.2 Attention Mechanism

The attention mechanism, central to Transformer models [27], computes a weighted representation of input elements based on their relevance to the query. It linearly transformed the input $Y \in \mathbb{R}^{n \times d}$ into three parts, *i.e.*, queries $Q = YW_Q \in \mathbb{R}^{n \times d_k}$, keys

$K = YW_K \in \mathbb{R}^{n \times d_k}$, and values $V = YW_V \in \mathbb{R}^{n \times d_v}$, where n denotes the sequence length and d, d_k, d_v represents the dimensions of inputs, queries(keys) and values. The scaled dot-product attention is applied on Q, K, V and can be formulated as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V, \qquad (3)$$

where the softmax function makes the rows of $\frac{QK^\top}{\sqrt{d_k}}$ sums to 1. The attention mechanism enables the model to focus on the most relevant parts of the input sequence, making it highly effective for sequence modeling tasks.

3.3 Link between Structural Causal Model and Attention Mechanism

The previous work [20] provides a bridge between self-attention and causal discovery. To be more specific, this section describes the details of why the attention mechanism can be considered to estimate the observed nodes of a SCM.

Initially, we explain that the covariance over the outputs of the attention has a similar formulation to the covariance over observed nodes in an SCM. For a linear SCM, $X = BX + \Lambda U$, then

$$X = (I - B)^{-1}\Lambda U \qquad (4)$$

where equation (4) represents a system with the outputs X, inputs U and weights $(I - B)^{-1}$. The covariance matrix of the output is

$$\begin{aligned} Cov(X) &= \mathbb{E}[(X - \mu_X)(X - \mu_X)^\top] \\ &= (I - B)^{-1}\Lambda Cov(U)((I - B)^{-1}\Lambda)^\top, \end{aligned} \qquad (5)$$

where $\mu_X = (I - B)^{-1}\Lambda\mu_U$, and $Cov(\cdot)$ means the covariance matrix.

On the other hand, an attention layer estimates the attention matrix A and a values matrix V from embeddings Y. The output embeddings are $Z = AV$ If we view V as a random variable with mean μ_V and covariance $Cov(V)$, then the output Z also has a covariance matrix given by
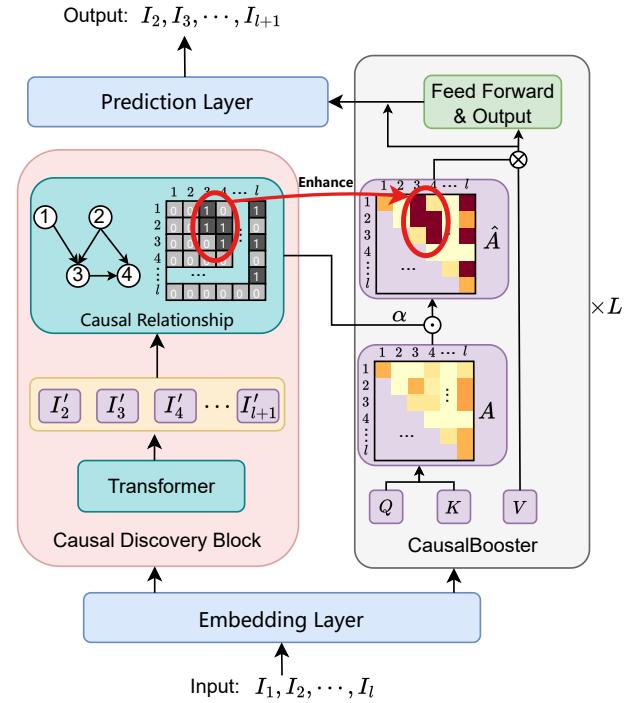
$$Cov(Z) = \mathbb{E}[(Z - \mu_Z)(Z - \mu_Z)^\top] = ACov(V)A^\top, \qquad (6)$$

where $\mu_Z = A\mu_V$. Comparing this to $Cov(X)$ in equation equation 5 shows a striking similarity: in each case, the output covariance is formed by applying a linear transformation (either $(I - B)^{-1}\Lambda$ in the SCM or A in the attention mechanism) to the input covariance ($Cov(U)$ or $Cov(V)$), then multiplying by its transpose.

Hence, the learned weight A in the attention mechanism plays the same linear mapping role as $(I - B)^{-1}\Lambda$ in SCM. In the SCM, the term Λ captures how exogenous inputs U influence the system, and $(I - B)^{-1}\Lambda$ encodes the dependence among observed variables. Analogously, in the attention mechanism, V represents the content being passed among tokens (similar to "inputs"), while A acts as a learned "adjacency matrix" that decides how different tokens interact with each other. Consequently, the attention mechanism can be viewed as estimating, in a data-driven way, relationships among observed nodes—mirroring how $(I - B)^{-1}\Lambda$ describes the relationships in a linear SCM.

4 Method

The overview of CausalRec is depicted in Fig. 2. The proposed model integrates a Causal Discovery Block to identify causal relations. To incorporate causality into the recommendation and further enhance sequential recommendation performance, we propose a CausalBooster that incorporates the discovered causal relations into the attention mechanism. Consistent with previous sequential recommendation methods, we include an Embedding Layer and a Prediction Layer to learn item embeddings and generate logits for the output (see details in Appendix 4.3). In summary, CausalRec comprises four main modules: the Embedding Layer, Causal Discovery Block, CausalBooster, and Prediction Layer.

Figure 2: Architecture of our proposed CausalRec. We propose a Causal Discovery Block to learn user behavior causality and incorporate it into our model with CausalBooster.

4.1 Causal Discovery Block

Identifying causal relations in recommendations faces identifiability concerns, which denote that multiple SCMs can produce the same observed data distribution, making it impossible to single out one true causal structure. Previous work, such as the causer [28], has attempted to solve the issues by clustering methods, but failed. To overcome these limitations, we propose an Item-level Causal Discovery Block and provide the identifiability proof.

4.1.1 Item-level Causal Discovery Block. Different from the causer, we directly mine the item's causal relationship. We leverage the SCM and estimate causal structures with a covariance matrix:

$$Cov(i,j) = \frac{1}{N}\sum_{k=1}^{N} x_{ki}x_{kj}^{\top} \in \mathbb{R}^{n \times n}, \qquad (7)$$

where $x_i \in \mathbb{R}^n$ is the vector of items in a user's sequence to avoid confusion with the identity matrix I symbol, n is the input sequence length, and N denotes the batch size. To address the identifiability issue, we use the final layer representation of the transformer as samples to calculate covariance. As illustrated in Section 3, the attention mechanism can estimate the linear SCM observed nodes, and the layer normalization mechanism naturally imposes an equal noise variance assumption across different dimensions. Since X follows linear SCM equations $X = BX + \Lambda U$, properties of the exogenous variables U—namely independence and finite variance—enable us to express $Cov(X)$ in terms of B and Λ. Therefore, B can be estimated using $Cov(X)$. Additionally, the incorporation of the transformer introduces the linear SCM and equal noise variance assumptions. These additional assumptions ensure that, with acyclicity and sparse constraints on $Cov(X)$ in continuous optimization, the DAG within the data generation process is theoretically identifiable [30]. This is due to the attention mechanism and the property of Layer Normalization, thereby resolving the identifiability issue in prior methods.

4.1.2 Identifiability Analysis. In this section, we provide the causal identifiability (DAG can be uniquely determined) of the Causal Discovery Block. As [20] provides a bridge between self-attention and causal discovery, and [27] demonstrates that the transformer can learn causal structure with gradient descent, here we show the identifiability of the outputs from the attention mechanism with the unique property of Layernorm. To the beginning, we provide the causal identifiability of linear SEM as Lemma 4.1: if the exogenous variables have equal variances, then the directed acyclic graph of this linear SEM can be uniquely identifiable. Here we denote $\mathcal{G}(B)$ as directed graph with adjacent weight matrix $(\beta_{j,k})B \in \mathbb{R}^{n \times n}$, vertex set $V = \{1,\ldots,n\}$ and edge $E(B)$ be the support of B $(E(B) = \{(k,j) : \beta_{j,k} \neq 0\})$. Let $X \sim (B_X, \sigma_X^2)$ denote the exogenous variables of linear SEM that have equal variance σ_X^2.

LEMMA 4.1 (IDENTIFIABILITY CONDITIONS FOR LINEAR SEMs [1, 17]). *Let $P(X), P(Y)$ be generated from a linear SEM (1) with DAG $\mathcal{G}(B_X), \mathcal{G}(B_Y)$ and true ordering $\pi_X, \pi_Y, X \sim (B_X, \sigma_X^2), Y \sim (B_Y, \sigma_Y^2)$ with both G_X, G_Y directed and acyclic. If $var(X) = var(Y)$, then $\mathcal{G}(B_X) = \mathcal{G}(B_Y)$, $B_X = B_Y$ and $\sigma_X^2 = \sigma_Y^2$. Then, DAG G is uniquely identifiable if exogenous variables have equal variances.*

A critical insight is that the self-attention mechanism, particularly in our context, can be formulated as a linear SEM, thus satisfying the preconditions of Lemma 4.1. Specifically, the Layer Normalization step ensures that the exogenous variables (derived from input embeddings) have identical variance. This allows us to formally state the identifiability of the learned causal structure.

PROPOSITION 4.2. *(Causal Identifiability of Self-Attention with LayerNorm) Let $P(Z)$ be generated from $Z = BZ + DV$ with pre-Layer normalization of diagonal matrix $D = diag(\frac{1}{\|V_1\|_2}, \ldots, \frac{1}{\|V_n\|_2})$ to make exogenous variables $\tilde{V} = DV$ with equal variances, $\mathcal{G}(Z)$ can be uniquely identified.*

PROOF. The output Z of the self-attention layer is a linear transformation of its input values V. For each output Z_i, we have: $Z_i = \sum_{j=1}^{n} \text{softmax}\left(\frac{Q_i K_j^T}{\sqrt{d_k}}\right) V_j = \sum_{j=1}^{n} A_{ij} V_j$, where A is the attention matrix. This confirms that the relationships among the components of Z are linear.

We can re-express this relationship with a linear SEM. For simplicity and without loss of generality, let's consider the case where the embedding dimension of V is 1. The structural model for the item representations Z can be written as $Z = BZ + \tilde{V}$, where the input value matrix V serves as the basis for the exogenous variables. The pre-Layer Normalization, represented by the diagonal matrix D, transforms V into $\tilde{V} = DV$. This normalization ensures that each component $\tilde{V}_i$ has unit variance, making them independent and identically distributed (i.i.d.) noise terms. By assuming that the causal graph $\mathcal{G}(B)$ is a DAG, the matrix $(I - B)$ is invertible. This allows for a unique solution for Z: $Z = (I - B)^{-1}\tilde{V}$ This equation describes a linear SEM where the output Z is generated from exogenous variables $\tilde{V}$ that have equal variances. According to Lemma 4.1, the underlying DAG $\mathcal{G}(B)$ is therefore uniquely identifiable from the covariance matrix of Z, which is given by $Cov(Z) = (I - B)^{-1}Cov(\tilde{V})((I - B)^{-1})^{\top}$. This completes the proof. □

In summary, the combination of self-attention (with causal masking ensuring the DAG property) and Layer Normalization satisfies the conditions for causal identifiability in linear SEMs. The autoregressive structure of the attention matrix A determines the causal ordering, and LayerNorm ensures the homogeneity of exogenous noise variances, thereby guaranteeing that the causal graph over the output representations Z is identifiable. Then, once we find a DAG during the training procedure, we say that the causal relationship matrix R in section 4.2 is uniquely identifiable.

To find a DAG, we introduce a general method to add acyclicity and sparse constraints in our optimization as NOTEARS [12]. One can also apply other acyclicity constraints which has less time complexity. The loss consists of two components:(1) the acyclicity constraint, which is to ensure that the learned causal graph is a directed acyclic graph. (2) Sparse Regularization, which is to encourage sparsity in causal relationships. The acyclicity constraint comes from the taylor expansion $trace(e^A) = trace(I) + trace(A) + trace(A^2) + \cdots$, where $A = W \odot W, W \in \mathbb{R}^{n \times n}$. Suppose W represents the adjacency matrix corresponding to graph $\mathcal{G}$ with n nodes and $(A^k)_{ij}$ denotes the number of k-step paths from node i to node j. Then $n = trace(I) = trace(A) + trace(A^2) + \cdots$ represents that there is no path from node i to node i, which indicates the graph $\mathcal{G}$ is a DAG. Therefore, the acyclicity constraint can be written as below:

$$\text{trace}(e^{W \odot W}) = n, \qquad (8)$$

where W denotes the covariance matrix $Cov(X)$, $\odot$ represents the Hadamard product, and n is the input sequence length. The idea of sparse regularization comes from the fact that in the user behavior sequence, the causal item corresponding to the target item should be sparse. Therefore, we incorporate the L_1 matrix norm on the covariance matrix as our sparse regularization to encourage sparse causal relations among items.

4.2 CausalBooster

The CausalBooster is designed to integrate causality into the sequential recommendation. It achieves this by stacking multiple CausalBoost Attention (CBA) Layers, each consisting of a CBA Layer and a feed-forward network. This structure allows the model to enhance relevant behaviors without discarding important information, addressing key challenges in incorporating causal relations.

4.2.1 CausalBoost Attention Layer.
Previous approaches [28] would filter out items with low causal relevance based on a learned causal matrix. However, that risks discarding user interest in sequences.

Instead, we introduce a multiplicative enhancement of attention. The original attention is calculated by $A = \text{softmax}(\frac{QK^\top}{\sqrt{d}})$ and $Z = AV$. In Section 4.1, the causal relationships are derived from B where $X = BX + \Lambda U$ denotes the linear SCM among items (also corresponding to V). Therefore, we form an adjusted attention matrix to apply the causal relation matrix on the unnormalized attention weight. This ensures no critical behavior is outright discarded and achieves numerical stability for smooth training:

$$\tilde{A}^l = A^l \odot (\mathbf{1}_n \mathbf{1}_n^\top + \alpha R), \tag{9}$$

where $\odot$ denotes the Hadamard product or element-wise product, $A^l \in \mathbb{R}^{n_{max} \times n_{max}}$ is the attention matrix of the l_{th} CBA Layer, $\mathbf{1}_n := \{1, \ldots, 1\}^\top \in \mathbb{R}^n$ means the all-ones vector and α is a scalar hyperparameter controlling how strongly causal relationships influence the attention. $R \in \mathbb{R}^{n \times n}$ denotes the learned causal relationship matrix (using the method in Section 4.1) where 1 denotes the items have a causal relation and 0 denotes not. We then apply a standard attention softmax and the prefix mask:

$$Z^l = \text{softmax}(\mathcal{M} + \tilde{A}^l)V^l, \tag{10}$$

where $V^l = X^l W_V^l$ represents the values in the attention mechanism, with X^l as the input to the l_{th}

$$\mathcal{M}(x, y) = \begin{cases} 0 & if \quad x \leq y, \\ -\infty & otherwise. \end{cases}$$

4.2.2 Feed-Forward Network and Output Layer.
To enrich the capabilities of the model's representation, a two-layer point-wise feed-forward network is incorporated after each CBA Layer. This network introduces nonlinearity and facilitates interaction across latent dimensions. It is formulated as:

$$\hat{X}^l = (\text{ReLU}(\tilde{X}^l W_1^l + b_1^l))W_2^l + b_2^l, \tag{11}$$

where $W_1^l, W_2^l \in \mathbb{R}^{D \times D}$ are the learnable parameter matrix and $b_1^l, b_2^l \in \mathbb{R}^D$ is the learnable parameter vector. Following the transformer architecture [27], we apply residual connections, layer normalization, and dropout layers to alleviate overfitting. The output of the layer is defined as:

$$X^{l+1} = \text{LayerNorm}(X^l + \text{Dropout}, (\hat{X}^l)), \tag{12}$$

where the Dropout and Layer Normalization are defined as:

$$\text{Dropout}(\boldsymbol{x}) = \boldsymbol{r} \odot \boldsymbol{x},$$
$$\text{LayerNorm}(\boldsymbol{x}) = \theta_1 \odot \frac{\boldsymbol{x} - \mu}{\sqrt{\sigma^2 + \epsilon}} + \theta_2, \tag{13}$$

where μ and σ are the mean and variance of $\boldsymbol{x}$, θ_1 and θ_2 are learned scaling factors and bias terms. $\boldsymbol{r}$ is the random vector and $\boldsymbol{r}_i \sim \text{Bernoulli}(p)$ with probability parameter p.

4.3 Embedding Layer and Prediction Layer

4.3.1 Embedding Layer.
Following previous practices, we first truncate the given user behavior interaction sequence $o = I_1, I_2, \cdots, I_n$ by remove the early item $I_i, i > n_{max}$ and pad empty items for a short sequence $o_j, n < N_{max}$ to obtain fixed sequence set $O = \{u_k, I_1^k, I_2^k, \cdots, I_{N_{max}}^k\}_{k=1}^n$, where n_{max} denotes the maximum sequence length. We use an item embedding matrix $M \in \mathbb{R}^{|\mathcal{V}| \times D}$, where D denotes the hidden representation dimension, to define the embedding of the sequence $E^k = M_{o_k}$. To make our sequence more sensitive to the position of the sequence, we define the positional embedding matrix $P \in \mathbb{R}^{N_{max} \times D}$ and add it to the sequence embedding. Our embedding layer is written as below:

$$E^k = \text{Dropout}(M_{o_k} + P). \tag{14}$$

4.3.2 Prediction Layer.
The Prediction Layer serves as the final component of the CausalRec. Suppose the user's preference score for an item i is calculated based on the interaction sequence processed by the CausalBooster. The calculation is performed using the dot product between the embedding of item i and the output of the Encoder, which quantifies the similarity between the user's preference and the item's representation. It can be defined as below:

$$\hat{y}_i = e_i^T I_{n_k}^L, \tag{15}$$

where $\hat{y}$ represents the output logits, e_v denotes the representation of item i which comes from the item embedding matrix M and $I_{n_k}^L$ is the output representation of L_{th} CausalBooster layer.

4.4 Training Loss

Given the training, user interaction set $O_{\text{train}} = \{u_k, I_1^k, I_2^k, \cdots, I_{n_k}^k\}_{k=1}^n$, the primary goal is to mine causal relations to enhance recommendation performance. To achieve this, the loss function is composed of two components: (1) Recommendation Loss ($\mathcal{L}_{\text{rec}}$): Captures the model's ability to predict the next item in a user sequence. (2) DAG Constraint ($\mathcal{L}_{DAG}$): Maintaining sparsity and acyclicity constraints in the learned causal graph. The final loss function is as below:

$$\mathcal{L} = \mathcal{L}_{rec} + \lambda * \mathcal{L}_{L_1} + \mathcal{L}_{DAG}, \tag{16}$$

where λ is the penalty coefficient for the L_1 sparsity term. This unified loss ensures that the model learns accurate recommendations while simultaneously identifying reliable causal relationships.

4.4.1 Recommendation Component.
For the recommendation component, we treat the sequential recommendation as a next-item prediction problem and use a cross-entropy loss function, which is widely adopted in related tasks [18, 26, 32]. It is defined as:

$$\mathcal{L}_{rec} = - \sum_{o_k \in O_{train}} \sum_i^{n_k} \sum_{s \in \mathcal{S}} [y_{i,s} log(\sigma(\hat{y}_{i,s})) + \tag{17}$$
$$(1 - y_{i,s})log(1 - \sigma(\hat{y}_{i,s}))],$$

where $\sigma(\cdot)$ denotes the sigmoid function, $\mathcal{S}$ is the item set, and $y_{i,s}$ means whether item s is the next item in the user sequence u_k.

4.4.2 DAG Component. As described in Section 4.1, the DAG component consists of two parts: an acyclicity constraint and an L_1 sparse penalty. For the user u_k with its learned causal W_k, incorporating the L1 penalty $\mathcal{L}_{L1} = \sum_{o_k \in O_{train}} ||W_k||_1$ is simple. However, integrating the acyclicity constraint is the opposite, as the acyclicity constraint $\text{trace}(e^{W \odot W}) = n$ is nonconvex. Hence, followed by the NOTEARS, we transform it into an unconstrained subproblem:

$$\mathcal{L}_{DAG} = \sum_{o_k \in O_{train}} \frac{\rho}{2}|h(W_k)|^2 + \beta|h(W_k)|, \qquad (18)$$

where $h(W_k) = \text{trace}(e^{W_k \odot W_k}) - n$, and W_k is the learned causal graph for user u_k. The β follows the rule $\beta \leftarrow \beta + \rho\kappa$ to update after each epoch, where $\kappa = mean_{o_k \in O_{train}} h(W_k)$. Following NOTEARS [33], we set ρ update rule $\rho \leftarrow \rho * \gamma_1$ if $\kappa \geq \gamma_2 \kappa^-$ after each epoch, where κ^- denotes the κ in the last epoch and initially set to 0.

4.5 Complexity Analysis

4.5.1 Model Complexity. The CausalBoost model is based on the Transformer architecture. Its computational complexity is $O(bl^2d + bd)$, where the first term corresponds to the attention mechanism's complexity and the second term corresponds to the feed-forward network (FFN) complexity. The variables b, l, and d represent the batch size, sequence length, and hidden dimension, respectively.

4.5.2 Loss Complexity. The primary contributor to the loss computation complexity is the acyclicity constraint, which relies on the matrix exponential. Prior methods such as NOTEARS[33], NOFEARS[5], and NOBEARS[13] (see Table 1) introduce optimizations to reduce computation. Our approach follows NOTEARS to compute the constraint, which has the worst complexity of $O(l^3)$. This can achieve the theoretically optimal complexity with our covariance matrix, Eq.(7). As l is set to 200 and b is set to 256, the constraint complexity is similar to the model complexity, and compared with SASRec, our runtime increases by only 0.35s per epoch on average. It should be noted that our approach can be further optimized by blocking the matrix with a permutation matrix and computing constraints on each block matrix, whose complexity is $O(l^2)$ or $O(\sum_i^m r_i^3) = O(r^3)$. r_i is the rank of the block matrix and $\sum_i^m r_i = l$, $r_i < l$ with the biggest rank $r = \max\{r_i\}_{i=1,...,m}$, and the number of blocks m. The time complexity comparison among our approach, NOTEARS, and its variants is shown in the Table 1.

Table 1: Time complexity of different acyclicity constraints (worst / average / best).

Case	NOTEARS	NOBEARS	NOFEARS	Ours (covariance matrix)
Worst	$O(\ell^3)$	$O(\ell^3)$	$O(\ell^3)$	$O(\ell^3)$
Average	$O(\ell^3)$	$O(\ell^2)$	$O(\ell^3)$	$O(\ell^3)$
Best	$O(\ell^3)$	$O(\ell^2)$	$O(\ell^3)$	$O(r^3)$

4.5.3 Overall Complexity. As above, the total best complexity mainly comes from the attention mechanism: $O(bl^2d + bd + r^3) = O(bl^2d)$.

5 Experiments

In this section, we present comprehensive experiments to evaluate the effectiveness of CausalRec in sequential recommendation tasks. We first outline the experimental setup and describe the baseline

models in Section 5.1. Next, we compare CausalRec with existing baselines to assess its performance. Then, to validate the role of causality in CausalRec, we conduct both quantitative and qualitative experiments, providing further insights into the advantages of incorporating causality into sequential recommendations.

5.1 Experiment details

This subsection outlines the datasets, baselines, and implementation details used in our experiments. These components provide a comprehensive foundation for evaluating the performance of CausalRec in sequential recommendation tasks.

5.1.1 Dataset. We conduct our sequential recommendation experiments on the following real-world datasets: Movielens-1M[1] is a popular movie recommendation dataset collected from GroupLens Research, which contains user ratings on movies. LastFM[2] is a music recommendation dataset that contains user interaction with music, such as artist listening records. Foursquare[3] is a location-based recommendation dataset including user check-ins of restaurants in Tokyo for about 10 months. KGRec-music[4] is a music recommendation dataset collected from songfacts.com and last.fm websites. Table 3 summarizes the statistical information of the above datasets.

5.1.2 Baselines. To evaluate the effectiveness of our model, we compare it against well-known baselines:

- BPR: BPR [19] is a well-known recommendation model for capturing user implicit feedback. It is combined with matrix factorization to model user-item preferences.
- GRU4Rec: GRU4Rec [7] is a sequential recommendation model based on gated recurrent units. It leverages sequential user interactions to predict the next item.
- STAMP: STAMP [14] is a sequential recommendation model that emphasizes short-term user preferences while integrating long-term memory. It uses attention mechanisms to capture recent interactions.
- Causer: Causer [28] is a sequential recommendation model incorporating the learned causality on user behavior sequences in the clustering level. Identifying the clustering causal relationships enhances the understanding of user intent and improves recommendation accuracy.
- VTRNN: VTRNN [35] is a sequential prediction model that combines visual and textual features using a recurrent neural network to capture multimodal contextual information.
- SASRec: SASRec [9] is a recommendation model based on self-attention mechanisms. It captures long-term dependencies in user behavior by modeling user interactions.
- BSARec: BSARec [23] is a sequential recommendation model that introduces an attentive inductive bias to enhance predictions beyond traditional self-attention mechanisms.

5.1.3 Implementation details. In our experiments, we first organize the interactions of each user according to the timestamp. Then, following the common practice, we use the last and the second

---

[1]https://grouplens.org/datasets/movielens/

[2]http://millionsongdataset.com/lastfm/

[3]https://www.kaggle.com/datasets/chetanism/foursquare-nyc-and-tokyo-checkin-dataset

[4]https://www.upf.edu/web/mtg/kgrec

Table 2: Overall comparison between baselines and our models. The best performance is highlighted in bold, and the suboptimal results are shown with a dashed line below. All the numbers are percentage values with "%" omitted.

Datasets	Movielen-1m		Foursquare		LastFM		KGRec-music	
Metric@10	NDCG↑	HR↑	NDCG	HR	NDCG	HR	NDCG	HR
BPR	7.33	16.34	14.20	29.09	11.22	17.00	16.54	33.03
GRU4Rec	6.45	13.54	22.87	38.69	4.80	10.46	14.61	22.43
STAMP	21.46	42.16	17.41	35.18	3.50	7.06	38.27	73.68
Causer	6.51	14.07	4.37	10.16	5.04	10.78	12.53	22.95
VTRNN	19.88	44.40	21.09	30.93	4.74	9.82	26.97	47.64
SASRec	59.18	82.25	20.75	39.12	13.57	18.95	75.37	93.21
BSARec	60.75	78.61	24.45	39.24	13.46	19.03	74.55	93.13
Ours	**72.40**↑ 11.65	**88.94**↑ 6.69	**39.94**↑ 15.49	**53.89**↑ 14.65	**16.36**↑ 2.79	**24.00**↑ 4.97	**79.70**↑ 4.33	**95.77**↑ 2.56

Table 3: Statistics of the datasets, where the "SeqLen" denotes the average sequence length of each user.

Dataset	#User	#Item	#Interaction	#Sparsity	#SeqLen
Movielen-1m	6040	3952	1000209	95.81%	163.5
Foursquare	1083	38333	91024	99.78%	82.05
LastFM	1892	17632	92834	99.72%	47.08
KGRec-music	5199	8640	751531	98.32%	142.55

last interactions of each user behavior sequence as validation sets and testing sets, while others are used as training sets. For our architecture, we use two CBA layers for the CausalRec. The optimizer is the Adam optimizer, the learning rate is 0.001, and the batch size is 256. The dropout rate for all datasets is set to 0.2, and the maximum sequence length is set to 200 for all datasets. The hyperparameters, α, λ are selected based on grid search, and the ranges are $\{10^{-8}, 10^{-7}, 10^{-6}, \cdots, 10^6, 10^7, 10^8\}$. Others in baseline models are set to the default value in the original paper.

For the evaluation, we take the widely used metrics, including HR and NDCG, for model evaluation. Specifically, suppose A_u and B_u are the set of items recommended to user u and the ones interacted with by them in the testing set. Z is the number of recommended items. $R(i)$ is the relevance score, where $R(i) = 1$ denotes the ith item belongs to B_u, otherwise $R(i) = 0$. Then the formulas for computing HR and NDCG are:

$$\text{HR@Z} = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \frac{\sum_{i=1}^{Z} R(i)}{|B_u|} \tag{19}$$

$$\text{DCG}_u@Z = \sum_{i=1}^{Z} \frac{R(i)}{\log_2(i+1)} \tag{20}$$

$$\text{NDCG@Z} = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \frac{\text{DCG}_u@Z}{\max_{u \in \mathcal{U}}(\text{DCG}_u@Z)}. \tag{21}$$

In our experiment, we set $Z = 10$. To avoid heavy computation on all user-item pairs, we follow the strategy in [9]. For each user, we sample 100 negative items and rank them with the ground truth.

5.2 Recommendation Performance

Table 2 presents the results across four datasets. Our model achieves superior performance on all datasets, demonstrating average improvements of 8.56% in NDCG@10 and 7.21% in HR@10 over the

best baseline. Notably, the Foursquare dataset shows 15.49% and 14.65% improvements in NDCG and HR. These results validate the effectiveness of incorporating item-level causality into sequential recommendations, enhancing target item prediction by focusing on item causal relations. While Causer also introduces causality, its suboptimal performance stems from: (1) less effective RNN-based architectures for long sequences compared to our Transformer-based model; (2) learning pseudo-causal relationships at the cluster level from side information, which poorly represents sparse and complex real item causal relationships; and (3) a lack of identifiability in its learned causal relationships, leading to history information loss when filtering items.

5.3 Causality Evaluation

To evaluate the effect of causality and correlation in improving recommendation system performance, we conduct the ablation study below. Table 4 shows the results, where CausalRec(w/o Attention) denotes we remove the attention mechanism, relying solely on the causal relation matrix, CausalRec(w/o Causality) represents we remove the causal relation matrix, making it equivalent to SASRec, CausalRec(w/o sparse) denotes removing the L_1 constraint in the loss function, and CausalRec(w/filter) in Table 5 denotes we use a filtering strategy instead of an enhancing strategy. The filtering strategy of CausalRec(w/filter) can be written as follows:
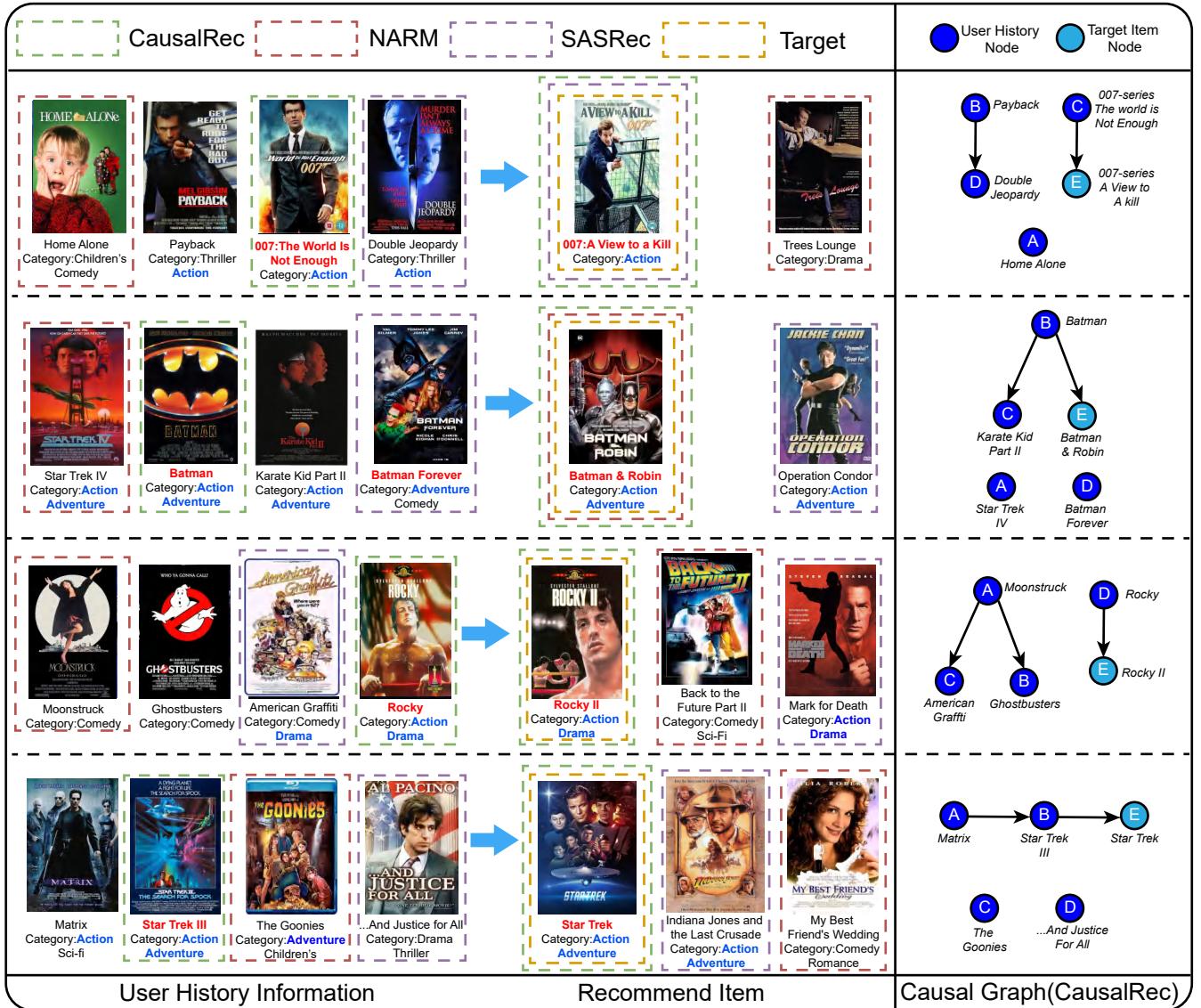
$$\tilde{A}^l = \text{softmax}(A^l + \mathcal{M}_R) \tag{22}$$

where all symbols are defined in Section 4.2. The $\mathcal{M}_R$ is defined as:

$$\mathcal{M}_R(x,y) = \begin{cases} -\infty & if \quad R_{x,y} \leq threshold \\ 0 & otherwise \end{cases} \tag{23}$$

where *threshold* is set to 0.9.

From the results, CausalRec(w/o Attention) underperforms CausalRec(w/o Causality) on KGRec-music, though causality still shows an advantage over correlation on the other three datasets. CausalRec(w/o Sparse) shows significant performance drops on Movielens-1M and Foursquare, emphasizing the necessity of sparse constraints. CausalRec(w/filter) improves performance on three datasets but significantly declines on KGRec-music (even below SASRec), indicating that improper fusion can degrade performance. To analyze our learned causality, we compared explanations generated by CausalRec, SASRec, and NARM (Fig. 3). Each row shows a user's

Figure 3: Visualization of user interaction histories (left), recommended items (center), and CausalRec's discovered causal graph (right). The film title in red on the left belongs to the same series as the target item in the middle (also shown in red), which is the user's actual choice that the model seeks to recommend. Categories shown in blue indicate the target item's category. Different colored boxes denote explanations from CausalRec (green), NARM (red), and SASRec (purple), while the orange box highlights the ground-truth (target) item.

Table 4: Comparison of CausalRec and its variants in terms of NDCG@10 and HR@10. All the numbers are percentage values with "%" omitted.

Dataset	Movielen-1m		Foursquare		Lastfm		KGRec-music	
Metric@10	NDCG	HR	NDCG	HR	NDCG	HR	NDCG	HR
CausalRec(w/o Attention)	$62.09_{\downarrow 10.31}$	$85.38_{\downarrow 3.56}$	$36.34_{\downarrow 3.60}$	$48.10_{\downarrow 5.79}$	$16.15_{\downarrow 0.21}$	$23.26_{\downarrow 0.74}$	$64.46_{\downarrow 15.24}$	$91.47_{\downarrow 4.30}$
CausalRec(w/o Causality)	$59.97_{\downarrow 12.43}$	$82.92_{\downarrow 6.02}$	$20.02_{\downarrow 19.92}$	$35.92_{\downarrow 17.97}$	$13.43_{\downarrow 2.93}$	$18.30_{\downarrow 5.70}$	$75.32_{\downarrow 4.38}$	$93.15_{\downarrow 2.62}$
CausalRec(w/o sparse)	$62.96_{\downarrow 9.44}$	$86.47_{\downarrow 2.47}$	$25.04_{\downarrow 14.90}$	$38.96_{\downarrow 14.93}$	$15.28_{\downarrow 1.08}$	$21.61_{\downarrow 2.39}$	$75.81_{\downarrow 3.89}$	$93.38_{\downarrow 2.39}$
CausalRec	**72.40**	**88.94**	**39.94**	**53.89**	**16.36**	**24.00**	**79.70**	**95.77**

Table 5: Comparison of filter strategy and Causalbooster in terms of NDCG@10 and HR@10. All the numbers are percentage values with "%" omitted.

Model		CausalRec(w/filter)	CausalRec
Dataset	Metric@10		
Movielen-1m	NDCG	$68.05_{\downarrow 4.35}$	**72.40**
	HR	$85.66_{\downarrow 3.28}$	**88.94**
Foursquare	NDCG	$30.36_{\downarrow 9.58}$	**39.94**
	HR	$45.06_{\downarrow 8.83}$	**53.89**
Lastfm	NDCG	$15.60_{\downarrow 0.76}$	**16.36**
	HR	$23.20_{\downarrow 0.80}$	**24.00**
KGRec-music	NDCG	$71.80_{\downarrow 7.90}$	**79.70**
	HR	$92.34_{\downarrow 3.43}$	**95.77**

history of actions (left to middle). The colored box on the left represents the explanation for the model recommendation, and the box on the right is the ground truth (orange) alongside models' recommendations: CausalRec (green), NARM (red), and SASRec (purple). We bold two elements: the film name if it is in the same series as the target, and the category if it matches that of the target film. We also show the learned causal graph of CausalRec on the right. In four examples, CausalRec consistently identifies another film from the same series as a causal explanation, which is intuitively sound. SASRec emphasizes category similarity, capturing thematic but less direct links, while NARM often provides unclear explanations with weak ties. These comparisons show that CausalRec more precisely identifies the causal link between user history and the recommendation. NARM or SASRec often locate semantically relevant items but miss the actual cause-and-effect relationship. These findings align with our experimental results, emphasizing that incorporating a causal relationship can substantially enhance the clarity and correctness of recommendation explanations.

6 Conclusion

We proposed CausalRec, a CausalBoost sequential recommendation model that integrates causality into attention mechanisms. To our knowledge, CausalRec is the first sequential recommendation model that integrates causal attention, effectively learns item-level causal relationships, and integrates them into the attention mechanism, enhancing predictions by emphasizing items with genuine causal effects on user preferences rather than mere correlations. In the transformer framework, CausalRec follows the causal identifiability condition as Lemma 4.1 due to the attention mechanism with layer normalization and achieves the best effect with a causal booster process. Experiments on four real-world datasets demonstrated CausalRec's superiority over state-of-the-art baselines in NDCG and HR metrics. The visual results show that CausalRec can capture underlying user patterns, such as a preference for film series, underscoring the interpretability benefits of modeling causality.

7 Ethical Considerations

CausalRec, a sequential recommendation model integrating causal attention, presents potential ethical risks, including biased recommendations due to implicit biases in historical user behavior data (which may disadvantage specific user groups), and security risks

of malicious attacks on causal graph data. Mitigation strategies involve bias auditing of training data using fairness-aware machine learning techniques, applying differential privacy to user interaction sequences and encrypted storage of causal graphs, conducting regular penetration testing on the model architecture, and establishing transparent audit trails for recommendation logic to ensure alignment with ethical principles and user welfare.

References

[1] Wenyu Chen, Mathias Drton, and Y Samuel Wang. 2019. On causal discovery with an equal-variance assumption. *Biometrika* 106, 4 (Sept. 2019), 973–980. https://doi.org/10.1093/biomet/asz049

[2] David Maxwell Chickering. 2002. Optimal structure identification with greedy search. *Journal of Machine Learning Research* 3 (2002), 507–554. https://www.jmlr.org/papers/volume3/chickering02a/chickering02a.pdf

[3] Diego Colombo and Marloes H Maathuis. 2014. Order-independent constraint-based causal structure learning. *The Journal of Machine Learning Research* 15, 1 (2014), 3741–3782.

[4] Ruifei Cui, Perry Groot, and Tom Heskes. 2016. Copula PC algorithm for causal discovery from mixed data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, Springer, Riva del Garda, Italy, 377–392.

[5] Yue Yu Dennis Wei, Tian Gao. 2020. DAGs with No Fears: A Closer Look at Continuous Optimization for Learning Bayesian Networks. In *Advances in Neural Information Processing Systems*. Curran Associates Inc., Red Hook, NY, USA, 3895–3906. https://arxiv.org/abs/2010.09133

[6] Olivier Goudet, Diviyan Kalainathan, Philippe Caillou, Isabelle Guyon, David Lopez-Paz, and Michèle Sebag. 2018. Causal Generative Neural Networks.

[7] Balázs Hidasi and Alexandros Karatzoglou. 2018. Recurrent Neural Networks with Top-k Gains for Session-based Recommendations. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management* (Torino, Italy) *(CIKM '18)*. Association for Computing Machinery, New York, NY, USA, 843–852. https://doi.org/10.1145/3269206.3271761

[8] Patrik O. Hoyer, Dominik Janzing, Joris M. Mooij, Jonas Peters, and Bernhard Schölkopf. 2008. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems*, Vol. 21. Curran Associates, Inc., New Orleans, USA, 689–696. https://proceedings.neurips.cc/paper/2008/file/f7664060cc52bc6f3d620bcedc94a4b6-Paper.pdf

[9] Wang-Cheng Kang and Julian McAuley. 2018. Self-Attentive Sequential Recommendation. In *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, Singapore, 197–206. https://doi.org/10.1109/ICDM.2018.00035

[10] Nan Rosemary Ke, Olexa Bilaniuk, Anirudh Goyal, Stefan Bauer, Hugo Larochelle, Chris Pal, and Yoshua Bengio. 2019. Learning Neural Causal Models from Unknown Interventions. arXiv:1711.08936

[11] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix Factorization Techniques for Recommender Systems. *Computer* 42, 8 (2009), 30–37. https://doi.org/10.1109/MC.2009.263

[12] Mathieu Lachapelle, Pierre-Luc Brouillard, Gianluca Bontempi, and Simon Lacoste-Julien. 2020. GraN-DAG: Gradient-based Neural DAG Learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*. OpenReview.net, Addis Ababa, Ethiopia. https://arxiv.org/abs/1906.02226

[13] Hao-Chih Lee, Matteo Danieletto, Riccardo Miotto, Sarah T Cherng, and Joel T Dudley. 2019. Scaling structural learning with NO-BEARS to infer causal transcriptome networks. In *Pacific Symposium on Biocomputing*. World Scientific, 5 Toh Tuck Link, Singapore 596224, 391–402.

[14] Qingyao Liu, Lixin Zou, Keping Yang, Jipeng Zhang, and Jie Zhang. 2018. 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 (KDD)*. ACM, London, United Kingdom, 1831–1839. https://dl.acm.org/doi/10.1145/3219819.3220082

[15] Shanlei Mu, Yaliang Li, Wayne Xin Zhao, Jingyuan Wang, Bolin Ding, and Ji-Rong Wen. 2022. Alleviating Spurious Correlations in Knowledge-aware Recommendations through Counterfactual Generator. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Madrid, Spain) *(SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 1401–1411. https://doi.org/10.1145/3477495.3531934

[16] Irene Ng, Kun Zhang, and Bryon Aragam. 2020. Masked Gradient-Based Causal Structure Learning. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., Virtual, 18464–18475.

[17] Gunwoong Park. 2020. Identifiability of Additive Noise Models Using Conditional Variances. *Journal of Machine Learning Research* 21, 75 (2020), 1–34. http://jmlr.org/papers/v21/19-664.html

[18] Ruihong Qiu, Chen Gao, Xiangnan He, and Yong Li. 2022. Contrastive Learning for Representation Degeneration Problem in Sequential Recommendation. In

Proceedings of the 15th ACM International Conference on Web Search and Data Mining (WSDM). ACM, Phoenix, Arizona, USA, 813–821. https://arxiv.org/pdf/2110.05730

[19] Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. 2010. Factorizing personalized Markov chains for next-basket recommendation. In *Proceedings of the 19th International Conference on World Wide Web* (Raleigh, North Carolina, USA) *(WWW '10)*. Association for Computing Machinery, New York, NY, USA, 811–820. https://doi.org/10.1145/1772690.1772773

[20] Raanan Y. Rohekar, Yaniv Gurwicz, and Shami Nisimov. 2023. Causal Interpretation of Self-Attention in Pre-Trained Transformers. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., New Orleans, Louisiana, USA, 31450 − 31465. https://arxiv.org/abs/2310.20307

[21] Guy Shani, David Heckerman, and Ronen I. Brafman. 2005. An MDP-Based Recommender System. *Journal of Machine Learning Research* 6, 43 (2005), 1265–1295. http://jmlr.org/papers/v6/shani05a.html

[22] Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, and Antti Kerminen. 2006. A Linear Non-Gaussian Acyclic Model for Causal Discovery. *Journal of Machine Learning Research* 7 (2006), 2003–2030. https://jmlr.org/papers/volume7/shimizu06a/shimizu06a.pdf

[23] Yehjin Shin, Jeongwhan Choi, Hyowon Wi, and Noseong Park. 2024. An Attentive Inductive Bias for Sequential Recommendation beyond the Self-Attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. AAAI Press, Vancouver, Canada, 8984–8992. https://arxiv.org/abs/2312.10325

[24] Peter Spirtes, Clark Glymour, and Richard Scheines. 2000. *Causation, Prediction, and Search* (2nd ed.). MIT Press, Cambridge, MA. https://mitpress.mit.edu/9780262527927/causation-prediction-and-search/

[25] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. BERT4Rec: Sequential Recommendation with Bidirectional Encoder Representations from Transformer. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management* (Beijing, China) *(CIKM '19)*. Association for Computing Machinery, New York, NY, USA, 1441–1450. https://doi.org/10.1145/3357384.3357895

[26] Zhen Tian, Xiang Wang, An Zhang, Fuli Feng, and Tat-Seng Chua. 2023. Frequency Enhanced Hybrid Attention Network for Sequential Recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. Association for Computing Machinery, New York, NY, USA, 78–88. https://arxiv.org/pdf/2304.09184

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*. Curran Associates Inc., Red Hook, NY, USA, 5998–6008.

[28] Zhenlei Wang, Xu Chen, Rui Zhou, Quanyu Dai, Zhenhua Dong, and Ji-Rong Wen. 2023. Sequential Recommendation with Causal Behavior Discovery. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*. IEEE, IEEE, Anaheim, CA, USA, 28–40. https://doi.org/10.1109/ICDE55852.2023.00011

[29] Song-Li Wu, Liang Du, Jia-Qi Yang, Yu-Ai Wang, De-Chuan Zhan, Shuang Zhao, and Zi-Xun Sun. 2024. RE-SORT: removing spurious correlation in multilevel interaction for CTR prediction. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence (UAI '24)*. JMLR.org, Barcelona, Spain, Article 178, 13 pages.

[30] An Zhang, Fangfu Liu, Wenchang Ma, Zhibo Cai, Xiang Wang, and Tat-Seng Chua. 2023. Boosting Differentiable Causal Discovery via Adaptive Sample Reweighting. In *Proceedings of the Eleventh International Conference on Learning Representations*. ICLR, International Conference on Learning Representations, Kigali, Rwanda.

[31] Kun Zhang and Aapo Hyvärinen. 2009. Distinguishing causes from effects using nonlinear acyclic causal models. In *Proceedings of the 15th International Conference on Neural Information Processing*. JMLR.org, Whistler, Canada, 157–164. https://www.cs.helsinki.fi/u/ahyvarin/papers/Zhang09NIPSworkshop.pdf

[32] Tingting Zhang, Pengpeng Zhao, Yanchi Liu, Victor S Sheng, Jiajie Xu, Deqing Wang, Guanfeng Liu, and Xiaofang Zhou. 2019. Feature-level Deeper Self-Attention Network for Sequential Recommendation. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*. AAAI Press, Macao, China, 4320–4326. https://www.ijcai.org/Proceedings/2019/0600.pdf

[33] Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. 2018. DAGs with NO TEARS: Continuous Optimization for Structure Learning. In *Advances in Neural Information Processing Systems*. Curran Associates Inc., Red Hook, NY, USA, 9472–9483. https://arxiv.org/abs/1803.01422

[34] Xun Zheng, Chen Dan, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. 2020. Learning Sparse Nonparametric DAGs. In *International Conference on Artificial Intelligence and Statistics*. PMLR, Breckenridge, Colorado, USA, 3414–3425.

[35] Yuyin Zhu, Shuang Li, Bing Li, Xian-Sheng Wang, and Hongyuan Zha. 2017. Variational Recurrent Neural Networks for Session-based Recommendation. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*. Association for Computing Machinery, New York, NY, USA, 3215–3221.