



# Intrinsic and Extrinsic Factor Disentanglement for Recommendation in Various Context Scenarios

YIXIN SU\*<sup>†</sup>, School of Computer Science and Technology, Huazhong University of Science and Technology, China

WEI JIANG\*, Alibaba Group, China

FANGQUAN LIN, Alibaba Group, China

CHENG YANG, Alibaba Group, China

SARAH M. ERFANI, The University of Melbourne, Australia

JUNHAO GAN, The University of Melbourne, Australia

YUNXIANG ZHAO<sup>‡</sup>, Laboratory of Advanced Biotechnology, Beijing Institute of Biotechnology, China

RUIXUAN LI, School of Computer Science and Technology, Huazhong University of Science and Technology, China

RUI ZHANG<sup>‡</sup>, School of Computer Science and Technology, Huazhong University of Science and Technology (www.ruizhang.info), China

In recommender systems, the patterns of user behaviors (e.g., purchase, click) may vary greatly in different contexts (e.g., time and location). This is because user behavior is jointly determined by two types of factors: *intrinsic factors*, which reflect consistent user preference, and *extrinsic factors*, which reflect external incentives that may vary in different contexts. Differentiating between intrinsic and extrinsic factors helps learn user behaviors better. However, existing studies have only considered differentiating them from a single, pre-defined context (e.g., time or location), ignoring the fact that a user's extrinsic factors may be influenced by the interplay of various contexts at the same time. In this paper, we propose the Intrinsic-Extrinsic Disentangled Recommendation (IEDR) model, a generic framework that differentiates intrinsic from extrinsic factors considering various contexts simultaneously, enabling more accurate differentiation of factors and hence the improvement of recommendation accuracy. IEDR contains a context-invariant contrastive learning component to capture intrinsic factors, and a disentanglement component to extract extrinsic factors under the interplay of various contexts. The two components work together to achieve effective factor learning. Extensive experiments on real-world datasets demonstrate

\*Both authors contributed equally to this research.

<sup>†</sup>Yixin Su did this work when he was an intern at Alibaba Group.

<sup>‡</sup>Corresponding authors.

---

Authors' Contact Information: Yixin Su, yixin.su@outlook.com, School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China; Wei Jiang, wwjiangwei@hotmail.com, Alibaba Group, Hangzhou, China; Fangquan Lin, fangquan.linfq@alibaba-inc.com, Alibaba Group, Hangzhou, China; Cheng Yang, charis.yangc@alibaba-inc.com, Alibaba Group, Hangzhou, China; Sarah M. Erfani, sarah.erfani@unimelb.edu.au, The University of Melbourne, Melbourne, Australia; Junhao Gan, junhao.gan@unimelb.edu.au, The University of Melbourne, Melbourne, Australia; Yunxiang Zhao, zhaoyx1993@163.com, Laboratory of Advanced Biotechnology, Beijing Institute of Biotechnology, Beijing, China; Ruixuan Li, rxli@hust.edu.cn, School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China; Rui Zhang, rayteam@yeah.net, School of Computer Science and Technology, Huazhong University of Science and Technology (www.ruizhang.info), Wuhan, China.

---

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, or post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

© 2025 Copyright held by the owner/author(s).

ACM 1558-2868/2025/3-ART

<https://doi.org/10.1145/3722553>

IEDR’s effectiveness in learning disentangled factors and significantly improving recommendation accuracy by up to 4% in NDCG.

CCS Concepts: • **Information systems** → **Recommender systems**; *Data mining*; *Personalization*; • **Computing methodologies** → **Knowledge representation and reasoning**.

Additional Key Words and Phrases: Recommender Systems, Intrinsic and Extrinsic Factors, Contrastive Learning, Disentanglement, Mutual Information

## 1 INTRODUCTION

Recommender systems [20, 27, 39, 58] aim to predict the probability of a user’s behavior (e.g., purchase, click) on a given item. This is a challenging task since a user’s behavior may vary significantly across different *contexts* (e.g., time, location, and social setting). For example, considering the context of social settings (e.g., alone vs. with friends), when recommending food, a user may prefer healthy food like steamed vegetables and salad when being alone, but may prefer more diverse food suitable for sharing like hot pot or pizza when gathering with friends. This context-dependent variation in user behaviors underscores their complex nature. Psychological research has devoted great efforts to understanding this phenomenon, and reveals that user behaviors are influenced by two types of factors: *intrinsic* and *extrinsic* factors [3, 35], distinguished by whether they can be influenced by context changes. An intrinsic factor, which is often stable for a user across different contexts, is an internal motivation for inherent satisfaction. In our food recommendation example, the preference for healthy food when eating alone could be driven by intrinsic factors such as personal health goals or taste preferences. In contrast, an extrinsic factor, which is an external motivation stimulated by the contexts, often varies when contexts change [26]. The choice of more diverse food when gathering with friends could be influenced by extrinsic factors such as the social setting. Therefore, to better understand user behaviors and provide more accurate recommendations, it is crucial yet challenging for recommender systems to effectively capture and differentiate between intrinsic and extrinsic factors in various contexts.

Existing studies that aim to differentiate between intrinsic and extrinsic factors consider only a single, pre-defined context, e.g., time [9, 53] or location [13, 16]. However, in reality, user behaviors are often influenced by the interplay of various contexts simultaneously. These methods may not be able to accurately capture user behaviors, especially when contexts change (an example will be given in the next paragraph). Moreover, these methods are designed specifically for the pre-defined context. For example, Li et al. [16] leverages location context to differentiate intrinsic and extrinsic factors. They incorporate a context-specific assumption into their model that the choice of a long geographical distance place is more influenced by intrinsic factors and vice versa. Consequently, it is difficult to extend these methods to scenarios where multiple types of contexts may affect the result. For instance, this location-specific assumption cannot be adapted to a social setting context.

Given these limitations, in this paper, we aim to capture and differentiate between intrinsic and extrinsic factors from various contexts, thereby enhancing the ability to learn user behaviors. To this end, we adopt an approach from a more fundamental perspective without introducing any context-specific assumptions. Under this general context condition, we first define intrinsic and extrinsic factors by focusing on whether these factors vary when contexts change. Following this definition, we propose an Intrinsic-Extrinsic Disentangled Recommendation (IEDR) model, a general framework that can effectively capture the interplay of various contexts and differentiate intrinsic and extrinsic factors within them. To illustrate the importance of accurately differentiating between intrinsic and extrinsic factors in scenarios with various contexts, consider the example in Figure 1. A user called Bob generally prefers healthy food but enjoys diverse food when gathering with friends (top left of the figure). The dataset happens to only contain Bob’s behaviors in cold weather (top right of the figure), where Bob has steamed vegetables (warm healthy food) when alone and hot pot (diverse option) with friends. Existing models differentiate between intrinsic and extrinsic factors from only one of the contexts, such as social settings (i.e.,

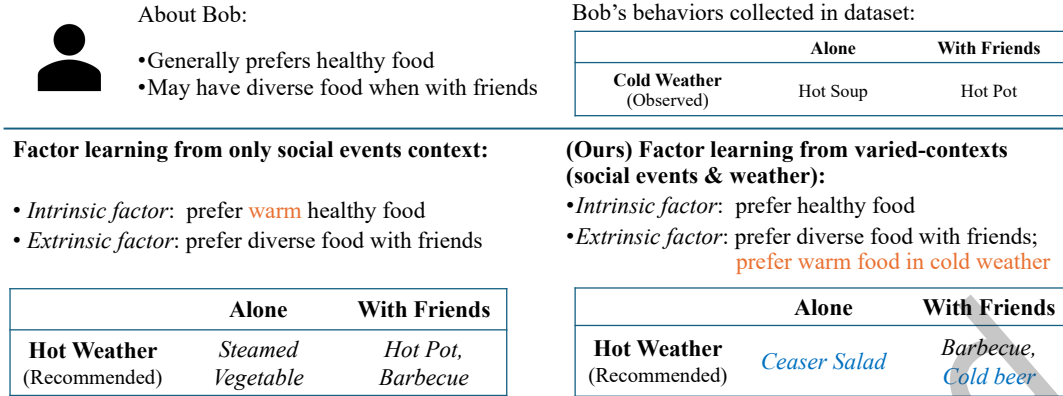


Fig. 1. An example to compare existing work (consider only the context of social settings) and our approach (consider various contexts) in learning intrinsic and extrinsic factors. The upper part shows the preference fact (upper left) and observed behaviors (upper right) of a user Bob. The bottom part shows the possible factor learning results and corresponding recommendations of existing work (bottom left) and our approach (bottom right).

alone vs. with friends) in this example. They might incorrectly identify warm food preference as Bob's intrinsic factor (lower left of the figure). This is because the model treats the weather context (i.e., cold vs. hot) as a regular feature rather than a context used for factor differentiation. The weather-dependent influence may show similar patterns across different social settings (e.g., warm foods are chosen either when alone or with friends), leading to weather-dependent extrinsic factors being mistakenly identified as intrinsic factors. In contrast, our model considers various contexts for differentiating the factors (lower right of the figure). Since a strong correlation may exist between weather and warm/cold food choices (e.g., most users may choose warm food in cold weather and cold food in hot weather), our model captures such weather-dependent preferences as extrinsic factors. Bob's choices of warm food all occur in cold weather, fitting well with the weather-dependent preference pattern (i.e., preferring warm food in cold weather). Therefore, our model can accurately capture such choices as being influenced by extrinsic factors. When in hot weather scenarios (shown in the bottom two tables of the figure), existing models (left table) may incorrectly recommend hot food due to misidentified intrinsic factors. In comparison, our model (right table) adapts to the weather context, recommending more suitable cold options like Caesar salad and cold beer.

The IEDR framework consists of two main modules: a recommendation prediction (RP) module and a contrastive intrinsic-extrinsic disentanglement (CIED) module. To better capture the interplay among different contexts, the RP module constructs various contexts into a graph structure, where each context is represented as a node and their interplay (interactions) is represented as edges, and a complete graph is constructed. By applying graph learning algorithms to this context graph, the model can comprehensively learn the complex relationships and mutual influences between contexts, enabling it to obtain more informative context representations. Similarly, user and item representations are obtained from their respective attributes (e.g., user gender, item category). The core innovation of IEDR lies in the CIED module, which leverages the synergy between a context-invariant *contrastive learning component* and a mutual information minimization-based *disentangling component* to effectively differentiate intrinsic and extrinsic factors into disentangled representations. The contrastive learning component captures user preference that is stable across contexts by contrasting user representations under different contextual conditions. Concurrently, the disentangling component employs a bidirectional mutual information minimization scheme to separate the extrinsic factors that vary with different contexts from the

intrinsic factors. By jointly optimizing these two components, IEDR ensures that the learned intrinsic factors are not only stable across different contexts but also well-separated from the extrinsic factors. This innovative approach enables IEDR to effectively learn disentangled intrinsic and extrinsic factors, capturing the complex user behavior patterns for recommendation in various context scenarios.

In this paper, we make the following contributions:

- We formally define intrinsic and extrinsic factors for recommender systems. Based on this definition, we propose IEDR, a novel framework that effectively learns intrinsic and extrinsic factors for more accurate recommendations. This is achieved by introducing two key components: a context-invariant contrastive learning component and a mutual information minimization-based disentangling component. These components work together to effectively capture the two types of factors from the interplay of various contexts. The implementation of IEDR is available at <https://github.com/ethanmock/IEDR>.
- We theoretically analyze the proposed methods from an information theory perspective, providing insights into the effectiveness of our approach. We also identify key challenges and propose principled solutions to avoid degenerating results and ensure robust disentanglement, thereby improving recommendation accuracy and stability.
- Extensive experiments on real-world datasets demonstrate that (1) IEDR significantly outperforms state-of-the-art methods by up to 4% in NDCG, and (2) the proposed CIED module effectively learns disentangled intrinsic and extrinsic factors, leading to improved recommendation accuracy.

## 2 RELATED WORK

This section summarizes the current research progress related to our work on factor disentanglement, feature interactions in recommender systems, and contrastive learning.

### 2.1 Factor disentanglement

Intrinsic and extrinsic factors are considered as two basic factors for individual decision-making in psychological research [3, 26, 35]. Recent recommender systems have borrowed the idea of capturing these two factors to achieve more accurate recommendations. For example, in the sequential recommendation, Hidasi et al. [13] leverage the recurrent neural networks to capture users' long- and short-term (LS-term) interests from their interacted item sequences. Yu et al. [53] propose a time-aware controller to capture the differences between LS-term interests for more accurate interest learning. Zheng et al. [56] further emphasize the disentanglement between the LS-term interests at different time scales to differentiate the LS-term interests. Ning et al. [23] demonstrate the effectiveness of embedding disentanglement by separating inter-domain and intra-domain knowledge. Wang et al. [41] propose a Causal Disentangled Recommendation framework to handle user preference shifts by modeling the interaction generation procedure using a causal graph. In point-of-interest recommendation, studies are leveraging spatial context to capture the intrinsic and extrinsic factors [16, 45]. However, all of the above studies focus on specific contexts. As a result, their factor learning approaches are hard to apply to other recommendation domains, which may result in a suboptimal solution if other contexts jointly influence these factors. Some studies learn users' various factors without knowing the meaning of each factor (i.e., implicit factor). They first define the number of factors (e.g., 4) to be learned, and then disentangle the representations of each pair of factors [21, 42]. Compared to previous studies that focus on specific contexts or learn implicit factors, our IEDR model provides a generic framework to explicitly learn intrinsic and extrinsic factors from various contexts, enabling effective modeling of the complex interplay between stable user preference and various contextual influences in real-world recommendation scenarios.

## 2.2 Feature interaction modeling

Many recommender systems leverage feature interactions to improve recommendation accuracy. One of the most common techniques is the factorization machine (FM) [25], which models feature interactions through dot product and achieves great success. Recent studies extend FM with deep neural networks for more powerful feature interaction modeling [12, 31, 46, 52]. The Wide & Deep model (WDL) [7] proposes a framework that combines shallow and deep modeling of features for recommendation. [11] combines FM and WDL by replacing the shallow part of WDL with an FM model. [30] leverages the relation reasoning power of graph neural networks for feature interaction modeling. We are the first work to represent various contexts as a feature graph, and leverage graph neural networks to capture the interplay of the contexts in a feature interaction modeling paradigm for unified context learning.

## 2.3 Contrastive learning

Contrastive learning has achieved great success in computer vision [6], neural language processing [24], graph learning [5, 55] and music learning [47]. Recently, contrastive learning has attracted attention in recommender systems. Yao et al. [48] conduct contrastive learning on users and items respectively on a two-tower framework to learn robust user and item representations. In addition, Wu et al. [44] propose a contrastive learning framework on a user-item bipartite graph to capture robust high-degree relationships between users and items. Ye et al. [49] leverage contrastive learning on perturbed embeddings to improve the robustness of neural graph collaborative filtering. Wang et al. [36] propose a general framework called ContraRec that unifies two kinds of contrastive learning tasks, context-target contrast and context-context contrast, for sequential recommendation. Some studies enhance recommendation through contrastive learning by mitigating popularity bias and promoting long-tail items with noise-based embedding augmentations [50, 51]. Zhang et al. [54] propose AdvInfoNCE to handle false negatives and improve generalization. Cai et al. [4] introduce LightGCL, using singular value decomposition to refine semantic structures and improve robustness. NCL incorporates structural and semantic neighbors as positive pairs for better user-item relationship learning [19]. The CETN model [17] addresses the challenge of capturing diverse and homogeneous feature interactions across semantic spaces by employing contrastive learning and self-supervised signals. These works use contrastive learning to enhance recommendation by addressing bias, improving robustness, and promoting long-tail items. Unlike previous works, we propose a context-invariant contrastive learning approach to capture stable intrinsic factors across various contexts, which is integrated with a mutual information minimization scheme to disentangle context-specific extrinsic factors.

## 3 PRELIMINARY

In this section, we introduce two key techniques that lay the foundation for our proposed method: the Statistical Interaction Graph Network (SIGN) [30] for effective feature interaction modeling, and the Variational Contrastive Log-ratio Upper Bound (vCLUB) [8] for mutual information estimation and minimization.

### 3.1 Statistical Interaction Graph Network (SIGN)

The statistical interaction graph network (SIGN) [30] explicitly models feature interactions through a graph neural network. Given a set of features (e.g., user/item attributes) of each data sample,  $\mathcal{Z} = \{z_1, z_2, \dots, z_n\}$ , SIGN regards  $\mathcal{Z}$  as a feature graph  $\mathcal{G}(\mathcal{Z}, \mathcal{E})$ , where  $\mathcal{Z}$  is the node set that each feature  $z_i$  is a node, and  $\mathcal{E}$  is the edge set containing all the combinations of pairwise feature interactions, with each feature interaction  $\langle z_i, z_j \rangle$  being an edge linking to corresponding nodes. Accordingly, user representation learning becomes a graph learning problem.

In SIGN, first, each feature  $z_i$  is mapped into a feature embedding  $z_i \in \mathbb{R}^d$  of  $d$  dimensions as the node embedding. The embeddings are first randomly initialized and are updated through training. Then, SIGN learns

the graph representation (e.g., a vector) using a function  $f$ :

$$f(\mathcal{G}) = \phi(\{\psi(\{e_{ij}h(z_i, z_j)\}_{j \in \mathcal{Z}})\}_{i \in \mathcal{Z}},$$

where  $\phi$  and  $\psi$  are aggregation functions (e.g., element-wise mean),  $h(\cdot) : \mathbb{R}^{2 \times d} \rightarrow \mathbb{R}^d$  is an MLP that models each feature interaction,  $e_{ij} \in \{0, 1\}$  is the edge indicator (since we use all pairwise feature interactions,  $e_{ij} = 1$  for all edges).  $f$  outputs the modeled graph representation  $\mathbf{u} \in \mathbb{R}^d$  of  $d$  dimensions.

### 3.2 Variational Contrastive Log-ratio Upper Bound (vCLUB) of Mutual Information

Given a set of sample pairs  $\{(A_i, B_i)\}_{i=1}^N$  drawn from an unknown distribution  $p(A, B)$  of random variables  $A$  and  $B$ . The vCLUB method [8] derives the upper bound of their mutual information  $I(A, B)$  as:

$$\mathcal{I}_{\text{vCLUB}}(A; B) := \mathbb{E}_{p(A, B)} [\log q_{\theta}(A|B)] - \mathbb{E}_{p(A)p(B)} [\log q_{\theta}(A|B)], \quad (1)$$

where  $p(A, B)$  is the joint distribution,  $p(A)p(B)$  is the marginal distribution,  $q_{\theta}(A|B)$  is a variational distribution of parameter  $\theta$  (e.g., an MLP) to predict  $A$  given  $B$ .

In an application of mutual information minimization, we aim to reduce the correlation between  $A_i$  and  $B_i$  by selecting an optimal parameter  $\sigma$  of the joint variational distribution  $p_{\sigma}(A, B)$ . vCLUB performs mutual information estimation and minimization in two steps iteratively. In the first step, to ensure Equation (1) holds as the upper bound,  $\theta$  is trained to make the log-likelihood function  $\mathcal{L}(A, B) := \frac{1}{N} \sum_{i=1}^N \log q_{\theta}(A_i|B_i)$  maximized (Theorem 3.2 of [8]). In the second step,  $\theta$  is frozen, and other parameters ( $\sigma$ ) are trained to minimize  $\mathcal{I}_{\text{vCLUB}}(A; B)$  so that the mutual information is minimized.

## 4 PROBLEM STATEMENT AND DEFINITIONS

Let  $\mathcal{U}$ ,  $\mathcal{V}$ , and  $\mathcal{C}$  denote the user set, item set, and context set, respectively. Each user  $u \in \mathcal{U}$  consists a set of user features  $u = \{z_1^u, z_2^u, \dots, z_p^u\}$  (e.g., user ID, gender). Similarly, each item  $v \in \mathcal{V}$  is represented by a set of item features  $v = \{z_1^v, z_2^v, \dots, z_q^v\}$  (e.g., branch, color). A context  $c \in \mathcal{C}$  is a set of context features  $c = \{z_1^c, z_2^c, \dots, z_m^c\}$ , denoting the context state when a user selects an item (e.g., weather, daytime). Let  $\mathcal{D}$  be a dataset containing  $N$  instances (i.e., data samples) of  $(u, v, c)$ , with a corresponding label  $y \in \{1, 0\}$  indicating whether or not the user  $u$  selects the item  $v$  under the context  $c$ . The recommendation task can be formulated as predicting the selection probability  $y' = p(u, v, c)$ . In our proposed IEDR model, the intrinsic factor  $\mathbf{o}_{in}$  and the extrinsic factor  $\mathbf{o}_{ex}$  are explicitly inferred for both users and items, and jointly leveraged to perform the prediction.

Next, we formally define intrinsic and extrinsic factors. We believe these two factors exist from both users' and items' perspectives. This is reasonable since a user selecting an item not only relates to the factors (motivations) of users, e.g., prefer *healthy food* (intrinsic factor) on *weekdays* (extrinsic factor), but also relates to the factors (attractiveness) of items, e.g., the Caesar salad is *healthy* (intrinsic factor) and is chosen more often when *the weather is hot* (extrinsic factor). In the following, we define intrinsic and extrinsic factors from the users' perspective only, as they are similar from the items' perspective.

**DEFINITION 1. (Intrinsic Factor and Extrinsic Factor)** Consider a user  $u$  and a set of contexts  $\mathcal{C}$ ; an **intrinsic factor** of the user is a factor that is invariant to the contexts in  $\mathcal{C}$ , i.e.,  $f_{in}(u, c) = f_{in}(u, c')$ , where  $f_{in}$  is a function learning intrinsic factor representations, and  $c$  and  $c'$  are two arbitrary contexts in  $\mathcal{C}$ . On the other hand, an **extrinsic factor** of the user is a factor that is different from its corresponding intrinsic factor, i.e.,  $\mathcal{I}(f_{in}(u, c), f_{ex}(u, c)) = 0$ , where  $\mathcal{I}$  computes the mutual information and  $f_{ex}$  learns extrinsic factor representations. Also, the extrinsic factor changes w.r.t. the context, i.e., there exist contexts  $c$  and  $c'$  in  $\mathcal{C}$  such that  $f_{ex}(u, c) \neq f_{ex}(u, c')$ .

In the definition,  $f_{in}(u, c) = f_{in}(u, c')$  shows the invariance of intrinsic factors. On the other hand,  $f_{ex}(u, c) \neq f_{ex}(u, c')$  shows that the extrinsic factors can be different if the contexts are different.

Table 1. Summary of notations used in the IEDR model.

Notation	Description
$U, V, C$	Sets of users, items, and contexts, respectively.
$z_i^u, z_i^v, z_i^c$	The $i^{th}$ feature representation of user $u$ , item $v$ , and context $c$ .
$\mathbf{u}, \mathbf{v}, \mathbf{c}$	User, item, and context representations.
$\mathbf{o}_{in}, \mathbf{o}_{ex}$	Intrinsic and extrinsic factor representations.
$\mathcal{L}_{RP}$	Recommendation prediction loss.
$\mathcal{L}_{CIKL}$	Context-invariant contrastive learning loss.
$\mathcal{L}_{bi-appr}$	Bidirectional approximation loss for disentanglement.
$\mathcal{L}_{Dis}$	Disentanglement loss.

In previous research (both in psychology [3, 35] and in recommender systems [13, 53]), intrinsic and extrinsic factors are considered all the factors influencing user behavior, and learning these two factors in a disentangled way has proven effective to analyze these behaviors [56]. Therefore, it leads to our factor learning objective based on Definition 1: leveraging the context-invariant property to ensure that  $f_{in}$  captures intrinsic factors, and disentangling the outputs of  $f_{in}(u, c)$  and  $f_{ex}(u, c)$  to ensure  $f_{ex}$  captures extrinsic factors (detailed in Section 5.2).

## 5 INTRINSIC-EXTRINSIC DISENTANGLED RECOMMENDATION MODEL

To effectively learn and disentangle intrinsic and extrinsic factors from various contexts, we propose our Intrinsic-Extrinsic Disentangled Recommendation (IEDR) Model. The overview of our model is visualized in Figure 2. More specifically, our proposed IEDR model consists of the following two modules, which will be detailed in the next subsections:

- A recommendation prediction (RP) module that takes a user and an item as input, and combines them with a set of contexts, to generate intrinsic and extrinsic factor representations for both the user and the item. The predicted probability  $y'$  is then jointly learned from these representations.
- A contrastive intrinsic-extrinsic disentangling (CIED) module is applied to both the user and the item sides to support the intrinsic and extrinsic factor learning. The module contains a context-invariant contrastive learning component and a disentangling component, to ensure the learned factors satisfy Definition 1.

For clarity and ease of understanding, Table 1 summarizes the key notations used throughout the IEDR model.

### 5.1 Recommendation Prediction (RP) Module

The recommendation prediction (RP) module is a symmetric structure that generates user intrinsic and extrinsic factor representations ( $\mathbf{o}_{in}^u, \mathbf{o}_{ex}^u$ ) from the user side, and generates item intrinsic and extrinsic factor representations ( $\mathbf{o}_{in}^v, \mathbf{o}_{ex}^v$ ) from the item side. On the user side, we first generate a user representation and a context representation based on user features and context features, respectively. Here, we use the SIGN model [30] to generate the representations (see Section 3.1 for details). SIGN has been proven effective in user/item/context representation learning through modeling feature interactions via graph neural networks. More formally, let  $f_u(u) : \mathbb{R}^{p \times d} \rightarrow \mathbb{R}^d$  be the function for SIGN-based feature modeling, where  $p$  is the number of user features.  $f_u(u)$  first maps each user feature  $z_i^u \in u$  into a  $d$ -dimensional feature embedding  $z_i^u$ . Then, it models these feature embeddings to output the user representation  $\mathbf{u}$ . Similarly, SIGN learns context representation  $\mathbf{c}$  through  $f_c$ . Next, a factor generation function  $f_{ie}^u(\mathbf{u}, \mathbf{c}) : \mathbb{R}^{2 \times d} \rightarrow \mathbb{R}^{2 \times d}$  (e.g., a neural network) takes the user representation and the context representation as input, and simultaneously generates a user intrinsic representation  $\mathbf{o}_{in}^u$  and a user extrinsic representations  $\mathbf{o}_{ex}^u$ . Here, the output is a  $2d$ -dimensional vector, with the first  $d$ -dimensional terms as  $\mathbf{o}_{in}^u$  and the

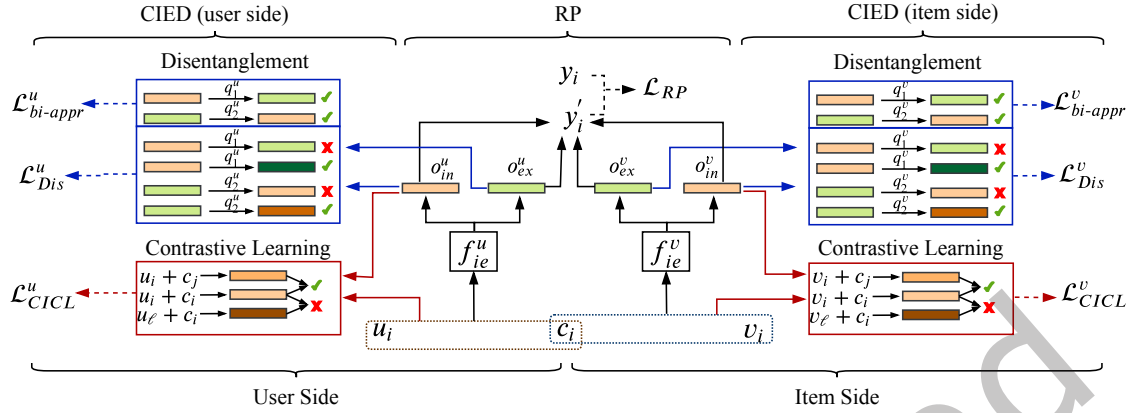


Fig. 2. An Overview of IEDR. It is a symmetric structure on the user side and the item side. The middle part (the black arrows) represents the recommendation prediction (RP) module (Section 5.1). It generates the intrinsic and extrinsic factor representations ( $\mathbf{o}_{in}$  and  $\mathbf{o}_{ex}$ ) for producing the recommendation prediction  $y'$ . The side parts are two contrastive intrinsic-extrinsic disentangling (CIED) modules. Each CIED includes a context-invariant contrastive learning component (the red arrows, Section 5.2.1), and a disentangling component (the blue arrows, Section 5.2.2) to ensure the success of the factor learning. The losses generated through these modules ( $\mathcal{L}_{RP}$ ,  $\mathcal{L}_{CICL}$ ,  $\mathcal{L}_{bi-app}$ ,  $\mathcal{L}_{Dis}$ ) will be optimized as a two-step multi-task training (Section 5.3.2).

rest as  $\mathbf{o}_{ex}^u$ . Note that without our CIED module (Section 5.2),  $\mathbf{o}_{in}^u$  and  $\mathbf{o}_{ex}^u$  are entangled. Currently, we name them  $\mathbf{o}_{in}^u$  and  $\mathbf{o}_{ex}^u$  to make it consistent with the following description. When equipped with CIED module,  $\mathbf{o}_{in}^u$  and  $\mathbf{o}_{ex}^u$  will be disentangled and represent intrinsic and extrinsic factors respectively. On the item side, a similar module structure is adopted. We use a different SIGN-based function for the item representation learning  $\mathbf{v} = f_v(v)$ , while using the same context representation as that on the user side. A factor-generating function  $f_{ie}^v(v, c)$  is applied to obtain the item intrinsic factor representation  $\mathbf{o}_{in}^v$  and extrinsic factor representation  $\mathbf{o}_{ex}^v$ .

Finally, we learn the prediction  $y' = f_{pred}(\mathbf{o}_{in}^u, \mathbf{o}_{ex}^u, \mathbf{o}_{in}^v, \mathbf{o}_{ex}^v)$ . We linearly combine the learned factors and use the dot product as the prediction function:  $f_{pred}(\mathbf{o}_{in}^u, \mathbf{o}_{ex}^u, \mathbf{o}_{in}^v, \mathbf{o}_{ex}^v) = (\mathbf{o}_{in}^u + \mathbf{o}_{ex}^u)^\top (\mathbf{o}_{in}^v + \mathbf{o}_{ex}^v)$ . A cross-entropy loss function is adopted to minimize the prediction error:  $\mathcal{L}_{RP}(u, v, c) := -y \log(y') + (1 - y) \log(1 - y')$ .

## 5.2 Contrastive Intrinsic-Extrinsic Disentangling (CIED) Module

The CIED module is designed to capture intrinsic and extrinsic factors from the representations generated by the RP module. The key idea is to integrate a context-invariant contrastive learning objective with a mutual information minimization scheme to simultaneously capture intrinsic factors that are stable across contexts and extrinsic factors that vary with different contextual conditions.

Specifically, CIED consists of two interrelated components: (1) a context-invariant contrastive learning component that encourages the model to learn intrinsic factors by contrasting user representations across different contexts, and (2) a bidirectional disentangling component that further separates the extrinsic factors from the learned intrinsic factors via a bidirectional mutual information minimization scheme. Next, we describe the two components in detail.

**5.2.1 Context-Invariant Contrastive Learning Component.** The context-invariant contrastive learning component is designed to learn intrinsic representations that are invariant across different contexts. The core idea is to maximize the agreement between the intrinsic representation pairs generated from the same user under different

contexts (positive pairs), while minimizing the agreement between those generated from different users under the same context (negative pairs). This contrastive objective encourages the model to capture the shared information across contexts as the intrinsic representation. More formally, we represent the intrinsic representations with the subscript  $(\mathbf{o}_{in}^u)_{ij}$  if it is generated through user  $u_i$  (from  $i$ -th data sample) and context  $c_j$  (from  $j$ -th data sample), i.e.,  $(\mathbf{o}_{in}^u)_{ij} = f_{ie}^u(\mathbf{u}_i, \mathbf{c}_j)$ . Inspired by InfoNCE [24], for the  $i$ -th data sample  $(u_i, v_i, c_i) \in \mathcal{D}$ , we calculate the objective function as follows:

$$\mathcal{L}_{\text{CICL}}^u(u_i, c_i) := -\log \frac{\exp(\text{sim}((\mathbf{o}_{in}^u)_{ii}, (\mathbf{o}_{in}^u)_{ij})/\tau)}{\sum_{u_t \in \mathcal{U}} \exp(\text{sim}((\mathbf{o}_{in}^u)_{ii}, (\mathbf{o}_{in}^u)_{ti})/\tau)}, \quad (2)$$

where  $(\mathbf{o}_{in}^u)_{ij}$  is generated from a user  $u_i$  and an arbitrary context  $c_j$ ,  $\text{sim}(\cdot)$  is the cosine similarity, and  $\tau$  is a temperature value.

The objective function is intuitive: one user should have the same intrinsic factor in different contexts, while different users can have their own personalized interests (different intrinsic factors).

**5.2.2 Disentangling Component.** To capture both the intrinsic and extrinsic factors, we need to disentangle extrinsic factors from intrinsic factors. The vCLUB method [8] can perform disentanglement through mutual information minimization. However, typical vCLUB is an asymmetric method, which may be less robust and lead to unsatisfactory disentanglement (detailed in Section 6.4). Therefore, we propose a bidirectional vCLUB approach that simultaneously minimizes the mutual information between intrinsic and extrinsic factors in both directions, leading to more robust and effective disentanglement.

In the bidirectional vCLUB, two variational distributions (e.g., approximated via neural networks)  $q_1^u(\mathbf{o}_{ex}^u | \mathbf{o}_{in}^u; \theta_1^u)$  and  $q_2^u(\mathbf{o}_{in}^u | \mathbf{o}_{ex}^u; \theta_2^u)$  are proposed with parameters  $\theta_1^u$  and  $\theta_2^u$ , to predict the two types of factors, respectively. Then a bidirectional vCLUB-based mutual information upper bound can be obtained as:<sup>1</sup>

$$\begin{aligned} \bar{\mathcal{I}}_{\text{bi-vCLUB}}(\mathbf{o}_{in}^u; \mathbf{o}_{ex}^u) := & \frac{1}{2} \left( \mathbb{E}_{p(\mathbf{o}_{in}^u, \mathbf{o}_{ex}^u)} [\log q_1^u(\mathbf{o}_{ex}^u | \mathbf{o}_{in}^u)] - \mathbb{E}_{p(\mathbf{o}_{in}^u)p(\mathbf{o}_{ex}^u)} [\log q_1^u(\mathbf{o}_{ex}^u | \mathbf{o}_{in}^u)] \right. \\ & \left. + \mathbb{E}_{p(\mathbf{o}_{in}^u, \mathbf{o}_{ex}^u)} [\log q_2^u(\mathbf{o}_{in}^u | \mathbf{o}_{ex}^u)] - \mathbb{E}_{p(\mathbf{o}_{ex}^u)p(\mathbf{o}_{in}^u)} [\log q_2^u(\mathbf{o}_{in}^u | \mathbf{o}_{ex}^u)] \right). \end{aligned} \quad (3)$$

By minimizing the upper bound  $\bar{\mathcal{I}}_{\text{bi-vCLUB}}(\mathbf{o}_{in}^u; \mathbf{o}_{ex}^u)$  as above, we minimize the mutual information between  $\mathbf{o}_{in}^u$  and  $\mathbf{o}_{ex}^u$ . Experimental results in Section 7.3.2 show that vCLUB is more robust and achieves better factor learning.

The optimization of the disentangling component is conducted in two iteratively steps. In the first step, we estimate the upper bound by training  $\theta_1^u$  and  $\theta_2^u$  to minimize the loss function  $\mathcal{L}_{\text{bi-appr}}^u(u_i, c_i) := -\frac{1}{2} \left( \log q_1^u((\mathbf{o}_{ex}^u)_{ii} | (\mathbf{o}_{in}^u)_{ii}) + \log q_2^u((\mathbf{o}_{in}^u)_{ii} | (\mathbf{o}_{ex}^u)_{ii}) \right)$ . Following [8], we use the mean squared error to optimize  $q_1^u$  and  $q_2^u$ . In the second step, we freeze  $\theta_1^u$  and  $\theta_2^u$ , and minimize the mutual information of  $\mathbf{o}_{in}^u$  and  $\mathbf{o}_{ex}^u$  by training other parameters to minimize the upper bound  $\mathcal{L}_{\text{Dis}}^u(u_i, c_i) = \bar{\mathcal{I}}_{\text{bi-vCLUB}}((\mathbf{o}_{in}^u)_{ii}; (\mathbf{o}_{ex}^u)_{ii})$ .

The context-invariant contrastive learning and disentanglement components in CIED are designed to work synergistically to learn meaningful intrinsic and extrinsic factors in the recommendation setting of various contexts. The contrastive learning component first learns context-invariant intrinsic factors by contrasting user representations across different contexts. These learned intrinsic factors then serve as a starting point for the disentanglement component to further separate the extrinsic factors via bidirectional mutual information minimization.

The seamless integration of these two components is crucial for the effectiveness of IEDR. By first learning context-invariant factors and then disentangling them from the extrinsic factors, CIED can effectively capture

<sup>1</sup>  $\bar{\mathcal{I}}_{\text{bi-vCLUB}}(\mathbf{o}_{in}^u; \mathbf{o}_{ex}^u)$  is the average of two vCLUB-based upper bounds of different directions. Therefore, it is obvious that  $\bar{\mathcal{I}}_{\text{bi-vCLUB}}(\mathbf{o}_{in}^u; \mathbf{o}_{ex}^u)$  is still an upper bound of  $\mathcal{I}(\mathbf{o}_{in}^u; \mathbf{o}_{ex}^u)$ .

the complex user behavior patterns influenced by various contextual conditions. Unlike existing methods, IEDR ensures context-agnostic learning of intrinsic and extrinsic factors in recommendations in scenarios of various contexts, and uniquely considers the interplay between these factors across various contexts, enhancing the model's effectiveness in complex, dynamic recommendation scenarios.

### 5.3 Implementation Details

**5.3.1 Iterative Optimization Procedure.** The CIED module is implemented as an iterative optimization procedure that alternates between the context-invariant contrastive learning and the disentanglement components.

In each iteration, the contrastive learning component first updates the model parameters to learn context-invariant intrinsic factors. Specifically, for each user  $u_i$  and context  $c_i$  in the current batch, we generate a positive pair  $(\mathbf{o}_{in}^u)_{ij}$  by either (1) randomly sampling a context  $c_j$  from the same batch, or (2) applying a high dropout rate to the original context representation  $c_i$ . We also generate  $L$  negative pairs  $(\mathbf{o}_{in}^u)_{li}$  by randomly sampling  $L$  users from the same batch. The contrastive loss  $\mathcal{L}_{\text{CICL}}^u(u_i, c_i)$  (Equation 2) is then computed and minimized to update the model parameters.

The learned intrinsic factors  $(\mathbf{o}_{in}^u)_{ii}$  are then fed into the disentangling component, which estimates and minimizes the mutual information between the intrinsic and extrinsic factors. We introduce two variational distributions  $q1^u(\mathbf{o}_{ex}^u | \mathbf{o}_{in}^u; \theta_1^u)$  and  $q2^u(\mathbf{o}_{in}^u | \mathbf{o}_{ex}^u; \theta_2^u)$ , parameterized by  $\theta_1^u$  and  $\theta_2^u$ , to estimate the bidirectional mutual information upper bound  $\mathcal{I}bi\text{-vCLUB}(\mathbf{o}_{in}^u; \mathbf{o}_{ex}^u)$  (Equation 3). The disentangling component is optimized in a two-step procedure: (1) estimating the mutual information upper bound by optimizing  $\theta_1^u$  and  $\theta_2^u$  to minimize the loss  $\mathcal{L}bi\text{-appr}^u(u_i, c_i)$ , and (2) minimizing the mutual information by optimizing the other parameters to minimize the upper bound  $\mathcal{L}Dis^u(u_i, c_i)$ .

The updated extrinsic factors  $(\mathbf{o}_{ex}^u)_{ii}$  are then used to refine the intrinsic factors in the next iteration of contrastive learning. This iterative process continues until convergence or a maximum number of iterations is reached.

**5.3.2 Multi-task Training.** We perform a two-step multi-task training to minimize the empirical risk of multiple components in IEDR. The two steps run alternatively until convergence. Appendix B provides the pseudo-code of the training procedure. In the first step, we freeze all the parameters except for  $\theta_1^u, \theta_2^u, \theta_1^v,$  and  $\theta_2^v$ , where  $\theta_1^v, \theta_2^v$  are the parameters of  $q1^v(\mathbf{o}_{ex}^v | \mathbf{o}_{in}^v; \theta_1^v)$  and  $q2^v(\mathbf{o}_{in}^v | \mathbf{o}_{ex}^v; \theta_2^v)$  in the disentangling component on the item side. We then minimize  $\mathcal{R}(\theta_1^u, \theta_2^u, \theta_1^v, \theta_2^v) = \frac{1}{N} \sum_{i=1}^N (\mathcal{L}_{bi\text{-appr}}^u(u_i, c_i) + \mathcal{L}_{bi\text{-appr}}^v(v_i, c_i))$ . In the second step, we freeze  $\theta_1^u, \theta_2^u, \theta_1^v,$  and  $\theta_2^v$ , and minimize the following function:

$$\arg \min \mathcal{R}(\omega) = \frac{1}{N} \sum_{i=1}^N \left( \mathcal{L}_{\text{RP}}(u_i, v_i, c_i) + \lambda_1 (\mathcal{L}_{\text{CICL}}^u(c_i, u_i) + \mathcal{L}_{\text{CICL}}^v(c_i, v_i)) + \lambda_2 (\mathcal{L}_{\text{Dis}}^u(u_i, c_i) + \mathcal{L}_{\text{Dis}}^v(v_i, c_i)) \right),$$

where  $\mathcal{L}_{bi\text{-appr}}^v, \mathcal{L}_{\text{CICL}}^v,$  and  $\mathcal{L}_{\text{Dis}}^v$  are the losses on the item side,  $\lambda_1$  and  $\lambda_2$  are the weight factors, and  $\omega$  are all the trainable parameters except for  $\theta_1^u, \theta_2^u, \theta_1^v,$  and  $\theta_2^v$ .

The multi-task training procedure ensures that the model learns to accurately predict recommendations while simultaneously learning disentangled intrinsic and extrinsic factors. The contrastive learning and disentanglement losses are integrated into the overall training objective, allowing the model to capture the complex user behavior patterns influenced by various contextual conditions.

## 6 DISCUSSION

In this section, we provide theoretical and practical discussions of IEDR from multiple perspectives, including the information theory foundation, time complexity analysis, trivial solution prevention, and potential problems of the vCLUB method used in the disentanglement component.

### 6.1 Theoretical Analysis: Context-invariant Contrastive Learning in Information Theory

In this section, we reason the context-invariant contrastive learning from the perspective of information theory. As formally defined in Theorem 1, optimizing Equation (2) is equivalent to maximizing the mutual information between the intrinsic representations and user representations, and simultaneously minimizing the mutual information between the intrinsic representations and the context representations. The theorem on the item side can be derived in the same fashion. The proof of this equivalence can be found in Appendix A.

**THEOREM 1 (EQUIVALENCE OF CONTRASTIVE LOSS  $\mathcal{L}_{CICL}^u$ ).** *Optimizing the contrastive loss is equivalent to solving:*

$$\operatorname{argmin} \sum_{i=1}^N \mathcal{L}_{CICL}^u(u_i, c_i) = \operatorname{argmax} \left( I(\mathbf{o}_{in}^u, \mathbf{u}) - I(\mathbf{o}_{in}^u, \mathbf{c}) \right). \quad (4)$$

Theorem 1 provides the perspective from information theory to understand the context-invariant contrastive learning procedure: the information of users that is not influenced by contexts (i.e., intrinsic factors) is kept in  $\mathbf{o}_{in}^u$ .

### 6.2 Time Complexity Analysis

The time complexity of IEDR is comparable to feature interaction-based recommender systems (e.g., AutoInt [28], SIGN [30]). The overhead of the alternative optimizing procedure for the disentanglement component is marginal in the whole optimizing procedure.

Specifically, the most time-consuming computations are the feature interaction learning to get user, item, and context representations, which need to conduct interaction modeling on every pair of feature interactions. This procedure has also been done on other feature interaction-based models. Therefore, the time complexity of the proposed module is comparable with those methods.

Our model takes additional computations on the contrastive learning component (CICL) and the disentangling component: (1) For the CICL component, we do not need to perform the feature interaction modeling again, but reuse the generated user, item and context representations, which saves the majority of the overhead. We only need to perform  $f_{ie} L + 1$  times, where  $L$  is the number of negative samples and  $f_{ie}$  is a one-hidden layer MLP. (2) For the disentangling component, we reuse the generated user/item/context representations as well. The first step in the two-step learning takes very little overhead. This is because this step only tries to optimize the parameters of the functions  $q_1$  and  $q_2$  (Equation (3)), which are two MLPs with one hidden layer. For each data sample, we only run  $q_1$  and  $q_2$  once using  $\mathbf{o}_{in}$  and  $\mathbf{o}_{ex}$ .

In summary, since all of the computations above do not need to perform feature interaction modeling (the most time-consuming procedure in all feature interaction-based models), the small imposed overhead is acceptable considering the effectiveness of our model in capturing accurate intrinsic and extrinsic factors. More empirical analysis can be found in Section 7.8.

### 6.3 Preventing the Trivial Solution of CIED

The two components in the CIED module, the contrastive learning component and the disentangling component, jointly ensure the success of the intrinsic and extrinsic factor representation learning. However, CIED may fall into a trivial solution:  $f_{ie}^u(\mathbf{u}, \mathbf{c})$  maps  $\mathbf{u}$  to  $\mathbf{o}_{in}^u$  without considering  $\mathbf{c}$ , and maps  $\mathbf{c}$  to  $\mathbf{o}_{ex}^u$  without considering  $\mathbf{u}$ . Although this trivial solution minimizes  $\mathcal{L}_{CICL}(u, c)$  and  $\mathcal{L}_{Dis}(u, c)$ ,  $\mathbf{o}_{in}^u$  (resp.  $\mathbf{o}_{ex}^u$ ) is not the intrinsic (resp. extrinsic) factor, but just a mapping of the user information (resp. context information). We prove that this trivial solution can be avoided by setting  $f_{ie}^u(\mathbf{u}, \mathbf{c})$  as a *non-linear* function, leading  $\mathbf{u}$  and  $\mathbf{c}$  to statistically interact.

**6.3.1 Statistical Interaction.** We first introduce the statistical interaction (or non-additive interaction), which ensures a joint influence of several variables on an output variable is not additive [34]. Based on [29],  $F(X)$  shows statistical interaction between variables  $x_i$  and  $x_j$  if  $\forall f_i, f_j, F(X)$  **cannot** be expressed as:

$$F(\mathbf{X}) \neq f_i(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) + f_j(x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_n). \quad (5)$$

More generally, if using  $\mathbf{v}_i \in \mathbb{R}^d$  to describe the  $i$ -th variable with a  $d$ -dimension vector [25, 30], e.g., variable embedding, each variable can be described in a vector form  $\mathbf{u}_i = x_i \mathbf{v}_i$ . Then, we define the pairwise statistical interaction in vector form by changing the Equation (5) into:

$$F(\mathbf{X}) \neq f_i(\mathbf{u}_1, \dots, \mathbf{u}_{i-1}, \mathbf{u}_{i+1}, \dots, \mathbf{u}_n) + f_j(\mathbf{u}_1, \dots, \mathbf{u}_{j-1}, \mathbf{u}_{j+1}, \dots, \mathbf{u}_n).$$

**6.3.2 Preventing the Trivial Solution.** Based on the definition of statistical interaction, we can express the trivial solution as that  $f_{ie}^u(\mathbf{u}, \mathbf{c})$  learns no statistical interaction between  $\mathbf{u}$  and  $\mathbf{c}$ :

$$f_{ie}^u(\mathbf{u}, \mathbf{c}) = \lambda_1 f_1(\mathbf{u}) + \lambda_2 f_2(\mathbf{c}), \quad (6)$$

where  $f_1$  outputs  $\mathbf{o}_{in}^u$ ,  $f_2$  outputs  $\mathbf{o}_{ex}^u$ , and  $\lambda$  are weight scalars.

To prevent the trivial solution, we need to ensure that function  $f_{ie}^u(\mathbf{u}, \mathbf{c})$  cannot be modeled in the form of Equation (6). Therefore, if  $\mathbf{u}$  and  $\mathbf{c}$  are modeled as a statistical interaction in  $f_{ie}^u(\mathbf{u}, \mathbf{c})$ , the trivial solution can be prevented. Since  $f_{ie}^u(\mathbf{u}, \mathbf{c})$  only takes  $\mathbf{u}$  and  $\mathbf{c}$  as inputs, we just need  $f_{ie}^u$  to be a non-additive model. That is,  $f_{ie}^u(\mathbf{u}, \mathbf{c})$  should contain a third term  $f_3(\mathbf{u}, \mathbf{c})$ :

$$f_{ie}^u(\mathbf{u}, \mathbf{c}) = \lambda_1 f_1(\mathbf{u}) + \lambda_2 f_2(\mathbf{c}) + \lambda_3 f_3(\mathbf{u}, \mathbf{c}),$$

where  $f_3$  is a non-additive model and  $\lambda_3 \neq 0$ .

Therefore, in the optimized situation,  $\mathbf{o}_{in}^u = \lambda_1 f_1(\mathbf{u})$  learns part of the information from users that do not interact with context information.  $\mathbf{o}_{ex}^u = \lambda_2 f_2(\mathbf{c}) + \lambda_3 f_3(\mathbf{u}, \mathbf{c})$  learns the context information ( $f_2(\mathbf{c})$ ) and the information that changes given different contexts ( $f_3(\mathbf{u}, \mathbf{c})$ ).

In Section 7.9, we empirically analyze how the trivial solution will influence the prediction performance.

## 6.4 Potential Problems of the Asymmetric vCLUB Method

The vCLUB-based mutual information minimization method proposed in [8] is an asymmetric method. In this section, we explain the possible reason that vCLUB is less robust and performs worse than our proposed bidirectional vCLUB method (*BiDis*).

Directly applying vCLUB leads to the parameter  $\theta_1^u$  of a variational distribution  $q_1^u(\mathbf{o}_{ex}^u | \mathbf{o}_{in}^u; \theta_1^u)$  being trained to approach the vCLUB-based upper bound in Equation (1) (Step 1). Then,  $\theta_1^u$  is frozen, and  $\mathbf{o}_{ex}^u, \mathbf{o}_{in}^u$  are trained to minimize  $\mathcal{I}(\mathbf{o}_{in}^u; \mathbf{o}_{ex}^u)$  via minimizing the upper bound  $\mathcal{I}_{vCLUB}(\mathbf{o}_{in}^u; \mathbf{o}_{ex}^u)$  (Step 2). However, this way of minimizing mutual information may result in an unexpected outcome: the mutual information may be minimized via making  $\mathbf{o}_{in}^u$  contain as little information as possible. To better illustrate the possible outcome, we design  $q_1^u$  as a linear function which is well trained in Step 1 to ensure Equation (1) is an upper bound of  $\mathcal{I}(\mathbf{o}_{in}^u; \mathbf{o}_{ex}^u)$ . Figure 3 shows how the unexpected result may occur. In Step 2,  $\mathbf{o}_{ex}^u, \mathbf{o}_{in}^u$  will be trained to minimize Equation (1). To achieve this goal, it ensures  $q_1^u$  cannot predict  $\mathbf{o}_{ex}^u$  given the corresponding  $\mathbf{o}_{in}^u$  from the joint distribution (the first term of Equation (1)), and at the same time ensures the output of  $q_1^u$  is similar to the other  $\mathbf{o}_{ex}^u$ 's from the marginal distribution (the second term of Equation (1)).

From  $\mathbf{o}_{in}^u$  perspective (blue circles), the goal can be achieved by pushing the  $\mathbf{o}_{in}^u$  to move from its original position (optimizing the first term of Equation (1)), and move towards the mean of the other  $\mathbf{o}_{in}^u$ 's (optimizing the second term of Equation (1)). From  $\mathbf{o}_{ex}^u$  perspective (red circles), the goal can be achieved by pushing the  $\mathbf{o}_{ex}^u$  away from its original position (optimizing the first term of Equation (1)) and the mean of the other  $\mathbf{o}_{ex}^u$ 's (optimizing the second term of Equation (1)).

This clusters all the  $\mathbf{o}_{in}^u$ 's together, making  $\mathbf{o}_{in}^u$ 's contain less information, while all the  $\mathbf{o}_{ex}^u$ 's try to split away from each other, making  $\mathbf{o}_{ex}^u$ 's contain more information. The mutual information minimization procedure is like

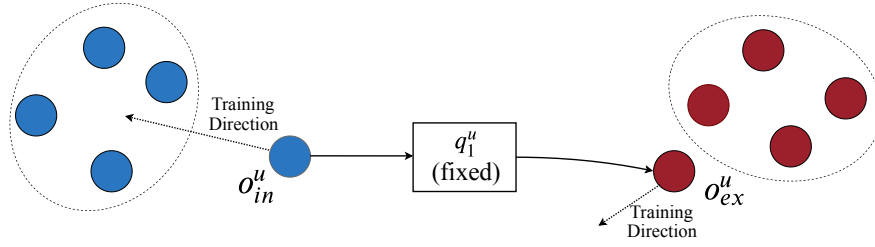


Fig. 3. An illustrative example demonstrating the potential problem of asymmetric learning in vCLUB. The blue circles are intrinsic representations, and the red circles are extrinsic representations. The dotted arrows are the directions that vCLUB will push  $\mathbf{o}_{in}^u$  and  $\mathbf{o}_{ex}^u$  to move toward their space.

“transferring” the information from  $\mathbf{o}_{in}^u$ ’s to  $\mathbf{o}_{ex}^u$ ’s, which is not what we expect. *BiDis*, however, is a symmetric disentangling method on  $\mathbf{o}_{in}^u$ ’s and  $\mathbf{o}_{ex}^u$ ’s that does not result in this issue. This may be why vCLUB performs worse and is less robust than our proposed symmetrical disentangling component.

## 7 EXPERIMENTS

We conduct extensive experiments to demonstrate the effectiveness of our model. In this section, we focus on 1) the recommendation performance of IEDR compared to the state-of-the-art methods; 2) the effectiveness of each component in IEDR; and 3) the ability to disentangle intrinsic and extrinsic factors of IEDR.

### 7.1 Experimental Setting

This section demonstrates the detailed experimental setting to evaluate our method, including the datasets, the baseline methods, and the implementation details.

**7.1.1 Datasets.** We evaluate our models in two scenarios with various contexts: a mobile app recommendation and a restaurant recommendation. In the mobile app recommendation, we use the Frappe [1] dataset that records mobile app usage logs. Each data sample logs users’ app usage in a certain context (e.g., weather, time, location). In the restaurant recommendation, we use the Yelp dataset [43]. Each data sample records users’ reviews of local restaurants. Due to the fact that a user usually goes to restaurants in the same city, geographic isolation appears in the dataset. Therefore, we select the records in New York City. We regard each record as a data sample that the user has been to the restaurant. We leverage the user/item features and context features (e.g., day of the week) to predict whether a user will go to a given restaurant in a specific context. We also evaluate our model on two Amazon datasets (Movies and CDs) [22], which have been used in sequential recommendation tasks [53]. The datasets contain user-item interactions with timestamps. For the sequential recommendation, we use the same IEDR model structure as that for the Frappe and Yelp datasets, but modify the data input to fit our model. More specifically, we do not directly learn behavior sequences, but consider each behavior as a data sample with time context information. That is, we consider the bucketed timestamp of each user behavior as a time context (we consider one month as a categorized time context). Therefore, behaviors in the same time interval have the same time context, indicating that these behaviors share some similar short-term (extrinsic) interests (e.g., item popularity). Note that our experiments are to evaluate our key motivation: learning better intrinsic/extrinsic factor representations. Therefore, our chosen four datasets have high-quality user feedback (e.g., review/comment-based), which is more suitable than other datasets that are larger but less accurate (e.g., click-through-based).

For each dataset, the users that have more than 5 records (Frappe and Yelp) or more than 20 records (Movies and CDs) are chosen. We use the last and the second last record of each user for testing validation, respectively. The rest are for training. Each of these data samples is considered a positive sample ( $y = 1$ ). For each positive data sample in the training set, we randomly sample 2 items (but keep the user and contexts) as negative samples ( $y = 0$ ), meaning the user did not select the 2 items in that context. For each test/validation data sample, we randomly choose 99 items as negative samples to ensure a more robust evaluation. The statistics of the datasets are shown in Table 2.

Table 2. Dataset statistics. “Count” refers to the number of users/items, and “Features” represents the number of different features (for User and Item, the number of features excludes the user/item IDs).

Datasets	Data Samples			User		Item		Context
	Train	Valid	Test	Count	Features	Count	Features	Features
Frappe	282,426	69,500	69,500	695	0	4,082	2,892	318
Yelp	518,208	633,600	633,600	6,336	24	12,902	66	13,034
Movies	2,305,362	39,663	1,322,100	13,221	0	49,189	161	193
CDs	879,030	16,392	546,400	5,464	0	16,184	209	195

**7.1.2 Baseline methods.** IEDR models the feature interactions of users, items, and contexts. Therefore, we compare our model with competitive feature interaction-based recommendation methods. The methods include attentional factorization machine (AFM) [46], neural factorization machine (NFM) [12], self-attention-based feature interaction model (AutoInt) [28], deep factorization machine (DeepFM) [11], wide & deep model (WDL) [7], improved deep & cross network (DCNv2) [40], input-aware factorization machine (IFM) [52], model-agnostic contrastive learning for CTR (CL4CTR) [37], and adaptive learning via Euler’s formula (EulerNet) [33]. We implement these methods using the DeepCTR package or their officially released code. The above methods model all the factors in a unified representation without considering the factors that affect user behavior.

Meanwhile, we compare IEDR with the methods that learn implicit factors. They are disentangled variational auto-encoder for recommendation (DisRec) [21] and disentangled graph collaborative filtering (DGCF) [42]. We implement these methods based on their released code. Note that since DisRec and DGCF models do not consider any feature, their task is to simply predict whether a user will select an item. IEDR and other baseline models, however, consider the user-item interactions in specific contexts (a user’s behavior in selecting an item may be different in different contexts). For DisRec and DGCF, to prevent the test data samples from appearing in the training set, we remove the data samples from the training set that appear in the test set (with different contexts in other models). For a fair comparison, we set the factor number to 4 for DisRec and DGCF. For sequential recommendation baselines, we compare our model with the models that consider LS-term interests. They are session-based recommender systems with recurrent neural networks (GRU4Rec) [13], Short-term and Long-term preference Integrated Recommender system (SLI-Rec) [53], and Contrastive learning framework of Long and Short-term interests for Recommendation (CLSR) [56]. We use the same MLP structure for feature interaction modeling and the same embedding size for features as our IEDR model.

**7.1.3 Implementation details.** In IEDR, all the MLPs have the same hidden structure: one hidden layer of 128 dimensions and a ReLU activation after that. The input and output sizes of MLPs vary based on their needs. We set the embedding dimension to 32 for all the features.  $f_{ie}$  is an MLP that outputs a 64-dimension vector, with the first 32 dimensions being the intrinsic factor representation and the last 32 dimensions being the extrinsic factor

Table 3. Comparing the prediction performance (in percentage) with the baselines. The best-performing results are in bold and the second best are underlined. The *Improv* and *p-value* rows show the relative improvements and the statistical significance of IEDR over the best-performed baselines, respectively.

	Frappe					Yelp				
	NDCG@5	NDCG@10	Recall@5	Recall@10	AUC	NDCG@5	NDCG@10	Recall@5	Recall@10	AUC
AFM	63.52	67.44	77.84	84.71	93.18	42.79	47.17	58.69	72.21	91.96
NFM	68.30	70.73	83.00	<u>90.40</u>	95.86	45.99	50.33	61.90	75.27	93.32
AutoInt	<u>69.45</u>	71.41	<u>84.04</u>	90.10	95.83	46.61	50.80	<u>63.72</u>	<u>76.55</u>	93.82
DeepFM	69.20	71.28	82.70	89.50	<u>96.09</u>	44.20	48.50	60.26	<u>73.55</u>	93.26
WDL	68.02	70.33	81.70	88.90	95.96	45.47	49.71	61.90	74.89	93.41
DCNv2	68.15	70.34	82.15	89.91	95.25	43.41	48.26	60.97	74.88	93.66
CL4CTR	68.36	70.51	82.23	89.82	95.48	45.05	49.80	63.24	76.29	93.54
EulerNet	68.87	70.68	83.30	90.36	95.88	44.81	49.54	63.33	76.08	93.47
IFM	66.91	69.13	80.90	87.60	95.32	46.74	50.86	63.04	75.69	<u>93.83</u>
SIGN	69.38	<u>71.49</u>	83.91	90.37	95.92	<u>46.80</u>	<u>50.94</u>	63.68	76.41	93.67
DisRec	56.81	60.07	67.42	76.29	85.51	34.82	37.90	48.29	63.17	84.01
DGCF	58.40	61.44	69.05	77.53	86.13	36.35	39.06	50.05	64.62	85.29
IEDR	<b>72.40</b>	<b>74.11</b>	<b>85.94</b>	<b>91.25</b>	<b>96.34</b>	<b>48.68</b>	<b>53.05</b>	<b>65.23</b>	<b>78.29</b>	<b>94.22</b>
<i>Improv</i>	4.24%	3.66%	2.26%	0.94%	0.26%	4.01%	4.14%	2.38%	2.28%	0.42%
<i>p-value</i>	0.25%	0.25%	0.25%	0.83%	3.72%	0.25%	0.25%	0.25%	0.25%	2.34%

representation. For the second (dropout-based) negative context-generating method in the context-invariant contrastive learning component, the dropout rate is set to 0.5. The number of negative pairs for contrastive learning is 40 for each data sample (note that the actual negative pairs will be doubled since both  $(\mathbf{o}_{in}^u)_{ii}$  and  $(\mathbf{o}_{in}^u)_{ij}$  will generate 40 negative pairs). The temperature  $\tau$  is set to 0.5. In the disentangling component,  $q_1$  and  $q_2$  are MLPs that output vectors that have the same dimension of intrinsic/extrinsic factor representations. The number of negative samples of the bidirectional vCLUB-based method is 5 for each direction. We set  $\lambda_1$  to 0.1 for the Frappe dataset and 0.01 for the Yelp dataset, and set  $\lambda_2$  to 0.1 for both datasets. The  $\lambda_1$  and  $\lambda_2$  are both 0.01 for the Movies and the CDs datasets.

The model structure of IEDR and its variations used in the experiments are detailed in Table 10 and Table 11. Note that the component structures of variations are the same as the IEDR if not specified.

## 7.2 Overall Performance

We evaluate the recommendation performance of our model, by comparing it with various baselines in two scenarios. In the first scenario, we learn intrinsic and extrinsic factors from various contexts. In the second scenario, we learn the factors from a specific (time) context and compare our model with sequential recommendation baselines. We use three common evaluation metrics for recommender systems: NDCG@ $k$ , Recall@ $k$  with  $k$  being 5 and 10, and AUC.

**7.2.1 Factor Learning from Specific Context.** We then evaluate IEDR on two Amazon datasets (Movies and CDs) [22] that contain only the time context. We compare with the state-of-the-art sequential recommendation baselines GRU4Rec [13], LSI-Rec [53], CLSR [56] and AutoMLP [18], that learn long-short term interests from the item sequences ordered by the time (discussed in Section 2). Also, we compare with state-of-the-art general sequential recommendation baselines, BERT4Rec [32], SASRec [15], S3Rec [57], TiSASRec [38]. In IEDR, we use the same model structure as that for the Frappe and Yelp datasets, but modify the data input to fit our model. More specifically, without directly learning behavior sequences, IEDR considers each behavior as a data sample

Table 4. Comparing the performance of IEDR and the baselines on time context-specific scenarios.

	Movies		CDs	
	NDCG@10	AUC	NDCG@10	AUC
GRU4Rec	25.18	77.11	19.41	78.86
SLI-Rec	26.85	78.69	20.27	79.37
CLSR	26.98	80.02	21.07	80.42
AutoMLP	26.73	79.91	20.32	79.60
BERT4Rec	25.27	78.24	19.53	79.13
SASRec	26.28	79.49	20.67	79.71
S3Rec	<b>27.04</b>	80.11	21.16	80.09
TiSASRec	26.84	79.93	<b>21.25</b>	80.18
AutoInt	22.27	77.78	18.25	77.65
DeepFM	23.13	78.50	19.18	78.26
SIGN	23.58	78.82	19.97	78.95
IEDR	26.68	<b>80.14</b>	20.95	<b>80.34</b>

with time context information, where the time context is the bucketed timestamp of each user behavior (one month as a categorized time context). We also run the best-performing baselines from Table 3 on the Amazon datasets. The experimental results are reported in Table 4.

From these results, we can see that our model achieves competitive accuracy compared to the sequential recommendation baselines. This proves the ability of our model to achieve state-of-the-art recommendation accuracy in the context-specific scenario, even compared with the models designed for the context. Moreover, our IEDR is more versatile and can be applied to various contexts. Finally, the feature interaction-based baselines do not disentangle intrinsic and extrinsic factors. Therefore, they perform worse than our models and sequential recommendation baselines on the Amazon datasets.

### 7.3 Effectiveness of Our Model’s Components

This section evaluates the components of IEDR. We only demonstrate the results in NDCG@10 since metrics show similar trends.

**7.3.1 Ablation Study of Contrastive Intrinsic-Extrinsic Disentangling Module.** To evaluate the contribution of the Contrastive Intrinsic-Extrinsic Disentangling (CIED) module, we compare IEDR against three variants: *noDis* (removes the disentanglement component), *noCL* (removes the context-invariant contrastive learning component), and *noCIED* (removes both components). The experiments are conducted on the Frappe and Yelp datasets, and the results are presented in Figure 4. The results highlight the synergistic contribution of the two components in IEDR. 1) *IEDR* achieves the best performance on both datasets (74.11 on Frappe and 53.05 on Yelp), with improvements over *noCIED* of 4.06 points on Frappe and 2.99 points on Yelp, exceeding the combined individual improvements of *noDis* and *noCL*. This indicates a cumulative effect, where the disentanglement component and CIED reinforce each other, ensuring stable intrinsic factors and effective separation of extrinsic factors. 2) The small improvement of *noCL* over *noCIED* on Frappe (0.16 points) highlights the limitations of relying solely on implicit factor disentanglement, particularly in datasets dominated by context features. These findings emphasize the importance of explicit factor learning through CIED, which ensures robust disentanglement and overall performance gains.

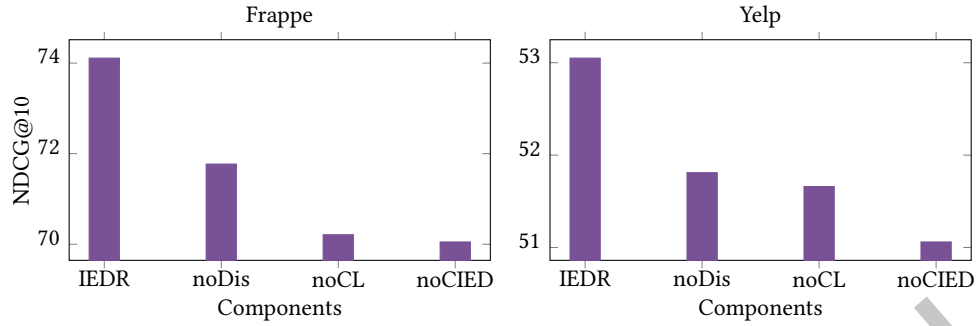


Fig. 4. Ablation studies results with different component(s) removed.

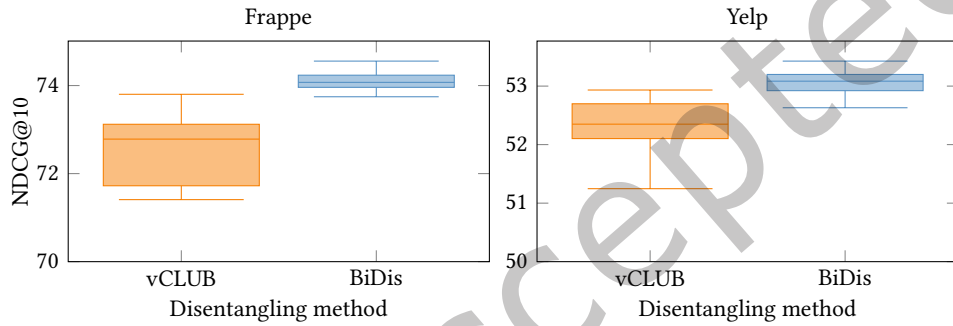


Fig. 5. The performance and variance statistics of vCLUB and BiDis.

**7.3.2 Disentangling Component Evaluation.** We propose a bidirectional vCLUB-based disentangling method (*BiDis*) to disentangle the intrinsic and extrinsic factors. In this section, we compare our *BiDis* method with the original vCLUB method (*vCLUB*) [8] in model performance. The results in Figure 5 highlight the superiority of our *BiDis* method over *vCLUB* in both performance and robustness. *BiDis* leverages bidirectional mutual information minimization, ensuring a more thorough and balanced disentanglement of intrinsic and extrinsic factors, as discussed in Section 5.2.2. This bidirectional approach avoids the instability and noise issues associated with *vCLUB*'s asymmetric optimization, resulting in more robust and consistent performance across datasets. Additionally, the visualization in Section 7.5.1 further demonstrates that *BiDis* produces clearer and more distinct factor separation, underscoring its effectiveness in real-world recommendation scenarios.

**7.3.3 Other Feature Modeling Methods.** In the RP module, although we use a SIGN-based method [30] to learn user, item, and context features, the module can use any feature modeling method. Here, we use other methods to evaluate whether our model still performs well. Specifically, we run our model with the other three variations using different feature modeling methods: 1) averaging feature embeddings (*MEAN*); 2) adding an MLP on top of the averaged feature embedding (*MLP*); and 3) modeling and aggregating feature interactions through a Bi-interaction layer proposed in [12] (*BI*). The results are shown in Figure 6. We report the results of each variation with and without the CIED module. From this figure, we can see that when equipped with the CIED module, all feature modeling methods perform better than those without the module. It shows that our proposed CIED module can learn intrinsic and extrinsic factors for more accurate recommendations when different feature

Table 5. Comparing the performance of IEDR<sub>sp</sub> with different dropout rates (for *NegGen2*).

	Frappe	Yelp
IEDR <sub>sp</sub> , p=0.1	70.68	52.03
IEDR <sub>sp</sub> , p=0.5	68.25	51.49
IEDR <sub>sp</sub> , p=0.1, noDis	70.56	52.02
IEDR <sub>sp</sub> , p=0.1, noCL	70.31	51.62
IEDR <sub>sp</sub> , p=0.1, noCIED	70.16	51.10
<b>IEDR</b>	<b>74.11</b>	<b>53.05</b>

modeling methods are applied. Meanwhile, the feature modeling methods can impact the performance. *MEAN* is just a linear aggregation of features, resulting in the worst performance. Both *MLP* and *BI* have better feature modeling ability and hence have better performance than *MEAN*. The *SIGN*-based feature modeling (*SIGN*) is the state-of-the-art feature interaction modeling method and archives the best performance.

#### 7.4 Comparing the Impact of Different Contrastive Learning Variations

To learn intrinsic factors, we propose a context-invariant contrastive learning method. However, directly generating intrinsic factor representations through user information seems to be a more direct way, i.e.,  $\mathbf{o}_{in}^u = f_{ie}^u(\mathbf{u})$ . However, we argue that the intrinsic factors learned this way could not guarantee the effectiveness of intrinsic factor learning. This is because the information in the learned intrinsic factor representations can vary with different contexts, since these factors have never been modeled w.r.t. the contexts.

In this section, we empirically evaluate our argument and show that our context-invariant contrastive learning method generates more accurate recommendations. To do so, we design a variation (IEDR<sub>sp</sub>) by splitting the intrinsic-extrinsic factor generation into two functions:  $\mathbf{o}_{in}^u = f_{in}^u(\mathbf{u})$ , and  $\mathbf{o}_{ex}^u = f_{ex}^u(\mathbf{u}, \mathbf{c})$ . Both  $f_{in}$  and  $f_{ex}$  have the same structure as  $f_{ie}$ , with the output dimension being a half to ensure the consistency of the factor representation dimension. The contrastive learning component does not consider context information but uses a standard InfoNCE-based contrastive learning for learning robust user/item representations following [48]. Table 5 illustrates the results of IEDR<sub>sp</sub> compared to our model with IEDR<sub>sp</sub> using different dropout rates ( $p = 0.1$  and  $p = 0.5$ ) in the contrastive learning component, and different component combinations (*noDis*, *noCL*, *noCIED*). The experiment demonstrates that our model outperforms the variation in recommendation accuracy. This proves that

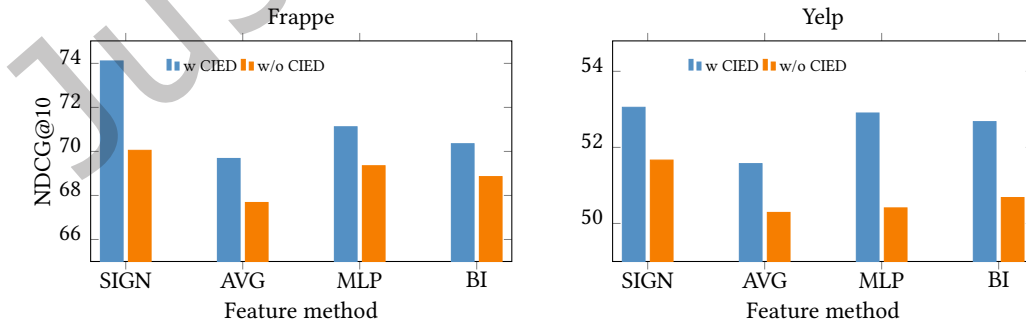


Fig. 6. Model performance when equipped with different feature modeling methods.

$IEDR_{sp}$  cannot ensure successful intrinsic factor learning and hence incurs a worse recommendation accuracy. Unlike  $IEDR$ ,  $IEDR_{sp}$  gains better performance with a lower dropout rate. This is because, in  $IEDR_{sp}$ , the dropout generates views representing the same user instead of different users, which is consistent with the conclusion in [10].

## 7.5 Disentanglement Verification

This section verifies the intrinsic and extrinsic factor disentangling ability of  $IEDR$ , including a visualization of the learned intrinsic and extrinsic representations and a case study to show the differences between these factors in users' decision-making.

**7.5.1 Intrinsic and Extrinsic Representation Visualization.** This section provides intrinsic and extrinsic representation visualizations of our model and three variations: 1) the contrastive learning component is removed (*noCL*); 2) the disentangling component is removed (*noDis*); and 3) the asymmetric disentanglement method (*vCLUB*) is used. Figure 7 compares these results. We include our main observations below:

- The intrinsic and extrinsic factors are perfectly disentangled with our CIED module ( $IEDR$ ).
- Without the disentangling component (*noDis*), the intrinsic and extrinsic disentangling procedure may not succeed. This is because there is no restriction on extrinsic representations. Therefore, the extrinsic representations can contain any information, including the information of the intrinsic factor.
- *noCL* has worse disentangling performance than  $IEDR$ , either. This is because the factors disentangled in *noCL* are implicit. The implicit factors only ensure the disentanglement between the factors of the same data sample, but not between the factors of other data samples. For example, some context information may be stored in the intrinsic representation in data sample 1 but be stored in the extrinsic representation in data sample 2.
- *noCIED* performs worst among all variations, which is reasonable since it does not distinguish the intrinsic and extrinsic representations.
- *vCLUB* performs disentanglement, but is not very stable in some situations. This is consistent with our analysis in Section 6.4.

**7.5.2 Case Study.** We conducted a case study to analyze the differences between the learned intrinsic and extrinsic factors. We randomly choose a user from the Frappe dataset and generate the intrinsic matching scores (the dot product of the user's intrinsic representation and the items' (apps) intrinsic representations) in two different contexts (Weekday and Weekend). The same for the extrinsic matching scores. We sort the matching scores for the intrinsic and extrinsic factors, respectively, and list the top 100 items. The results are in Figure 8. Note that the top 100 items for intrinsic and extrinsic factors are different. According to Figure 8, from weekday to weekend, the extrinsic scores vary a lot, while the intrinsic scores remain invariant. These observations demonstrate that, in different contexts, the user has different intrinsic factors, as well as consistent intrinsic factors.

Then, we show how intrinsic and extrinsic factors may have different impacts on users' choices. Table 6 lists the categories of the items with the 10 highest intrinsic/extrinsic scores for two users, respectively. we can observe that users have individual intrinsic interests that indicate their real hobbies, e.g., *User1* prefers sports and fitness apps, while *User2* prefers gaming apps. On the other hand, extrinsic factors give a higher rank to the items based on the contexts (Weekday), e.g., Tools (Google Search) and Communication (Gmail) rank highest in *User1*'s extrinsic scores.

## 7.6 Different Negative Context Generation Methods

We propose two negative context-generating methods in the contrastive learning component: 1) sample other contexts; 2) use a large dropout rate on the original context. We evaluate the two methods in this section. Table

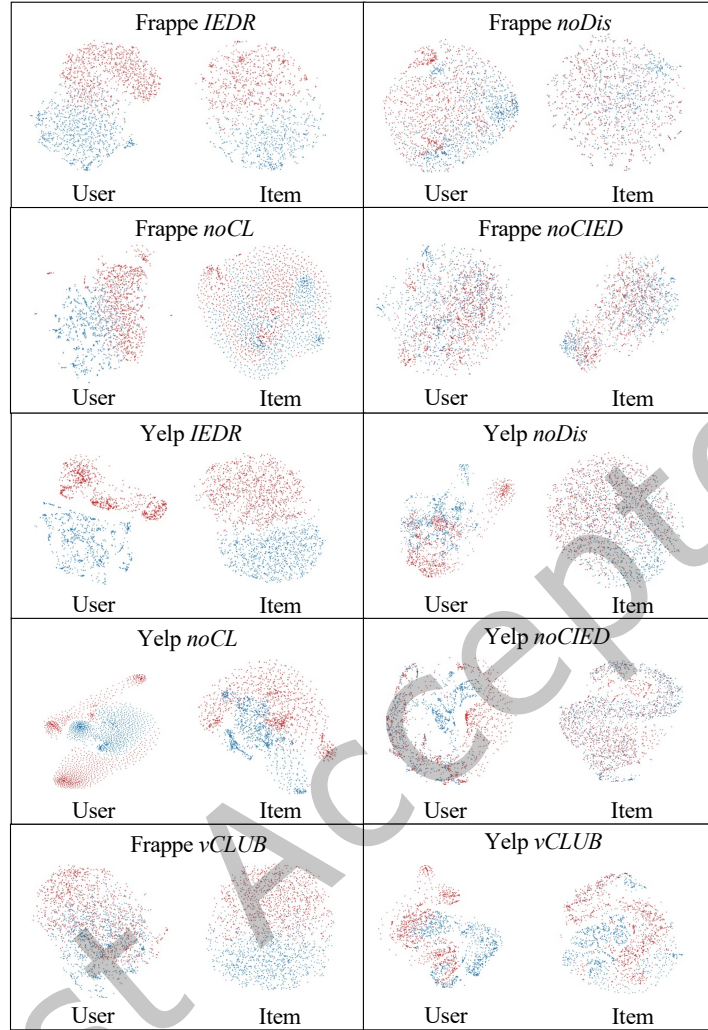


Fig. 7. The complete intrinsic-extrinsic disentanglement visualizations in t-SNE. The blue dots are intrinsic representations, and the red dots are extrinsic representations.

7 shows the results of our model when using only *NegGens1*, only *NegGens2*, and *NegGen1&2*. We can see that *NegGen1* results in a better performance than using *NegGen2*. This is because *NegGen1* uses true context representations, which are consistent with what may appear in the test samples. Meanwhile, we see that *NegGen1&2* results in the best performance. This is because *NegGen2* provides more unseen (randomly generated) context representations, which strengthens the generalization ability of our model. Next, we evaluate *NegGen2* with different dropout rates in Figure 9. The best performance can be achieved when the dropout rates range from 0.5 to 0.7. This is consistent with our claim in Section 5.2.1. The reason is that a small dropout rate (e.g., 0.1) pushes the generated context representation too close to the original one; hence it cannot be considered a different context. However, a relatively large dropout rate (e.g., 0.9) loses too much information; hence, it is no

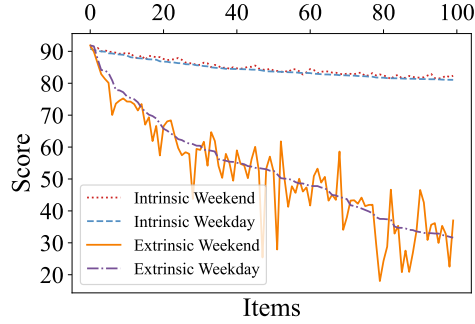


Fig. 8. A user’s top 100 intrinsic and extrinsic scores in different contexts (Weekend vs. Weekday).

Table 6. Items (in category) of the highest intrinsic and extrinsic scores for different users in Weekday.

Rank	User1		User2	
	Intrinsic	Extrinsic	Intrinsic	Extrinsic
1	Photography	Tools	Cards&Casino	Communication
2	Sports	Communication	Productivity	Tools
3	Health&Fitness	Media&Video	Cards&Casino	News&Magazines
4	Tools	Personalization	Sports Games	Tools
5	Health&Fitness	Communication	Brain&Puzzle	Communication
6	Personalization	Casual	Communication	News&Magazines
7	Personalization	Music&Audio	Tools	Personalization
8	Communication	News&Magazines	Sports	Media&Video
9	Personalization	Communication	Arcade&Action	Tools
10	Health&Fitness	Travel&Local	Tools	Communication

Table 7. Comparing the performance of IEDR using different negative context generating methods (for the contrastive learning component).

	Frappe	Yelp
NegGen1	73.01	52.49
NegGen2	71.50	51.82
NegGen1&2	<b>74.11</b>	<b>53.05</b>

longer a valid context representation. In addition, for *NegGen1&2* of all the dropout rates, the results consistently outperform those that only use *NegGen2*.

### 7.7 Effectiveness of Model Hyperparameters

We evaluate our model with different hyperparameter settings, including embedding dimensions, number of negative samples, and loss weight values. Below, we summarize our observations.

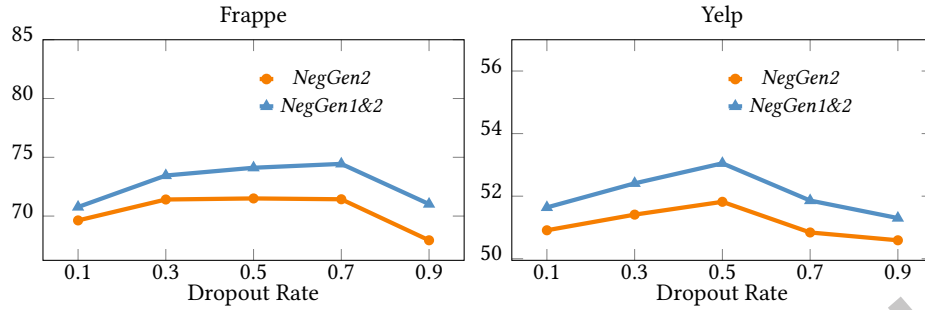


Fig. 9. The performance of different dropout rates for method 2 (NegGen2).

**7.7.1 Embedding Dimension.** We run our model with different feature embedding dimensions. The results are in Figure 10. The embedding dimension poses a trade-off between the expression ability and efficiency. From the figure, we can see that larger dimensions result in better recommendation accuracy. However, the improvement is not significant when the dimension is larger than 32. A larger dimension may even reduce the performance due to the overfitting problem (e.g., dimension 256 for the Frappe dataset).

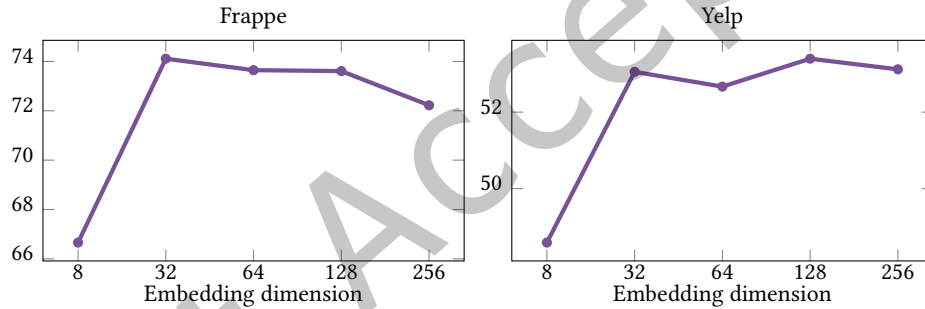


Fig. 10. The performance of different embedding dimensions.

**7.7.2 The Number of Negative Sample and Loss Weight.** The contrastive learning and disentangling components are both contrastive-based methods that require negative sampling. This section evaluates how the number of negative samples influences performance. We also compare the influence of different loss weights of the two components. We run our model with different negative sample numbers and loss weights for the two components, respectively. From Figure 11, we can see that a large loss weight, or a large number of negative samples does not necessarily result in a better performance. Both components should be fine tuned to generate the best outcome. Generally, a very large or small loss weight may make the multi-task training unbalanced, harming the final performance. For the number of negative samples, a small number will make contrastive learning insufficient, while a large number may cause overfitting.

## 7.8 Empirical Analysis of Time Complexity

We summarize the overall time consumption of IEDR and several feature interaction-based baseline models in Table 8. The results are recorded by running the models for one batch (batch size 1024) on the Frappe dataset on a

Table 8. The overall time consumption of different models in one batch training.

Model	Time (ms)
DCNv2	34.40
AutoInt	37.53
SIGN	40.41
IEDR	44.61

Table 9. The time consumption of critical procedures in IEDR in one batch training.

Procedure	Time (ms)
Graph Learning (Feature Interaction Modeling)	14.16
CICL	8.05
Disentangling (step 1)	0.16
Disentangling (step 2)	1.93
Optimization (step 1)	2.21
Optimization (step 2)	8.52

machine with CPU:12th Gen Intel(R) Core(TM) i9-12900K, RAM: 32GB, GPU: NVIDIA GeForce RTX 3090. We can see that our model's overall time consumption is only slightly higher than the other baselines. Next, we summarize the time cost of critical procedures in IEDR in Table 9. The first four rows are model forwarding procedures, and the last two rows are model (alternative) optimizing procedures. Table 9 shows the feature interaction modeling procedure takes most of the time, which is consistent with our analysis in Section 6.2. CICL and disentangling forward procedures (rows 2-4) do not pose a large overhead since they reuse the feature interaction modeling results. Optimization (step 1) updates the parameters of the model's disentangling component ( $q_1$  and  $q_2$ ), which produces little overhead (2.21 ms) and is negligible in the whole procedure.

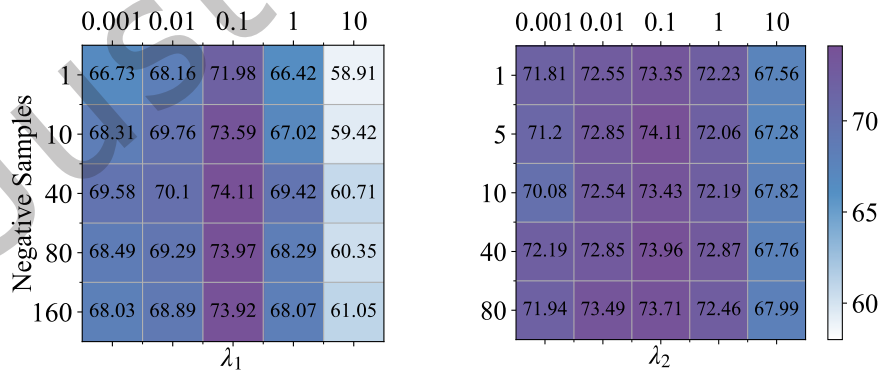


Fig. 11. The performance of different numbers of negative samples and the loss weights in the risk minimization function for the contrastive learning component (left) and the disentangling component (right), respectively.

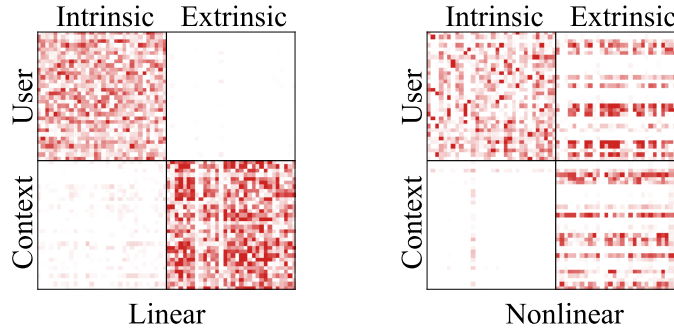


Fig. 12. Visualization of  $f_{ie}$  weights for the *Linear* and *Nonlinear* models.

### 7.9 Empirical Analysis of Falling Into Trivial Solutions

As discussed in Section 6.3, our model may fall into a trivial solution if  $f_{ie}^u(\mathbf{u}, \mathbf{c})$  is a linear mapping method. To evaluate how the trivial solution influences our model in learning the factors, we run our model with  $f_{ie}$  being linear. Specifically, we concatenate  $\mathbf{u}$  and  $\mathbf{c}$  and feed them into an MLP without a hidden layer or activation (a linear mapping), making it easy to fall into the trivial solution. We call this variation *Linear*. Then, we avoid this by simply adding a nonlinear activation function (ReLU) activation after the linear mapping. We call this variation *Nonlinear*. Figure 12 shows the weight values of  $f_{ie}$  of the two variations. The color shows the weights mapping from user/context representations to intrinsic/extrinsic representations. The darker the color, the larger the weight (the more information of user/context is mapped into intrinsic/extrinsic representations). The figure shows that in the *Linear* variation, user information is largely mapped into intrinsic representation (user-intrinsic block) but not extrinsic representation (user-extrinsic block). Context information is largely mapped into extrinsic representation (context-extrinsic block) but not intrinsic representation (context-intrinsic block). This means that the *Linear* variation falls into the trivial solution. On the contrary, in the *Nonlinear* variation, user information is mapped into extrinsic representation (user-extrinsic block), showing that the extrinsic representation contains both user and context information. Figure 13 shows the performance of the two variations. We can see that the *Linear* model performs worse than the *Nonlinear* model. It proves that learning intrinsic and extrinsic factors results in a better performance than simply mapping user and context information into two representations, respectively (the trivial solution).

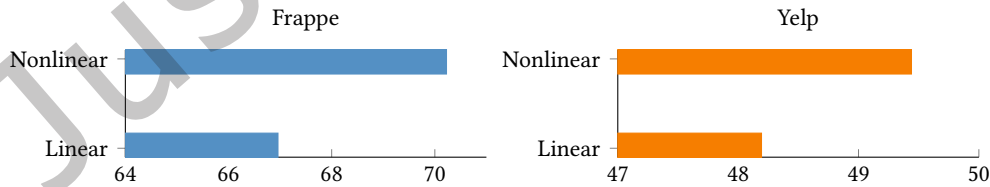


Fig. 13. Comparing the performance of the *Linear* and *Nonlinear* models on different datasets.

## 8 CONCLUSION

To enhance recommendation accuracy, we proposed IEDR, a novel framework that effectively differentiates and captures intrinsic and extrinsic factors from the interplay of various contexts. IEDR leverages a context-invariant

contrastive learning component and a mutual information minimization-based disentangling component to capture consistent user preference and external motivation that may vary across contexts. Extensive experiments on real-world datasets demonstrated IEDR’s effectiveness in learning disentangled factors and significantly improving recommendation accuracy by up to 4% in NDCG. Following this work, we may explore learning more fine-grained intrinsic and extrinsic factors (e.g., multiple intrinsic factors) so that can capture nuanced user interests and generalize our methods to broader applications, e.g., improving the diversity of recommendations. Also, we may explore how to disentangle intrinsic and extrinsic factors when context features are not available.

## ACKNOWLEDGMENTS

This work was financially supported by the National Natural Science Foundation of China (Grant No. 62436003 and 62306333), ARC Discovery Project (Grant No. DP230102908 to Junhao Gan), and ARC Discovery Early Career Researcher Award (DECRA) (Grant No. DE220100680 to Sarah M. Erfani).

## REFERENCES

- [1] Linas Baltrunas, Karen Church, Alexandros Karatzoglou, and Nuria Oliver. 2015. Frappe: Understanding the Usage and Perception of Mobile App Recommendations In-the-wild. *arXiv preprint arXiv:1505.03014* (2015).
- [2] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. 2018. Mutual Information Neural Estimation. In *ICML*. 531–540.
- [3] Roland Bénabou and Jean Tirole. 2003. Intrinsic and Extrinsic Motivation. *The Review of Economic Studies* (2003), 489–520.
- [4] Xuheng Cai, Chao Huang, Lianghao Xia, and Xubin Ren. 2023. LightGCL: Simple Yet Effective Hraph Contrastive Learning for Recommendation. In *ICLR*.
- [5] Han Chen, Ziwen Zhao, Yuhua Li, Yixiong Zou, Ruixuan Li, and Rui Zhang. 2023. CSGCL: Community-strength-enhanced Graph Contrastive Learning. In *IJCAI*. 2059–2067.
- [6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *ICML*. 1597–1607.
- [7] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishu Aradhya, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, et al. 2016. Wide & Deep Learning for Recommender Systems. In *Recsys*. 7–10.
- [8] Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. 2020. Club: A Contrastive Log-ratio Upper Bound of Mutual Information. In *ICML*. PMLR, 1779–1788.
- [9] Jiasheng Duan, Peng-Fei Zhang, Ruihong Qiu, and Zi Huang. 2023. Long Short-term Enhanced Memory for Sequential Recommendation. *World Wide Web* 26, 2 (2023), 561–583.
- [10] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *EMNLP*. 6894–6910.
- [11] Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. 2017. DeepFM: a Factorization-machine based Neural Network for CTR Prediction. In *IJCAI*. 1725–1731.
- [12] Xiangnan He and Tat-Seng Chua. 2017. Neural Factorization Machines for Sparse Predictive Analytics. In *SIGIR*. 355–364.
- [13] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2016. Session-based Recommendations with Recurrent Neural Networks. In *ICLR*. 1–10.
- [14] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. 2019. Learning Deep Representations by Mutual Information Estimation and Maximization. In *ICLR*. 1–14.
- [15] Wang-Cheng Kang and Julian McAuley. 2018. Self-Attentive Sequential Recommendation. In *ICDM*. IEEE, 197–206.
- [16] Huayu Li, Yong Ge, Defu Lian, and Hao Liu. 2017. Learning User’s Intrinsic and Extrinsic Interests for Point-of-Interest Recommendation: A Unified Approach. In *IJCAI*. 2117–2123.
- [17] Honghao Li, Lei Sang, Yi Zhang, Xuyun Zhang, and Yiwen Zhang. 2024. CETN: Contrast-enhanced Through Network for Click-Through Rate Prediction. *TOIS* 43, 1 (2024), 1–34.
- [18] Muyang Li, Zijian Zhang, Xiangyu Zhao, Wanyu Wang, Minghao Zhao, Runze Wu, and Ruocheng Guo. 2023. Automlp: Automated MLP for Sequential Recommendations. In *WWW*. 1190–1198.
- [19] Zihan Lin, Changxin Tian, Yupeng Hou, and Wayne Xin Zhao. 2022. Improving Graph Collaborative Filtering with Neighborhood-enriched Contrastive Learning. In *WWW*. 2320–2329.
- [20] Qijiong Liu, Jieming Zhu, Yanting Yang, Quanyu Dai, Zhaocheng Du, Xiao-Ming Wu, Zhou Zhao, Rui Zhang, and Zhenhua Dong. 2024. Multimodal Pretraining, Adaptation, and Generation for Recommendation: A Survey. In *SIGKDD*. 6566–6576.

- [21] Jianxin Ma, Chang Zhou, Peng Cui, Hongxia Yang, and Wenwu Zhu. 2019. Learning Disentangled Representations for Recommendation. In *NeurIPS*. 5712–5723.
- [22] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based Recommendations on Styles and Substitutes. In *SIGIR*. 43–52.
- [23] Wentao Ning, Xiao Yan, Weiwen Liu, Reynold Cheng, Rui Zhang, and Bo Tang. 2023. Multi-domain Recommendation with Embedding Disentangling and Domain Alignment. In *CIKM*. 1917–1927.
- [24] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation Learning with Contrastive Predictive Coding. *arXiv preprint arXiv:1807.03748* (2018).
- [25] Steffen Rendle. 2010. Factorization Machines. In *ICDM*. 995–1000.
- [26] Richard M Ryan and Edward L Deci. 2000. Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions. *Contemporary Educational Psychology* (2000), 54–67.
- [27] Xiaoteng Shen, Rui Zhang, Xiaoyan Zhao, Jieming Zhu, and Xi Xiao. 2024. PMG: Personalized Multimodal Generation with Large Language Models. In *WWW*. 3833–3843.
- [28] Weiping Song, Chence Shi, Zhiping Xiao, Zhijian Duan, Yewen Xu, Ming Zhang, and Jian Tang. 2019. AutoInt: Automatic Feature Interaction Learning via Self-attentive Neural Networks. In *CIKM*. 1161–1170.
- [29] Daria Sorokina, Rich Caruana, Mirek Riedewald, and Daniel Fink. 2008. Detecting Statistical Interactions with Additive Groves of Trees. In *ICML*. 1000–1007.
- [30] Yixin Su, Rui Zhang, Sarah Erfani, and Zhenghua Xu. 2021. Detecting Beneficial Feature Interactions for Recommender Systems. In *AAAI*. 4357–4365.
- [31] Yixin Su, Rui Zhang, Sarah M. Erfani, and Junhao Gan. 2021. Neural graph matching based collaborative filtering. In *SIGIR*. 849–858.
- [32] 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 *CIKM*. 1441–1450.
- [33] Zhen Tian, Ting Bai, Wayne Xin Zhao, Ji-Rong Wen, and Zhao Cao. 2023. EulerNet: Adaptive Feature Interaction Learning via Euler’s Formula for CTR Prediction. In *SIGIR*. 1376–1385.
- [34] Michael Tsang, Hanpeng Liu, Sanjay Purushotham, Pavankumar Murali, and Yan Liu. 2018. Neural Interaction Transparency (NIT): Disentangling Learned Interactions for Improved Interpretability. In *NeurIPS*. 5804–5813.
- [35] Robert J Vallerand. 1997. Toward a Hierarchical Model of Intrinsic and Extrinsic Motivation. *Advances in Experimental Social Psychology* (1997), 271–360.
- [36] Chenyang Wang, Weizhi Ma, Chong Chen, Min Zhang, Yiqun Liu, and Shaoping Ma. 2023. Sequential Recommendation with Multiple Contrast Signals. *TOIS* 41, 1 (2023), 1–27.
- [37] Fangye Wang, Yingxu Wang, Dongsheng Li, Hansu Gu, Tun Lu, Peng Zhang, and Ning Gu. 2023. Cl4ctr: A Contrastive Learning Framework for CTR Prediction. In *WSDM*. 805–813.
- [38] Jianling Wang, Raphael Louca, Diane Hu, Caitlin Cellier, James Caverlee, and Liangjie Hong. 2020. Time to Shop for Valentine’s Day: Shopping Occasions and Sequential Recommendation in E-commerce. In *WSDM*. 645–653.
- [39] Jinpeng Wang, Ziyun Zeng, Yunxiao Wang, Yuting Wang, Xingyu Lu, Tianxiang Li, Jun Yuan, Rui Zhang, Hai-Tao Zheng, and Shu-Tao Xia. 2023. MISSRec: Pre-training and Transferring Multi-modal Interest-aware Sequence Representation for Recommendation. In *MM*. 6548–6557.
- [40] Ruoxi Wang, Rakesh Shivanna, Derek Cheng, Sagar Jain, Dong Lin, Lichan Hong, and Ed Chi. 2021. DCN V2: Improved Deep & Cross Network and Practical Lessons for Web-scale Learning to Rank Systems. In *WWW*. 1785–1797.
- [41] Wenjie Wang, Xinyu Lin, Liuhui Wang, Fuli Feng, Yunshan Ma, and Tat-Seng Chua. 2023. Causal Disentangled Recommendation Against User Preference Shifts. *TOIS* 42, 1 (2023), 1–27.
- [42] Xiang Wang, Hongye Jin, An Zhang, Xiangnan He, Tong Xu, and Tat-Seng Chua. 2020. Disentangled Graph Collaborative Filtering. In *SIGIR*. 1001–1010.
- [43] Jiancan Wu, Xiangnan He, Xiang Wang, Qifan Wang, Weijian Chen, Jianxun Lian, and Xing Xie. 2022. Graph Convolution Machine for Context-aware Recommender System. *Frontiers of Computer Science* (2022), 1–12.
- [44] Jiancan Wu, Xiang Wang, Fuli Feng, Xiangnan He, Liang Chen, Jianxun Lian, and Xing Xie. 2021. Self-supervised Graph Learning for Recommendation. In *SIGIR*. 726–735.
- [45] Yuxia Wu, Ke Li, Guoshuai Zhao, and QIAN Xueming. 2020. Personalized Long-and Short-term Preference Learning for Next POI Recommendation. *TKDE* (2020), 2301–2304.
- [46] Jun Xiao, Hao Ye, Xiangnan He, Hanwang Zhang, Fei Wu, and Tat-Seng Chua. 2017. Attentional Factorization Machines: Learning the Weight of Feature Interactions via Attention Networks. In *IJCAI*. 3119–3125.
- [47] Dong Yao, Zhou Zhao, Shengyu Zhang, Jieming Zhu, Yudong Zhu, Rui Zhang, and Xiuqiang He. 2022. Contrastive Learning with Positive-negative Frame Mask for Music Representation. In *WWW*. 2906–2915.
- [48] Tiansheng Yao, Xinyang Yi, Derek Zhiyuan Cheng, Felix Yu, Ting Chen, Aditya Menon, Lichan Hong, Ed H Chi, Steve Tjoa, Jieqi Kang, et al. 2021. Self-supervised Learning for Large-scale Item Recommendations. In *CIKM*. 4321–4330.

- [49] Haibo Ye, Xinjie Li, Yuan Yao, and Hanghang Tong. 2023. Towards Robust Neural Graph Collaborative Filtering via Structure Denoising and Embedding Perturbation. *TOIS* 41, 3 (2023), 1–28.
- [50] Junliang Yu, Xin Xia, Tong Chen, Lizhen Cui, Nguyen Quoc Viet Hung, and Hongzhi Yin. 2023. XSimGCL: Towards Extremely Simple Graph Contrastive Learning for Recommendation. *TKDE* 36, 2 (2023), 913–926.
- [51] Junliang Yu, Hongzhi Yin, Xin Xia, Tong Chen, Lizhen Cui, and Quoc Viet Hung Nguyen. 2022. Are Graph Augmentations Necessary? Simple Graph Contrastive Learning for Recommendation. In *SIGIR*. 1294–1303.
- [52] Yantao Yu, Zhen Wang, and Bo Yuan. 2019. An Input-aware Factorization Machine for Sparse Prediction.. In *IJCAI*. 1466–1472.
- [53] Zeping Yu, Jianxun Lian, Ahmad Mahmoody, Gongshen Liu, and Xing Xie. 2019. Adaptive User Modeling with Long and Short-Term Preferences for Personalized Recommendation. In *IJCAI*. 4213–4219.
- [54] An Zhang, Leheng Sheng, Zhibo Cai, Xiang Wang, and Tat-Seng Chua. 2024. Empowering Collaborative Filtering with Principled Adversarial Contrastive Loss. In *NeurIPS*. 6242–6266.
- [55] Rui Zhang, Bayu Distiawan Trisedya, Miao Li, Yong Jiang, and Jianzhong Qi. 2022. A Benchmark and Comprehensive Survey on Knowledge Graph Entity Alignment via Representation Learning. *VLDB* 31, 5 (2022), 1143–1168.
- [56] Yu Zheng, Chen Gao, Jianxin Chang, Yanan Niu, Yang Song, Depeng Jin, and Yong Li. 2022. Disentangling Long and Short-Term Interests for Recommendation. In *WWW*. 2256–2267.
- [57] Kun Zhou, Hui Wang, Wayne Xin Zhao, Yutao Zhu, Sirui Wang, Fuzheng Zhang, Zhongyuan Wang, and Ji-Rong Wen. 2020. S3-rec: Self-supervised Learning for Sequential Recommendation with Mutual Information Maximization. In *CIKM*. 1893–1902.
- [58] Jieming Zhu, Quanyu Dai, Liangcai Su, Rong Ma, Jinyang Liu, Guohao Cai, Xi Xiao, and Rui Zhang. 2022. Bars: Towards open Benchmarking for Recommender Systems. In *SIGIR*. 2912–2923.

## A PROOF OF THEOREM 1

PROOF. Since the mutual information is not explicitly intractable, we approximate the right side of Equation (4) with a lower bound (i.e., MINE [2]) and an upper bound (i.e., CLUB [8]) of mutual information, respectively. More formally,

$$\mathcal{I}(\mathbf{o}_{in}^u, \mathbf{u}) \geq \mathcal{I}_{MINE}(\mathbf{o}_{in}^u, \mathbf{u}) := \mathbb{E}_{p(\mathbf{o}_{in}^u, \mathbf{u})} [\log p(\mathbf{o}_{in}^u, \mathbf{u})] - \log \mathbb{E}_{p(\mathbf{o}_{in}^u)p(\mathbf{u})} [p(\mathbf{o}_{in}^u, \mathbf{u})], \quad (7)$$

$$\mathcal{I}(\mathbf{o}_{in}^u, \mathbf{c}) \leq \mathcal{I}_{CLUB}(\mathbf{o}_{in}^u, \mathbf{c}) := \mathbb{E}_{p(\mathbf{o}_{in}^u, \mathbf{c})} [\log p(\mathbf{o}_{in}^u | \mathbf{c})] - \mathbb{E}_{p(\mathbf{o}_{in}^u)p(\mathbf{c})} [\log p(\mathbf{o}_{in}^u | \mathbf{c})]. \quad (8)$$

With the approximated terms above, proving Equation. (4) turns to verify:

$$\arg \min \sum_{i=1}^N \mathcal{L}_{CICL}(u_i, c_i) = \arg \max (\mathcal{I}_{MINE}(\mathbf{o}_{in}^u, \mathbf{u}) - \mathcal{I}_{CLUB}(\mathbf{o}_{in}^u, \mathbf{c})).$$

By minimizing  $\mathcal{L}_{CICL}$ , we aim to make  $(\mathbf{o}_{in}^u)_{ii}$  similar to  $(\mathbf{o}_{in}^u)_{ij}$ . This procedure can be interpreted in probability as: increasing the probability of  $f_{ie}^u(\mathbf{u}_i, \mathbf{c}_j)$  to predict  $(\mathbf{o}_{in}^u)_{ii}$ . Therefore, maximizing the  $\exp(\text{sim}((\mathbf{o}_{in}^u)_{ii}, (\mathbf{o}_{in}^u)_{ij})/\tau)$  in Equation (2) is equivalent to maximizing  $p((\mathbf{o}_{in}^u)_{ii} | \mathbf{u}_i, \mathbf{c}_j)$  ( $\exp(\cdot)$  is monotone increasing so that does not influence the conclusion). Similar to the above conclusion, minimizing  $\exp(\text{sim}((\mathbf{o}_{in}^u)_{ii}, (\mathbf{o}_{in}^u)_{\ell i})/\tau)$  is equivalent to minimizing  $p((\mathbf{o}_{in}^u)_{ii} | \mathbf{u}_\ell, \mathbf{c}_i)$ . Therefore, we have

$$\begin{aligned} & - \sum_{i=1}^N \mathcal{L}_{CICL}(u_i, c_i) \\ &= \sum_{i=1}^N \log \frac{\exp(\text{sim}((\mathbf{o}_{in}^u)_{ii}, (\mathbf{o}_{in}^u)_{ij})/\tau)}{\sum_{u_\ell \in \mathcal{U}} \exp(\text{sim}((\mathbf{o}_{in}^u)_{ii}, (\mathbf{o}_{in}^u)_{\ell i})/\tau)} \\ &= \sum_{i=1}^N \log[\exp(\text{sim}((\mathbf{o}_{in}^u)_{ii}, (\mathbf{o}_{in}^u)_{ij})/\tau)] - \sum_{i=1}^N \log[\sum_{u_\ell \in \mathcal{U}} \exp(\text{sim}((\mathbf{o}_{in}^u)_{ii}, (\mathbf{o}_{in}^u)_{\ell i})/\tau)] \\ &= \sum_{i=1}^N \log[p((\mathbf{o}_{in}^u)_{ii} | \mathbf{u}_i, \mathbf{c}_j)] - \sum_{i=1}^N \log[\sum_{u_\ell \in \mathcal{U}} p((\mathbf{o}_{in}^u)_{ii} | \mathbf{u}_\ell, \mathbf{c}_i)]. \end{aligned}$$

Equation (2) only samples one context  $c_j$  for each data point. However, during the training, all contexts in  $C$  are expected to be sampled. If we count all contexts, we have

$$\begin{aligned} & \sum_{i=1}^N \log[p((\mathbf{o}_{in}^u)_{ii} | \mathbf{u}_i, \mathbf{c}_j)] - \sum_{i=1}^N \log[\sum_{u_\ell \in \mathcal{U}} p((\mathbf{o}_{in}^u)_{ii} | \mathbf{u}_\ell, \mathbf{c}_i)] \\ &= \sum_{i=1}^N \sum_{c_j \in C} \log[p((\mathbf{o}_{in}^u)_{ii} | \mathbf{u}_i, \mathbf{c}_j)] - \sum_{i=1}^N \log[\sum_{u_\ell \in \mathcal{U}} p((\mathbf{o}_{in}^u)_{ii} | \mathbf{u}_\ell, \mathbf{c}_i)] \\ &= \mathbb{E}_{p(\mathbf{o}_{in}^u, \mathbf{u})p(\mathbf{c})} [\log p(\mathbf{o}_{in}^u | \mathbf{u}, \mathbf{c})] - \mathbb{E}_{p(\mathbf{o}_{in}^u, \mathbf{c})} \log \mathbb{E}_{p(\mathbf{u})} [p(\mathbf{o}_{in}^u | \mathbf{u}, \mathbf{c})]. \end{aligned} \quad (9)$$

Equation (9) is the probability form of the objective function of the context-invariant counteractive learning component (Equation (2)). Equation (9) maximizes the likelihood  $p(\mathbf{o}_{in}^u | \mathbf{u}, \mathbf{c})$  given the joint distribution of users

and intrinsic factors, with the marginal distribution of contexts. Meanwhile, it minimizes the likelihood  $p(\mathbf{o}_{in}^u | \mathbf{u}, \mathbf{c})$  given the joint distribution of contexts and intrinsic factors, with the marginal distribution of the user.<sup>2</sup>

From Equation (9), we further have:

$$\begin{aligned}
& \mathbb{E}_{p(\mathbf{o}_{in}^u, \mathbf{u})p(\mathbf{c})} [\log p(\mathbf{o}_{in}^u | \mathbf{u}, \mathbf{c})] - \mathbb{E}_{p(\mathbf{o}_{in}^u, \mathbf{c})} \log \mathbb{E}_{p(\mathbf{u})} [p(\mathbf{o}_{in}^u | \mathbf{u}, \mathbf{c})] \\
& \stackrel{(a)}{=} \mathbb{E}_{p(\mathbf{o}_{in}^u, \mathbf{u})p(\mathbf{c})} [\log p(\mathbf{o}_{in}^u | \mathbf{u}, \mathbf{c})] - \mathbb{E}_{p(\mathbf{o}_{in}^u, \mathbf{c})p(\mathbf{u})} [\log p(\mathbf{o}_{in}^u | \mathbf{u}, \mathbf{c})] \\
& = \mathbb{E}_{p(\mathbf{o}_{in}^u, \mathbf{u})p(\mathbf{c})} [\log p(\mathbf{o}_{in}^u | \mathbf{u}, \mathbf{c})] - \mathbb{E}_{p(\mathbf{o}_{in}^u, \mathbf{c})p(\mathbf{u})} [\log p(\mathbf{o}_{in}^u | \mathbf{u}, \mathbf{c})] + (\mathbb{E}_{p(\mathbf{u})} [\log p(\mathbf{u})] - \mathbb{E}_{p(\mathbf{u})} [\log p(\mathbf{u})]) \\
& = \mathbb{E}_{p(\mathbf{o}_{in}^u, \mathbf{u})p(\mathbf{c})} [\log p(\mathbf{o}_{in}^u | \mathbf{u}, \mathbf{c})p(\mathbf{u})] - \mathbb{E}_{p(\mathbf{o}_{in}^u, \mathbf{c})p(\mathbf{u})} [\log p(\mathbf{o}_{in}^u | \mathbf{u}, \mathbf{c})] - \mathbb{E}_{p(\mathbf{u})} [\log p(\mathbf{u})] \\
& = \mathbb{E}_{p(\mathbf{o}_{in}^u, \mathbf{u})p(\mathbf{c})} [\log p(\mathbf{o}_{in}^u | \mathbf{u}, \mathbf{c})] - \mathbb{E}_{p(\mathbf{o}_{in}^u, \mathbf{c})p(\mathbf{u})} [\log p(\mathbf{o}_{in}^u | \mathbf{u}, \mathbf{c})] - \mathbb{E}_{p(\mathbf{u})} [\log p(\mathbf{u})] \\
& = \mathbb{E}_{p(\mathbf{o}_{in}^u, \mathbf{u})p(\mathbf{c})} [\log p(\mathbf{o}_{in}^u | \mathbf{u}, \mathbf{c})] - \mathbb{E}_{p(\mathbf{o}_{in}^u, \mathbf{c})p(\mathbf{u})} [\log p(\mathbf{o}_{in}^u | \mathbf{u}, \mathbf{c})] \\
& \quad - \mathbb{E}_{p(\mathbf{u})} [\log p(\mathbf{u})] + \left( \mathbb{E}_{p(\mathbf{o}_{in}^u)p(\mathbf{u})p(\mathbf{c})} [\log p(\mathbf{o}_{in}^u | \mathbf{u}, \mathbf{c})] - \mathbb{E}_{p(\mathbf{o}_{in}^u)p(\mathbf{u})p(\mathbf{c})} [\log p(\mathbf{o}_{in}^u | \mathbf{u}, \mathbf{c})] \right) \\
& = \mathbb{E}_{p(\mathbf{o}_{in}^u, \mathbf{u})p(\mathbf{c})} [\log p(\mathbf{o}_{in}^u | \mathbf{u}, \mathbf{c})] - \mathbb{E}_{p(\mathbf{o}_{in}^u)p(\mathbf{u})p(\mathbf{c})} [\log p(\mathbf{o}_{in}^u | \mathbf{u}, \mathbf{c})] - \mathbb{E}_{p(\mathbf{u})} [\log p(\mathbf{u})] \\
& \quad - \mathbb{E}_{p(\mathbf{o}_{in}^u, \mathbf{c})p(\mathbf{u})} [\log p(\mathbf{o}_{in}^u | \mathbf{u}, \mathbf{c})] + \mathbb{E}_{p(\mathbf{o}_{in}^u)p(\mathbf{u})p(\mathbf{c})} [\log p(\mathbf{o}_{in}^u | \mathbf{u}, \mathbf{c})] \\
& = \mathbb{E}_{p(\mathbf{o}_{in}^u, \mathbf{u})p(\mathbf{c})} [\log p(\mathbf{o}_{in}^u | \mathbf{u}, \mathbf{c})] - \mathbb{E}_{p(\mathbf{o}_{in}^u)p(\mathbf{u})p(\mathbf{c})} [\log p(\mathbf{o}_{in}^u | \mathbf{u}, \mathbf{c})] \\
& \quad - \left( \mathbb{E}_{p(\mathbf{o}_{in}^u, \mathbf{c})p(\mathbf{u})} [\log p(\mathbf{o}_{in}^u | \mathbf{u}, \mathbf{c})] - \mathbb{E}_{p(\mathbf{o}_{in}^u)p(\mathbf{u})p(\mathbf{c})} [\log p(\mathbf{o}_{in}^u | \mathbf{u}, \mathbf{c})] \right) \\
& = \mathbb{E}_{p(\mathbf{c})} \left( \mathbb{E}_{p(\mathbf{o}_{in}^u, \mathbf{u})} [\log p(\mathbf{o}_{in}^u | \mathbf{u}, \mathbf{c})] - \mathbb{E}_{p(\mathbf{o}_{in}^u)p(\mathbf{u})} [\log p(\mathbf{o}_{in}^u | \mathbf{u}, \mathbf{c})] \right) \\
& \quad - \mathbb{E}_{p(\mathbf{u})} \left( \mathbb{E}_{p(\mathbf{o}_{in}^u, \mathbf{c})} [\log p(\mathbf{o}_{in}^u | \mathbf{u}, \mathbf{c})] - \mathbb{E}_{p(\mathbf{o}_{in}^u)p(\mathbf{c})} [\log p(\mathbf{o}_{in}^u | \mathbf{u}, \mathbf{c})] \right). \tag{10}
\end{aligned}$$

(a): In the second term, pushing the log inside the expectation does not change the minimizer.

Comparing Equation (7) and the first term of Equation (10), they both act like classifiers whose objectives maximize the expected log-ratio of the joint distribution over the product of marginal distributions [14]. Therefore, maximizing this term in Equation (10) will have the same effect as maximizing Equation (7). We can interpret the first term of Equation (10) as maximizing the mutual information between users and the corresponding intrinsic factor, conditioned on a given context. Similarly, maximizing the negative of the second term of Equation (10) will have the same effect of minimizing Equation (8), which can be interpreted as minimizing the mutual information between contexts and the corresponding intrinsic factors, conditioned on a given user.

Therefore, we can conclude that:

$$\begin{aligned}
& \arg \min \sum_{(u_i, v_i, c_i) \in \mathcal{D}} \mathcal{L}_{\text{CICL}}(u_i, c_i) \\
& = \arg \max \mathcal{I}_{\text{MINE}}(\mathbf{o}_{in}^u, \mathbf{u}) - \mathcal{I}_{\text{CLUB}}(\mathbf{o}_{in}^u, \mathbf{c}).
\end{aligned}$$

□

<sup>2</sup>Note that only if  $f_{ie}^u(\mathbf{u}, \mathbf{c})$  is a many-to-one (or one-to-one) mapping then Equation (9) and Equation (2) will be equivalent. Otherwise, given a sample pair  $(\mathbf{u}, \mathbf{c})$ ,  $f_{ie}^u(\mathbf{u}, \mathbf{c})$  may have different  $\mathbf{o}_{in}^u$  outputs (i.e., one-to-many). In this situation, the first term of Equation (9) cannot guarantee that the same user with different context will have the same intrinsic factor (i.e., they may have various intrinsic factor representations while still meet the objective of the first term of Equation (9)). We use an MLP as  $f_{ie}^u(\mathbf{u}, \mathbf{c})$ , which is a many-to-one mapping function. Therefore, we can ensure the equivalence between Equation (9) and Equation (2).

Table 10. Implementation details of different variations on the recommendation prediction module. “-” represent the operation is the same as our original IEDR setting.

Variation	Recommendation Prediction Module	
	feature model <sup>3</sup>	$f_{ie}$ <sup>4</sup>
IEDR	$\phi(\psi(\{MLP(z_i^u \odot z_j^u)\}_{j \in u}))_{i \in u} \rightarrow \mathbf{u}$	$MLP(\mathbf{u} \circ \mathbf{c}) \rightarrow [\mathbf{o}_{in}^u, \mathbf{o}_{ex}^u]$
AVG	$\psi(z_i^u)_{i \in u} \rightarrow \mathbf{u}$	-
MLP	$MLP(\psi(z_i^u)_{i \in u}) \rightarrow \mathbf{u}$	-
BI	$\psi(z_i^u \odot z_j^u)_{i,j \in u} \rightarrow \mathbf{u}$	-
Linear	-	$\mathbf{W}[\mathbf{u}, \mathbf{c}] \rightarrow [\mathbf{o}_{in}^u, \mathbf{o}_{ex}^u]$
Nonlinear	-	$\sigma(\mathbf{W}[\mathbf{u}, \mathbf{c}]) \rightarrow [\mathbf{o}_{in}^u, \mathbf{o}_{ex}^u]$
IEDR <sub>sp</sub>	-	$MLP_1(\mathbf{u}) \rightarrow \mathbf{o}_{in}^u, MLP_2(\mathbf{u} \circ \mathbf{c}) \rightarrow \mathbf{o}_{ex}^u$

Table 11. Implementation details of different variations of the contrastive intrinsic-extrinsic disentanglement module. “-” represents the operation as the same as our original IEDR setting. × represents the variation that does not contain the component.

Variation	Contrastive Intrinsic-Extrinsic Disentangling Module	
	Contrastive Learning Component <sup>5</sup>	Disentangling Component
IEDR	positive sample: $f_{ie}^u(\mathbf{u}_i, \mathbf{c}_j) \rightarrow (\mathbf{o}_{in}^u)_{ij}$ negative sample: $f_{ie}^u(\mathbf{u}_\ell, \mathbf{c}_i) \rightarrow (\mathbf{o}_{in}^u)_{\ell i}$ $f_{ie}^u(\mathbf{u}_\ell, \mathbf{c}_j) \rightarrow (\mathbf{o}_{in}^u)_{\ell j}$ $c_j = \text{randChoice}(\text{NegGen1}, \text{NegGen2})$	$MLP_{\theta_1}(\mathbf{o}_{in}^u) \rightarrow (\mathbf{o}_{ex}^u)' (q_1^u)$ $MLP_{\theta_2}(\mathbf{o}_{ex}^u) \rightarrow (\mathbf{o}_{in}^u)' (q_2^u)$
noDis	-	×
noCL	×	-
noCIED	×	×
NegGen1	$c_j$ is generated from <i>NegGen1</i>	-
NegGen2	$c_j$ is generated from <i>NegGen2</i>	-
NegGen1&2	-	-
vCLUB	-	$MLP_{\theta_1}(\mathbf{o}_{in}^u) \rightarrow (\mathbf{o}_{ex}^u)' (q_1^u)$
BiDis	-	-
IEDR <sub>sp</sub>	positive sample: $\text{dropout}((\mathbf{o}_{in}^u)_i) \rightarrow (\mathbf{o}_{in}^u)^p$ negative sample: $\text{dropout}((\mathbf{o}_{in}^u)_\ell) \rightarrow (\mathbf{o}_{in}^u)^n$	-

<sup>3</sup>Here we use user representation learning as an example. The item and context learning have the same structure.  $\phi, \psi$  are both element-wise averaging functions and  $\odot$  is the element-wise product.

<sup>4</sup>Here we use user factor learning as an example.  $\circ$  is a flexible operation to combine two vectors, i.e.,  $\circ$  is an element-wise product for the Frappe dataset, and an element-wise summation for the Yelp dataset.  $[\cdot, \cdot]$  is the concatenation operation.  $\mathbf{W}$  is a linear transformation matrix,  $\sigma$  is a ReLU activation.

<sup>5</sup>For IEDR<sub>sp</sub>, the positive samples  $(\mathbf{o}_{in}^u)^p$  are generated through a dropout of the intrinsic representation of the user, and the negative samples  $(\mathbf{o}_{in}^u)^n$  are generated through a dropout of intrinsic representations of random users.

## B ALGORITHM

This section provides the training process of our IEDR model in Algorithm 1. In each epoch, we use the batch stochastic gradient descent method.

Just Accepted

---

**Algorithm 1** Batch stochastic gradient descent training of IEDR.
 

---

```

1: Input:  $\mathcal{D} = \{(u_i, v_i, c_i)\}_{i=1:N}$  with the corresponding true label  $y_i$  for each data sample.
2: Hyperparameters:  $B$ : batch size;  $L$ : negative sample number for the context-invariant contrastive learning component;  $L_{dis}$ : negative sample number for the disentangling component.
3: Parameters:  $\theta_1^u, \theta_2^u, \theta_1^v, \theta_2^v$ : parameters for  $q_1^u, q_2^u, q_2^v, q_2^v$ , respectively;  $\omega$ : parameters of IEDR except for  $\theta_1^u, \theta_2^u, \theta_1^v, \theta_2^v$ .
4: function CONTRASTIVELEARNING_USER( $\{(u_i, c_i)\}_{i=1:B}$ )
5:   for  $i = 1, \dots, B$  do
6:      $(\mathbf{o}_{in}^u)_{ii} \leftarrow f_{ie}^u(\mathbf{u}_i, \mathbf{c}_i)$ 
7:      $ContextGen \leftarrow RandomChoice(NegGen1, NegGen2)$ 
8:      $c_j \leftarrow ContextGen(c_i)$ 
9:      $(\mathbf{o}_{in}^u)_{ij} \leftarrow f_{ie}^u(\mathbf{u}_i, \mathbf{c}_j)$  ▷ Generate positive samples.
10:    for  $\ell = 1, \dots, L$  do ▷ Generate negative samples.
11:       $u_{\ell_1} \leftarrow randomChoice(\{u_i\}_{i=1:B}), (\mathbf{o}_{in}^u)_{\ell_1 i} = f_{ie}^u(\mathbf{u}_{\ell_1}, \mathbf{c}_i)$ 
12:       $u_{\ell_2} \leftarrow randomChoice(\{u_i\}_{i=1:B}), (\mathbf{o}_{in}^u)_{\ell_2 j} = f_{ie}^u(\mathbf{u}_{\ell_2}, \mathbf{c}_j)$ 
13:    end for
14:     $\mathcal{L}_{CICL}(u_i, c_i) \leftarrow$  Equation (4) based on the above positive and negative samples
15:  end for
16:  return  $average(\{\mathcal{L}_{CICL}(u_i, c_i)\}_{i=1:B})$ 
17: end function
18: function CONTRASTIVELEARNING_ITEM( $\{(v_i, c_i)\}_{i=1:B}$ )
19:   Symmetric to CONTRASTIVELEARNING_USER.
20: end function
21: function DISENTANGLEMENT_USER( $\{(u_i, c_i)\}_{i=1:B}$ )
22:   for  $i = 1, \dots, B$  do
23:      $(\mathbf{o}_{in}^u)_{ii}, (\mathbf{o}_{ex}^u)_{ii} \leftarrow f_{ie}^u(\mathbf{u}_i, \mathbf{c}_i)$ 
24:      $(\mathbf{o}_{ex}^u)_{ii}^{pred} \leftarrow q_{\theta_1}((\mathbf{o}_{in}^u)_{ii}), (\mathbf{o}_{in}^u)_{ii}^{pred} \leftarrow q_{\theta_2}((\mathbf{o}_{ex}^u)_{ii})$  ▷ Generate positive samples.
25:      $a_{pos}^{\rightarrow} \leftarrow MSE((\mathbf{o}_{ex}^u)_{ii}, (\mathbf{o}_{ex}^u)_{ii}^{pred}), a_{pos}^{\leftarrow} \leftarrow MSE((\mathbf{o}_{in}^u)_{ii}, (\mathbf{o}_{in}^u)_{ii}^{pred})$ 
26:      $a_{neg}^{\rightarrow} \leftarrow 0, a_{neg}^{\leftarrow} \leftarrow 0$ 
27:     for  $j = 1, \dots, L_{dis}$  do ▷ Generate negative samples.
28:        $(\mathbf{o}_{in}^u)_r, (\mathbf{o}_{ex}^u)_r \leftarrow randomChoice(\{((\mathbf{o}_{in}^u)_{ii}, (\mathbf{o}_{ex}^u)_{ii})\}_{i=1:B})$ 
29:        $(\mathbf{o}_{ex}^u)_r^{pred} = q_{\theta_1}((\mathbf{o}_{in}^u)_r), (\mathbf{o}_{in}^u)_r^{pred} = q_{\theta_2}((\mathbf{o}_{ex}^u)_r)$ 
30:        $a_{neg}^{\rightarrow} \leftarrow a_{neg}^{\rightarrow} + MSE((\mathbf{o}_{ex}^u)_{ii}, (\mathbf{o}_{ex}^u)_r^{pred})$ 
31:        $a_{neg}^{\leftarrow} \leftarrow a_{neg}^{\leftarrow} + MSE((\mathbf{o}_{in}^u)_{ii}, (\mathbf{o}_{in}^u)_r^{pred})$ 
32:     end for
33:      $(\mathcal{L}_{bi-appr})_i \leftarrow \frac{1}{2}(a_{pos}^{\rightarrow} + a_{pos}^{\leftarrow})$ 
34:      $(\mathcal{L}_{Dis})_i \leftarrow \frac{1}{2}(\frac{a_{neg}^{\rightarrow} + a_{neg}^{\leftarrow}}{N_{dis}} - (a_{pos}^{\rightarrow} + a_{pos}^{\leftarrow}))$ 
35:   end for
36:   return  $average(\{(\mathcal{L}_{bi-appr})_i\}_{i=1:B}), average(\{(\mathcal{L}_{Dis})_i\}_{i=1:B})$ 
37: end function
38: function DISENTANGLEMENT_ITEM( $\{(v_i, c_i)\}_{i=1:B}$ )
39:   Symmetric to DISENTANGLEMENT_USER.
40: end function
41:

```

---

**Algorithm 1** Batch stochastic gradient descent training of IEDR (continued).

---

```

42: shuffle( $\{(u_i, v_i, c_i)\}_{i=1:N}$ )
43: for each batch  $\{(u_i, v_i, c_i)\}_{i=1:B}$  do
44:   for  $i = 1, \dots, B$  do                                     ▶ Line 45-47 are the recommendation prediction module.
45:      $u_i \leftarrow f_u(u_i), v_i \leftarrow f_v(v_i), c_i \leftarrow f_c(c_i)$ 
46:      $(\sigma_{in}^u)_{ii}, (\sigma_{ex}^u)_{ii} \leftarrow f_{ie}^u(\mathbf{u}_i, \mathbf{c}_i), (\sigma_{in}^v)_{ii}, (\sigma_{ex}^v)_{ii} \leftarrow f_{ie}^v(\mathbf{v}_i, \mathbf{c}_i)$ 
47:      $y'_i \leftarrow f_{pred}((\sigma_{in}^u)_{ii}, (\sigma_{ex}^u)_{ii}, (\sigma_{in}^v)_{ii}, (\sigma_{ex}^v)_{ii})$ 
48:      $(\mathcal{L}_{RP})_i \leftarrow \text{CrossEntropy}(y'_i, y_i)$ 
49:   end for
50:    $\mathcal{L}_{RP} \leftarrow \text{average}(\{\mathcal{L}_{RP}\}_i)_{i=1:B}$ 
51:    $\mathcal{L}_{CICL}^u \leftarrow \text{CONTRASTIVELEARNING\_USER}(\{(u_i, c_i)\}_{i=1:B})$ 
52:    $\mathcal{L}_{CICL}^v \leftarrow \text{CONTRASTIVELEARNING\_ITEM}(\{(v_i, c_i)\}_{i=1:B})$ 
53:    $\mathcal{L}_{bi-appr}^u, \mathcal{L}_{Dis}^u \leftarrow \text{DISENTANGLEMENT\_USER}(\{(u_i, c_i)\}_{i=1:B})$ 
54:    $\mathcal{L}_{bi-appr}^v, \mathcal{L}_{Dis}^v \leftarrow \text{DISENTANGLEMENT\_ITEM}(\{(v_i, c_i)\}_{i=1:B})$ 
55:   Freeze  $\omega$ , update  $\theta_1^u, \theta_2^u, \theta_1^v, \theta_2^v$  through minimizing  $\mathcal{R}(\theta_1^u, \theta_2^u, \theta_1^v, \theta_2^v)$            ▶ Step 1
56:   Freeze  $\theta_1^u, \theta_2^u, \theta_1^v, \theta_2^v$ , update  $\omega$  through minimizing  $\mathcal{R}(\omega)$            ▶ Step 2
57: end for

```

---