i^{2}VAE: Interest Information Augmentation with Variational Regularizers for Cross-Domain Sequential Recommendation

Xuying Ning[*1] **Wujiang Xu**[*†2] **Tianxin Wei**[1] **Xiaolei Liu**[3]

[1]University of Illinois Urbana-Champaign, Urbana, IL, USA
[2]Rutgers University, New Brunswick, NJ, USA
[3]Independent Researcher

Abstract

Cross-Domain Sequential Recommendation (CDSR) leverages user behaviors across multiple domains to mitigate data sparsity and cold-start challenges in Single-Domain Sequential Recommendation. Existing methods primarily rely on shared users (overlapping users) to learn transferable interest representations. However, these approaches have limited information propagation, benefiting mainly overlapping users and those with rich interaction histories while neglecting non-overlapping (cold-start) and long-tailed users, who constitute the majority in real-world scenarios. To address this issue, we propose i^{2}VAE, a novel variational autoencoder (VAE)-based framework that enhances user interest learning with mutual information-based regularizers. i^{2}VAE improves recommendations for cold-start and long-tailed users while maintaining strong performance across all user groups. Specifically, cross-domain and disentangling regularizers extract transferable features for cold-start users, while a pseudo-sequence generator synthesizes interactions for long-tailed users, refined by a denoising regularizer to filter noise and preserve meaningful interest signals. Extensive experiments demonstrate that i^{2}VAE outperforms state-of-the-art methods, underscoring its effectiveness in real-world CDSR applications. Code and datasets are available at `https://github.com/YennNing/I2VAE`.

1 INTRODUCTION

With the rise of various sequential models, single-domain sequential recommendation (SDSR) [Hidasi et al., 2015, Kang and McAuley, 2018, Sun et al., 2019, Tang and Wang, 2018, Wang et al., 2020, Xu et al., 2024b,a, Wei et al., 2024] has gained increased attention due to its ability to model users' dynamic interests in recommendation systems. However, these SDSR models often suffer from the long-standing data sparsity problem [Lin et al., 2021, Xu et al., 2023a, Wei and He, 2022], where users have few interactions in a domain to learn their preferences effectively. To address this issue, cross-domain sequential recommendation (CDSR) methods [Ma et al., 2019, 2022, Cao et al., 2022a, Li et al., 2021] have been proposed to leverage abundant data from other relevant domains to improve recommendation performance in a data-scarce domain.

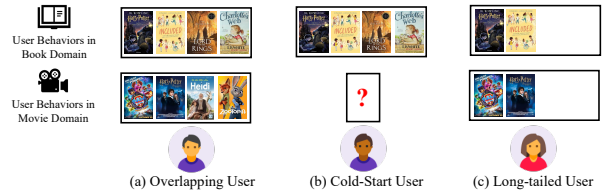


Figure 1: Previous CDSR methods [Li et al., 2021, Ma et al., 2019, Cao et al., 2022a] rely on domain-shared information from overlapping users (Fig. (a)) but fail to effectively model cold-start users' interests (Fig. (b)) and capture long-tail users' preferences with sparse interactions (Fig. (c)). In contrast, i^{2}VAE disentangles cross-domain interests and refines intra-domain interest representations, enhancing recommendations for all user groups.

The core idea of CDSR is to extend SDSR methods by designing cross-domain information transfer modules to capture domain-shared interests that can be transferred across domains. Researchers have employed various techniques, including gating mechanisms [Ma et al., 2019, Sun et al., 2021], attention mechanisms [Li et al., 2021, Xu et al., 2023b], graph neural networks [Guo et al., 2021, Ma et al., 2022], and contrastive learning [Cao et al., 2022a, Xu et al., 2023b], to learn user interests that can be transferred across

[*]Equal contribution.
[†]Correspondence to: Wujiang Xu <wujiang.xu@rutgers.edu>.

domains. Although it appears promising, we find that previous CDSR methods [Ma et al., 2019, Li et al., 2021, Man et al., 2017, Cao et al., 2022a] heavily rely on overlapping users (Figure 1(a)) with rich historical behaviors. They can only effectively learn users' cross-domain interests when a majority (over 70%, as noted by [Xu et al., 2023c]) of users are overlapping across domains, and only adequately capture intra-domain interests when users have extensive historical interactions. However, these conditions are rarely met in real-world recommendation tasks [Xu et al., 2023c,b], where the majority of users are cold-start[1] (Figure 1(b)) or long-tailed[2] (Figure 1(c)). For instance, on platforms like Taobao or Amazon, there are very few overlapping users compared to the entire user base (including overlapping and non-overlapping users), and most users have very few interaction records (a.k.a. long-tailed users). In such cases, these CDSR methods, trained on the sparse historical behaviors of a few overlapping users, often show poor generalization, especially when inferring the intra-domain and cross-domain interests of cold-start and long-tailed users. This presents the **primary challenge**: how to enhance the CDSR model's performance in real-world recommendation scenarios, where most users are either cold-start or long-tailed?

Suffering from cold-start users, previous CDSR and CDR works have often utilized cross-domain modules that learn a mapping function from one domain to another based on the historical behaviors of overlapping users [Kang et al., 2019, Zhu et al., 2021, Salah et al., 2021, Shi and Wang, 2019]. Most recently, some researchers [Xu et al., 2023a,c] construct user-user graph to propagate the cross-domain information for the cold-start users. However, cross-domain information is only learned in domains where users exhibit behaviors, and the density of the constructed graph heavily depends on the density of these user interactions. Therefore, the **1st sub-challenge** is: *how to transfer relevant cross-domain information for cold-start users who lack historical interest data in a domain?* In addition, to address insufficient interest learning of the predominant presence of long-tailed users in real-world scenarios, MACD [Xu et al., 2023b] employs attention mechanisms to discern latent interests from users' auxiliary sequential behavior data. However, such auxiliary behaviors may not always be available in real-world scenarios. Therefore, the **2nd sub-challenge** is: *how to unearth and leverage the latent interest information of long-tailed users, thereby improving model performance in practical CDSR scenarios?*

To address these challenges, we propose Interest Information Augmentation with Variational Regularizers, named as i^{2}VAE, a novel framework that integrates a pseudo-sequence generator, variational autoencoders, and interest-enhancing regularizers (cross-domain, disentangling, and denoising). Our framework effectively explores latent interest information for long-tailed users while transferring disentangled cross-domain interest information for both overlapping and cold-start users, significantly improving performance in real-world scenarios. Recent works [Cao et al., 2022b,c] have explored cross-domain interest transfer but remain limited in scope. DisenCDR [Cao et al., 2022b] assumes full user overlap and applies mutual information for disentanglement, making it ineffective for cold-start users. CDRIB [Cao et al., 2022c] leverages information bottleneck and contrastive learning but struggles to capture long-tailed users' preferences. In contrast, our framework integrates information augmentation and denoising via mutual information, enabling robust learning for both cold-start and long-tailed users in practical scenarios. The main contributions of our work can be summarized as follows:

- **Novel Framework.** We propose i^{2}VAE, a framework based on variational autoencoders that enhances Interest Information through variational regularizers. It uncovers latent interests in long-tailed users and facilitates cross-domain interest transfer for overlapping and cold-start users, significantly improving real-world recommendation performance.
- **Cold-Start Adaptation.** We design new data pathways and introduce cross-domain and disentangling regularizers to jointly model and separate users' intra- and cross-domain interests in partially overlapped CDSR scenarios.
- **Long-Tail Interest Enhancement.** We introduce a pseudo-sequence generator combined with a denoising regularizer to enrich sparse interaction histories while eliminating noises in pseudo-sequences.
- **Empirical and theoretical validation.** We show that i^{2}VAE achieves state-of-the-art performance across all user types, including long-tailed and cold-start users, in real-world cross-domain scenarios. Additionally, we provide rigorous theoretical derivations to support its effectiveness.

2 PRELIMINARY

This section introduces the CDSR problem, where the model uses user behavior from two domains to predict true interests. We also present the concept of mutual information.

2.1 PROBLEM FORMULATION

In this work, we consider a real-world CDSR scenario that includes a fraction of overlapping users, and a majority of long-tailed and cold-start users across two domains, namely domain X and domain Y. The recommendation data is represented by $D^X = (\mathcal{U}^X, \mathcal{V}^X, \mathcal{E}^X)$ and $D^Y = (\mathcal{U}^Y, \mathcal{V}^Y, \mathcal{E}^Y)$, where $\mathcal{U}^{\cdot}$, $\mathcal{V}^{\cdot}$, and $\mathcal{E}^{\cdot}$ are the sets of users, items, and interaction edges, respectively. For a given user, we denote the sequeneces of user-item interaction in chronological order, as $S^X = [v_1^X, v_2^X, \cdots, v_{|S^X|}^X]$

[1]Cold-start users have interactions in only one domain.

[2]Long-tailed users are those with a few interaction records.

and $S^Y = [v_1^Y, v_2^Y, \cdots, v_{|S^Y|}^Y]$, where $|\cdot|$ is the length of user behavior. The objective of CDSR is to predict the next item that each user will purchase based on the user's previous behavior in two domains [Cao et al., 2022c, Xu et al., 2023b].

2.2 PRELIMINARY OF MUTUAL INFORMATION

Mutual Information Maximization [McGill, 1954, Hjelm et al., 2018, Belghazi et al., 2018] is a key mathematical tool for ensuring the robustness of interest augmentation. For random variables X and Y, the *mutual information* [Järvelin and Kekäläinen, 2002, Paninski, 2003] $I(X;Y)$, measuring how much X reduces uncertainty in Y, is defined as: $I(X;Y) = H(X) - H(X|Y)$, where $H(X)$ and $H(X|Y)$ denote the entropy [Rényi, 1961, Cover et al., 1991] of X and the conditional entropy of X given Y. For three random variables X, Y, and Z, the *interaction information* [McGill, 1954, Bell, 2003] $I(X;Y;Z)$, is defined as:

$$I(X;Y;Z) = I(X;Y) - I(X;Y|Z), \tag{1}$$

which captures shared information beyond pairwise mutual information. Notably, all variables are symmetric in these definitions.

3 METHODOLODGY

This section introduces the pseudo-sequence generator, inference and generation procedures of i^{2}VAE, and its interest-enhancing regularizers. Figure 2 provides an overview.

3.1 PSEUDO-SEQUENCE GENERATOR

Behavior sparsity poses a significant challenge in real-world CDSR scenarios, making it difficult for models to capture users' within-domain interests, let alone cross-domain interests. Thus, the pseudo-sequence generator (PSG) *augments user behaviors* by serving as a fast retrieval model that efficiently generates candidate items aligned with user interests but not yet interacted with. To balance efficiency and effectiveness, we use LightGCN [He et al., 2020] with iterative recall to generate pseudo-sequences while avoiding the high computational cost of sequential recommendation models.

The pseudo-sequence generation process consists of three steps: (1) Unified Item Set Construction: Items from both domains are remapped into a unified item set, $\mathcal{V} = \mathcal{V}^X \cup \mathcal{V}^Y$, ensuring no duplicates. A user-item bipartite graph $\mathcal{G}$ is built based on the interaction data $\mathcal{E} = \mathcal{E}^X \cup \mathcal{E}^Y$. (2) Embedding Learning & Recall: Using LightGCN, user and item embeddings are learned as $\boldsymbol{E}_{\mathcal{U}} \in \mathbb{R}^{|\mathcal{U}|\times d}$ and $\boldsymbol{E}_{\mathcal{V}} \in \mathbb{R}^{|\mathcal{V}|\times d}$, where d is the embedding dimension. Predicted user preferences are represented by the rating matrix $\boldsymbol{R} = \boldsymbol{E}_{\mathcal{U}}\boldsymbol{E}_{\mathcal{V}}^T$. The recall process iteratively selects candidate items based on $\boldsymbol{R}$, updates the interaction graph with pseudo-interactions, and refines embeddings through message propagation. (3) Pseudo-Sequence Generation: Users' original interaction sequences S^X and S^Y are expanded with high-scoring items from the recall process that have not been interacted with. For domain X, the pseudo-sequence is defined as $\widetilde{S}^X = S^X \cup \{\widetilde{v}_1, \ldots, \widetilde{v}_{T'}\}$, where the top T' items $\widetilde{v}_i$ are selected by the iterative recall process, satisfying $\widetilde{v}_i \in \mathcal{V}^X \setminus S^X$. A similar process is applied to domain Y simultaneously.

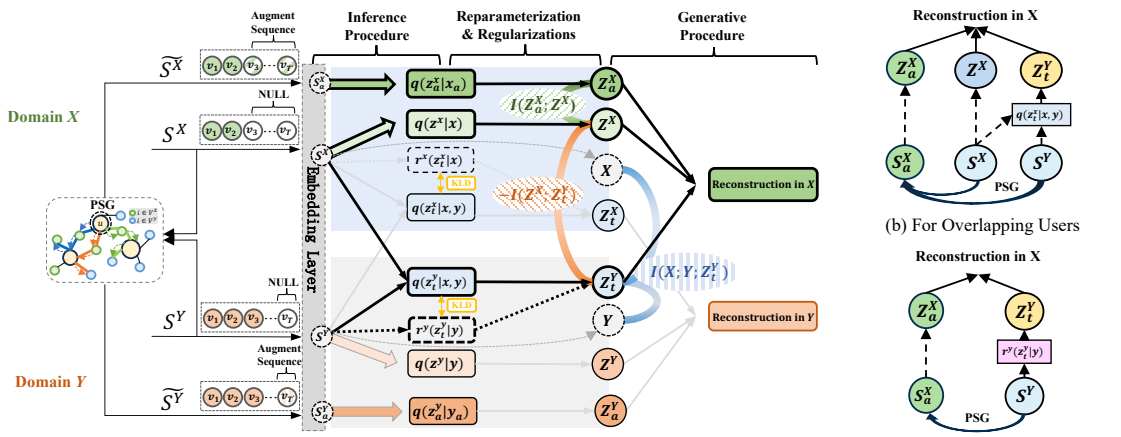
3.2 EMBEDDING LAYER

We employ embedding layers $\boldsymbol{E}^X \in \mathbb{R}^{|\mathcal{V}^X|\times d}$ and $\boldsymbol{E}^Y \in \mathbb{R}^{|\mathcal{V}^Y|\times d}$, along with the self-attention layer from SASRec [Kang and McAuley, 2018], to derive interest representations for both *real* sequences (S^X and S^Y) and *pseudo* sequences ($\widetilde{S}^X$ and $\widetilde{S}^Y$). Specifically, the representations are computed as $\boldsymbol{S^X} = \odot\big(\{\mathbf{h}_{S_1^X}, \cdots, \mathbf{h}_{S_T^X}\}\big)$, $\boldsymbol{S^Y} = \odot\big(\{\mathbf{h}_{S_1^Y}, \cdots, \mathbf{h}_{S_T^Y}\}\big)$, $\boldsymbol{S_a^X} = \odot\big(\{\mathbf{h}_{\widetilde{S}_1^X}, \cdots, \mathbf{h}_{\widetilde{S}_{T'}^X}\}\big)$, $\boldsymbol{S_a^Y} = \odot\big(\{\mathbf{h}_{\widetilde{S}_1^Y}, \cdots, \mathbf{h}_{\widetilde{S}_{T'}^Y}\}\big)$, where $\mathbf{h}_{S_t^X}$ and $\mathbf{h}_{S_t^Y}$ denote the hidden representations of items at position t in sequences S^X and S^Y, and $\odot$ represents mean pooling over the time dimension, producing embeddings in $\mathbb{R}^d$. Here, T and T' are the maximum lengths of the real and pseudo sequences, respectively. For sequences longer than T (T'), only the most recent T (T') actions are retained; for shorter sequences, 'padding' items are added to the left to meet the required length. $\boldsymbol{S^X}$ and $\boldsymbol{S^Y}$ represent the user interests derived from real sequences, while $\boldsymbol{S_a^X}$ and $\boldsymbol{S_a^Y}$ capture the *augmented interests* obtained from pseudo sequences.

3.3 GENERATION AND INFERENCE OF I^{2}VAE

In line with previous CDR studies [Cao et al., 2022b,a], we adopt variational autoencoder as the foundamental architecture due to its ability to model interest decomposition and transfer. We assume that a user's interests in domains X and Y are represented by $\boldsymbol{X}$ and $\boldsymbol{Y}$, which capture the user's true preferences in each domain. These interests follow a bivariate distribution $P_D(\boldsymbol{X}, \boldsymbol{Y})$, and our goal is to reconstruct them in P_D to predict the next items the user will interact with. In our i^{2}VAE, to map the initial interest representations ($\boldsymbol{S^X}, \boldsymbol{S^Y}, \boldsymbol{S_a^X}, \boldsymbol{S_a^Y}$) to their true interests, we design six d-dimensional latent variables in VAE, three for each domain, representing different aspects of user interests to facilitate interest reconstruction:

- $\boldsymbol{Z^X}$, $\boldsymbol{Z^Y}$: *Domain-specific interests* in domain X and Y.
- $\boldsymbol{Z_t^Y}$, $\boldsymbol{Z_t^X}$: *Transferable interests* that capture cross-domain user preferences; $\boldsymbol{Z_t^Y}$ represents the interests in domain Y that are transferable to X, and vise versa for $\boldsymbol{Z_t^X}$.



(a) Framework Overview

(b) For Overlapping Users

(b) For Cold-start Users in Domain X

Figure 2: Figure (a) presents an overview of i^{2}VAE, highlighting computational pathways (thick black lines) and key interest-enhancing regularizers $I(\boldsymbol{Z_a^X};\boldsymbol{Z^X})$, $-I(\boldsymbol{Z^X};\boldsymbol{Z_t^Y})$, and $I(\boldsymbol{X};\boldsymbol{Y};\boldsymbol{Z_t^Y})$ for predictions in domain X. The dashed component $r^y(z_t^y|y)$ replaces the cross-domain module $q(z_t^y|x,y)$ for cold-start users in X. The model is symmetric for domain Y (grey lines). Figures (b) and (c) depict simplified data pathways for overlapping and cold-start users in domain X.

• $\boldsymbol{Z_a^X}$, $\boldsymbol{Z_a^Y}$: *Augmented domain-specific interests* derived from pseudo sequnces in domain X and Y.

3.3.1 Inference Procedure

To model the six representations described above, we use the VAE [Kingma, 2013], which encodes different aspects of a user's interests into latent variables. We assume that the six latent variables are conditionally independent given $\boldsymbol{X}$ and $\boldsymbol{Y}$. Moreover, $\boldsymbol{Z_t^Y}$ and $\boldsymbol{Z_t^X}$ represent cross-domain information, while $\boldsymbol{Z^X}$, $\boldsymbol{Z_a^X}$, $\boldsymbol{Z^Y}$, and $\boldsymbol{Z_a^Y}$ represent domain-specific information that only correlates with one domain. With the previously mentioned assumptions, we can obtain $q(z^x|x,y) = q(z^x|x)$[3], which is applicable to other domain-specific latent variables. Based on this assumption, we factorize q_ϕ as follows.

$$\begin{aligned} &q(z^x, z^y, z_t^y, z_t^x, z_a^x, z_a^y|x,y) \\ &= \underbrace{q(z^x|x)q(z_t^y|x,y)q(z_a^x|x)}_{\textit{Domain X}}\underbrace{q(z^y|y)q(z_t^x|x,y)q(z_a^y|y)}_{\textit{Domain Y}} \end{aligned} \tag{2}$$

Resembling standard VAE's inference process, each $q(\cdot)$ in the factorization above corresponds to a VAE's encoder, and it is assumed to follow a Gaussian distribution. The mean ($\mu^{\cdot}$) and standard deviation vectors ($\sigma^{\cdot}$) of the distribution are generated by their respective encoders. In the equation above, ϕ represents the learnable parameters of all encoders. Details regarding the implementation of domain-specific and cross-domain encoders are provided in Appendix D. Then, using the reparameterization trick, latent variables are generated as follows. For z^x, we have:

$$z^x = \mu^x + \sigma^x \odot \epsilon, \quad \epsilon \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \tag{3}$$

[3]To keep equations compact, we use lowercase, non-bold symbols in prior and posterior expansions to represent their bold uppercase counterparts (e.g., z_t^y for $\boldsymbol{Z_t^Y}$, and x for $\boldsymbol{X}$).

with other latent variables generated similarly.

3.3.2 Generative Procedure

We assume that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are conditionally independent given the six latent variables. Additionally, each domain is associated with only three latent variables. For instance, reconstructing $\boldsymbol{X}$ only involves the domain-specific interest representation $\boldsymbol{Z^X}$, the augmented interest representation $\boldsymbol{Z_a^X}$, and the cross-domain representation $\boldsymbol{Z_t^Y}$ for information transferred from domain Y to X. Therefore, we have $p_\theta(x|z^x, z_t^y, z_a^x, z^y, z_t^x, z_a^y) = p_{\theta_x}(x|z^x, z_t^y, z_a^x)$, which similarly applies to domain Y. Based on this assumption, our generative distribution can be structured as:

$$\begin{aligned} p_\theta(x,y) = \int & p_{\theta_x}(x|z^x, z_t^y, z_a^x)p_{\theta_y}(y|z^y, z_t^y, z_a^y)p(z^x)p(z^y) \\ & \cdot p(z_a^x)p(z_a^y)p(z_t^x)p(z_t^y)dz^x dz^y dz_a^x dz_a^y dz_t^x dz_t^y. \end{aligned}$$

Here, $\theta = \{\theta_x, \theta_y\}$ denotes the parameters of the VAE decoders. The prior distributions $p(\cdot)$ for cross-domain latent variables (z_t^x and z_t^y) are set as standard normal distributions, $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, while domain-specific priors vary in mean and standard deviation to reflect domain distinctions.

3.3.3 Evidence Lower Bound of i^{2}VAE

Similar to standard VAE approaches, we derive the Evidence Lower Bound (ELBO) of the optimization objective [Kingma, 2013, Salah et al., 2021] based on the assumptions in the generation and inference. It comprises two reconstruction terms ($\mathbb{E}_{q_\phi}[\cdot]$), aiming to reconstruct the user's true interests in domains X and Y, and six Kullback-Leibler (KL) divergences ($D_{KL}[\cdot]$), which regularize the

latent variable distributions to align with their priors:

$$\begin{aligned}\log p(x,y) &\geq \text{ELBO} \\ &= \mathbb{E}_{q_\phi}[\log p(x|z^x, z_t^y, z_a^x)] + \mathbb{E}_{q_\phi}[\log p(y|z^y, z_t^x, z_a^y)] \\ &- D_{KL}[q(z^x|x)\|p(z^x)] - D_{KL}[q(z^y|y)\|p(z^y)] \\ &- D_{KL}[q(z_t^y|x,y)\|p(z_c^y)] - D_{KL}[q(z_t^x|x,y)\|p(z_c^x)] \\ &- D_{KL}[q(z_a^x|x)\|p(z_a^x)] - D_{KL}[q(z_a^y|y)\|p(z_a^y)].\end{aligned} \tag{4}$$

For the reconstruction term in domain X, an MLP functions as the VAE decoder, reconstructing the user's true interest representation $\boldsymbol{X} \in \mathbb{R}^d$ from three domain-specific latent variables. Each item in domain X is associated with an embedding $\boldsymbol{E}_{v_i}^X \in \mathbb{R}^d$, which is used to estimate the user's preference for an item via their dot product: $\hat{r}_{ui}^X = \boldsymbol{X}_u \cdot \boldsymbol{E}_{v_i}^X$. The reconstruction term is optimized by minimizing the Binary Cross-Entropy Loss between the predicted interaction probability $\hat{r}_{ui}^X$ and the true interaction label r_{ui}^X. The same process applies to domain Y.

3.4 INTEREST-ENHANCING REGULARIZERS

Using an unconstrained VAE model alone cannot guarantee reliable interest augmentation. We identify two key aspects of this reliability: (a) Predicting interactions in the cold-start domain relies *solely* on cross-domain transferable interests, which must be *separate* from domain-specific interests in the other domain. (b) Augmenting intra-domain interests may introduce noise due to inaccuracies in the PSG, thus a *denoising mechanism* is needed to filter out these deviations and preserve the true intra-domain augmented interests. To address these reliability challenges, we propose three interest-enhancing regularizers—cross-domain, disentangling, and denoising—illustrated using domain X, with symmetric designs for domain Y omitted for brevity.

3.4.1 Cross-domain Regularizer $I(\boldsymbol{X};\boldsymbol{Y};\boldsymbol{Z_t^Y})$

We aim for the the transferable interest representation $\boldsymbol{Z_t^Y}$ to capture user interests that can transfer from domain Y to X, acting as a cross-domain signal, such as *shared themes or topics across domains like books and movies.* To achieve this, we propose the first regularizer—maximizing the interaction information $I(\boldsymbol{X};\boldsymbol{Y};\boldsymbol{Z_t^Y})$—which quantifies the interdependence and shared information among these variables. Using the formal definition of interaction information (Eq. (1)), we expand $I(\boldsymbol{X};\boldsymbol{Y};\boldsymbol{Z_t^Y})$ as:

$$I(\boldsymbol{X};\boldsymbol{Y};\boldsymbol{Z_t^Y}) = I(\boldsymbol{X};\boldsymbol{Z_t^Y}) - I(\boldsymbol{X};\boldsymbol{Z_t^Y}|\boldsymbol{Y}). \tag{5}$$

The above equation can be intuitively explained from two perspectives: (1) it increases $I(\boldsymbol{X};\boldsymbol{Z_t^Y})$, ensuring $\boldsymbol{Z_t^Y}$ captures information relevant to $\boldsymbol{X}$; (2) it minimizes $I(\boldsymbol{X};\boldsymbol{Z_t^Y}|\boldsymbol{Y})$, which helps restrict $\boldsymbol{Z_t^Y}$ to information inferred from domain Y. By doing so, the cross-domain latent variable, $\boldsymbol{Z_t^Y}$, effectively encodes shared and transferable user-interest signals across the two domains.

3.4.2 Disentangling Regularizer $-I(\boldsymbol{Z^X};\boldsymbol{Z_t^Y})$

For cold-start users in domain X, only $\boldsymbol{Z_t^Y}$, the cross-domain representation, is utilized for reconstruction, while all domain-specific representations in X are masked. This design requires $\boldsymbol{Z_t^Y}$ to focus *exclusively* on transferable cross-domain interests without contamination from domain-specific information in Y, as such contamination could lead to negative transfer [Zhang et al., 2023, Park et al., 2023].

To achieve this, we propose our second regularizer-minimizing $I(\boldsymbol{Z^X};\boldsymbol{Z_t^Y})$-ensuring that $\boldsymbol{Z_t^Y}$ remains disentangled and dedicated to cross-domain information. We further propose *a generalizable proposition* that simplifies intractable disentangling regularizers by breaking them down into manageable components. Distinct from Cao et al. [2022b,a], our *derivation of its optimizable terms* in Section 3.5.1 further ensures the reliability of interest extraction for cold-start users.

Proposition 1. *Let $\boldsymbol{Z^X} \in \mathbb{R}^d$ represent the domain-specific interest in domain X, and $\boldsymbol{Z_t^Y} \in \mathbb{R}^d$ denote the cross-domain transferable interest from domain Y to X. To effectively disentangle domain-specific and cross-domain interests, we aim to minimize their mutual information, $I(\boldsymbol{Z^X};\boldsymbol{Z_t^Y})$, which is equivalent to:*

$$\max\Big\{-I(\boldsymbol{X};\boldsymbol{Z^X}) - I(\boldsymbol{X};\boldsymbol{Z_t^Y}) + I(\boldsymbol{X};\boldsymbol{Z^X},\boldsymbol{Z_t^Y})\Big\}. \tag{6}$$

The proof is given in the Appendix A.1. Eq. (6) above can be intuitively explained as follows: $\boldsymbol{Z^X}$ and $\boldsymbol{Z_t^Y}$ are required to be jointly informative to domain X (the third term), and the total amount of information in $\boldsymbol{Z_t^Y}$ and $\boldsymbol{Z^X}$ will be penalized (the first and second terms). Thus, maximizing Eq. (6) will naturally encourage $\boldsymbol{Z_t^Y}$ and $\boldsymbol{Z^X}$ to encode the distinct, non-overlapping information that can be informative to domain X.

3.4.3 Denoising Regularizer $I(\boldsymbol{Z^X};\boldsymbol{Z_a^X})$

The third regularizer aims to maximize the mutual information $I(\boldsymbol{Z^X};\boldsymbol{Z_a^X})$ between the representations of S^X and $\widetilde{S}^X$. This ensures that the pseudo-sequences retain relevant intra-domain interest information while filtering out noise introduced by unreliable recalled items from PSG. Specifically, maximizing $I(\boldsymbol{Z^X};\boldsymbol{Z_a^X})$ reduces the uncertainty of $\boldsymbol{Z^X}$ given $\boldsymbol{Z_a^X}$, encouraging $\boldsymbol{Z_a^X}$ to capture the true underlying information in $\boldsymbol{Z^X}$ and eliminate irrelevant noise. When user behavior is sparse, it is crucial to leverage rich information from the pseudo-sequence rather than excessively denoise it. Conversely, when user behavior is abundant, enhancing the denoising process improves the model's ability to capture useful augmentations. Therefore, we introduce a noise-adaptive weight λ_d^X for the denoising regularizer $I(\boldsymbol{Z^X};\boldsymbol{Z_a^X})$, which increases with the richness of user interactions: $\lambda_d^X = \exp(aL^X/T) - b$, where L^X is the length

of the user's unpadded historical behavior in domain X, and T is the maximum length of S^X. Constants a and b both set to 0.8 in our study.

3.5 OPTIMIZATION DETAILS

3.5.1 Derivation of Regularizers' Lower Bounds

Combining the previous three interest-enhancing regularizers for domain X, we have the following optimization objectives to enforce robustness:

$$\max \left\{ \underbrace{\overbrace{I(\boldsymbol{X};\boldsymbol{Y};\boldsymbol{Z_t^Y})}^{\text{Cross-domain}} + \overbrace{[-I(\boldsymbol{Z^X};\boldsymbol{Z_t^Y})]}^{\text{Disentangling}} + \overbrace{I(\boldsymbol{Z^X};\boldsymbol{Z_a^X})}^{\text{Denoising}}}_{I^X_{\text{interest}}} \right\}, \quad (7)$$

with the derivations in Eq. (5) - (6), it is equivalent to:

$$\max \left\{ \underbrace{I(\boldsymbol{X};\boldsymbol{Z^X},\boldsymbol{Z_t^Y}) - I(\boldsymbol{X};\boldsymbol{Z^X}) - I(\boldsymbol{X};\boldsymbol{Z_t^Y}|\boldsymbol{Y}) + I(\boldsymbol{Z^X};\boldsymbol{Z_a^X})}_{I^X_{\text{interest}}} \right\}. \quad (8)$$

Although mutual information possesses elegant mathematical properties, the terms above cannot be directly optimized, requiring innovative derivations to *establish tractable and reliable lower bounds*. Below, we provide derivations and explanations for bounds $\mathcal{L}_{(\cdot)}$.

- **Derivations for** $I(\boldsymbol{X};\boldsymbol{Z^X},\boldsymbol{Z_t^Y}) \geq \mathcal{L}_{I(\boldsymbol{X};\boldsymbol{Z^X},\boldsymbol{Z_t^Y})}$:

$$I(\boldsymbol{X};\boldsymbol{Z^X},\boldsymbol{Z_t^Y}) \geq H(\boldsymbol{X}) + \mathbb{E}_{p_D q(z^x|x) q(z_t^y|x,y)} \left[\log p(x|z^x, z_t^y)\right] \quad (9)$$

The detailed derivation is provided in Section A.2. Maximizing Eq. (9) plays a critical role in aligning $p(x|z^x, z_t^y)$ with $q(x|z^x, z_t^y)$, effectively functioning as a reconstruction term. This alignment ensures that the learned latent representations $\boldsymbol{Z^X}$ and $\boldsymbol{Z_t^Y}$ are capable of accurately reconstructing the original user preferences $\boldsymbol{X}$. To avoid redundancy in the objectives, we approximate this term in practice with $\mathbb{E}_{q_\phi}[\log p(x|z^x, z_t^y, z_a^x)]$ from Eq. (4). This approximation not only avoids overlapping objectives but also simplifies computation. For clarity, we refer to this term as $\mathcal{L}_{I(\boldsymbol{X};\boldsymbol{Z^X},\boldsymbol{Z_t^Y})}$.

- **Derivations for** $-I(\boldsymbol{X};\boldsymbol{Z^X}) \geq \mathcal{L}_{-I(\boldsymbol{X};\boldsymbol{Z^X})}$:

$$\begin{aligned} -I(\boldsymbol{X};\boldsymbol{Z^X}) &= -\mathbb{E}_{p_D(x)} \left[D_{\text{KL}} \left[q(z^x|x) \| q(z^x)\right]\right] \\ &\geq -\mathbb{E}_{p_D(x)} \left[D_{\text{KL}} \left[q(z^x|x) \| p(z^x)\right]\right]. \end{aligned} \quad (10)$$

This formula represents reducing the mutual information between $\boldsymbol{X}$ and its latent representation $\boldsymbol{Z^X}$ to ensure that $\boldsymbol{Z^X}$ retains only the necessary information about $\boldsymbol{X}$. By replacing $q(z^x)$ with the prior distribution $p(z^x)$, a more tractable KL divergence lower bound is obtained for practical model training. We denote the derived term as $\mathcal{L}_{-I(\boldsymbol{X};\boldsymbol{Z^X})}$.

- **Derivations for** $-I(\boldsymbol{X};\boldsymbol{Z_t^Y}|\boldsymbol{Y}) \geq \mathcal{L}_{-I(\boldsymbol{X};\boldsymbol{Z_t^Y}|\boldsymbol{Y})}$:

$$-I(\boldsymbol{X};\boldsymbol{Z_t^Y}|\boldsymbol{Y}) \geq -\mathbb{E}_{p_D(x,y)} \left[D_{KL}\left(q(z_t^y|x,y) \| r^y(z_t^y|y)\right)\right] \quad (11)$$

The derived bound is denoted as $\mathcal{L}_{-I(\boldsymbol{X};\boldsymbol{Z_t^Y}|\boldsymbol{Y})}$, with detailed derivations provided in Section A.3. Minimizing $-I(\boldsymbol{X};\boldsymbol{Z_t^Y} \mid \boldsymbol{Y})$ aligns the auxiliary distribution $r^y(z_t^y|y)$ with the cross-domain representation $q(z_t^y|x,y)$. This is particularly useful for cold-start users in domain X, where $q(z_t^y|x,y)$ becomes unreliable. In such cases, as illustrated in Figure 2, $r^y(z_t^y|y)$, relying solely on domain Y, substitutes $q(z_t^y|x,y)$ during inference to ensure the robustness for cold-start users.

- **Derivations for** $I(\boldsymbol{Z^X};\boldsymbol{Z_a^X}) \geq \mathcal{L}_{I(\boldsymbol{Z^X};\boldsymbol{Z_a^X})}$:

$$I(\boldsymbol{Z^X};\boldsymbol{Z_a^X}) = -\mathbb{E}_{q(z_a^x|z^x,x)} \left[D_{KL}\left(q(z^x|x) \| q(z_a^x|x)\right)\right] + \epsilon. \quad (12)$$

Detailed derivation can be found in Section A.4. Since ϵ is intractable, we optimize $I(\boldsymbol{Z^X};\boldsymbol{Z_a^X})$ by solely maximizing the first term, denoted as the lower bound $\mathcal{L}_{I(\boldsymbol{Z^X};\boldsymbol{Z_a^X})}$.

3.5.2 Overall Optimization Objectives

The interest-enhancing regularizers for the reconstruction of domain Y are symmetrical. Referring to Eqs. (9)-(12), we can obtain optimization objectives for domain Y: $\mathcal{L}_{I(\boldsymbol{Y};\boldsymbol{Z^Y},\boldsymbol{Z_t^X})}$, $\mathcal{L}_{-I(\boldsymbol{Y};\boldsymbol{Z^Y})}$, $\mathcal{L}_{-I(\boldsymbol{Y};\boldsymbol{Z_t^X}|\boldsymbol{X})}$, and $\mathcal{L}_{I(\boldsymbol{Z^Y};\boldsymbol{Z_a^Y})}$.

To jointly optimize domains X and Y, we derive the overall optimization objective by combining the ELBO in Eq. (4) with the tractable lower bounds of $I^{(\cdot)}_{\text{interest}}$ for each domain X and Y. This is achieved using balancing weights λ_a and adaptive denoising weight $\lambda_d^{(\cdot)}$ for the denoising regularizer and λ_c for the cross-domain and disentangling regularizers. The overall optimization objective is formulated as follows.

$$\begin{aligned} &\max_{\theta,\phi} \left\{ \log p(x,y) + I^X_{\text{interest}} + I^Y_{\text{interest}} \right\} \\ \geq &\max_{\theta,\phi} \Big\{ \text{ELBO} + \lambda_a \lambda_d^X \mathcal{L}_{I(\boldsymbol{Z^X};\boldsymbol{Z_a^X})} + \lambda_a \lambda_d^Y \mathcal{L}_{I(\boldsymbol{Z^Y};\boldsymbol{Z_a^Y})} \\ &\quad + \lambda_c \big(\mathcal{L}_{I(\boldsymbol{X};\boldsymbol{Z^X},\boldsymbol{Z_t^Y})} + \mathcal{L}_{-I(\boldsymbol{X};\boldsymbol{Z^X})} + \mathcal{L}_{-I(\boldsymbol{X};\boldsymbol{Z_t^Y}|\boldsymbol{Y})} \big) \\ &\quad + \lambda_c \big(\mathcal{L}_{I(\boldsymbol{Y};\boldsymbol{Z^Y},\boldsymbol{Z_t^X})} + \mathcal{L}_{-I(\boldsymbol{Y};\boldsymbol{Z^Y})} + \mathcal{L}_{-I(\boldsymbol{Y};\boldsymbol{Z_t^X}|\boldsymbol{X})} \big) \Big\}. \end{aligned}$$

4 EXPERIMENTS

In this section, we conduct experiments to evaluate the performance of our i^{2}VAE. Experiments in this section intend to answer the following research questions (RQs): **RQ1**: How does i^{2}VAE perform compared to other baseline methods in the CDSR task across different user types, including long-tailed and cold-start users? **RQ2**: How do the different modules of i^{2}VAE contribute to the performance improvement of our method? **RQ3**: Can i^{2}VAE consistently achieve strong performance across varying user-item interaction densities and different numbers of overlapping users, and how do hyperparameter settings impact its performance?

Table 1: Experimental Results (%) across different types of users, including long-tailed(tailed), cold-start, and all users, on "Cloth-Sport" and "Phone-Elec" CDSR datasets. Due to space constraints, the full experimental results, including the "Game-Video" dataset, are presented in Appendix E. The **best** and **second-best** average performances are highlighted.

Datasets	User Types	Metric	SDR			CDR-sequential			CDR			Ours	↑(%)
			Multi-VAE	SVAE	SASRec	DASL	PiNet	C^2DSR	DisenCDR	SA-VAE	CDRIB	i^2VAE	
Cloth	Tailed	NDCG	2.31±0.08	2.07±0.16	2.09±0.20	2.28±0.17	2.11±0.17	2.29±0.09	2.20±0.06	**2.40±0.14**	2.27±0.10	**2.52±0.07***	5.00
		HR	4.15±0.12	3.99±0.34	3.99±0.31	**4.46±0.24**	4.15±0.36	4.38±0.32	4.21±0.15	4.43±0.22	4.21±0.21	**4.73±0.16***	6.05
	Cold-start	NDCG	3.23±0.29	**3.41±0.35**	2.86±0.50	3.28±0.18	3.04±0.66	3.08±0.65	2.95±0.29	3.25±0.24	3.00±0.39	**3.45±0.35***	1.17
		HR	6.08±0.38	**6.39±0.43**	5.64±0.86	6.27±0.40	5.58±1.24	5.96±0.95	5.58±0.23	5.89±0.46	5.52±0.58	**6.77±0.51***	5.95
	All	NDCG	2.20±0.10	2.18±0.10	2.11±0.15	2.43±0.09	2.18±0.10	2.40±0.07	2.30±0.05	**2.52±0.10**	2.39±0.06	**2.59±0.11***	2.78
		HR	4.25±0.12	4.28±0.28	4.15±0.30	**4.82±0.11**	4.24±0.18	4.63±0.18	4.48±0.18	4.81±0.19	4.56±0.09	**5.00±0.17***	3.73
Sport	Tailed	NDCG	3.02±0.14	2.90±0.16	2.80±0.21	**3.31±0.36**	2.96±0.15	3.06±0.15	3.14±0.10	3.29±0.15	3.12±0.11	**3.36±0.10***	1.51
		HR	6.06±0.50	5.76±0.34	5.72±0.22	6.29±0.57	5.89±0.25	6.03±0.22	5.94±0.19	**6.31±0.42**	5.81±0.10	**6.66±0.17***	5.55
	Cold-start	NDCG	4.33±0.49	4.19±0.24	3.92±0.40	4.48±0.42	4.24±0.31	4.65±0.53	4.93±0.18	**5.05±0.29**	4.95±0.25	**5.43±0.29***	7.52
		HR	8.89±0.66	8.62±0.73	7.88±0.62	8.15±0.94	8.28±0.27	9.36±0.54	8.54±0.59	**9.63±0.73**	9.09±0.43	**10.30±0.46***	6.96
	All	NDCG	3.79±0.07	3.89±0.15	3.70±0.18	3.93±0.14	3.68±0.08	4.13±0.17	4.21±0.13	**4.31±0.13**	4.27±0.09	**4.40±0.10***	2.20
		HR	7.23±0.27	7.27±0.29	7.00±0.17	7.45±0.32	6.90±0.33	7.84±0.33	7.90±0.22	**8.06±0.26**	7.88±0.33	**8.53±0.14***	5.80
Phone	Tailed	NDCG	3.58±0.13	3.51±0.11	3.41±0.13	**4.10±0.15**	3.68±0.17	3.81±0.17	4.01±0.14	4.06±0.31	4.01±0.23	**4.23±0.21***	3.17
		HR	6.90±0.34	6.74±0.20	6.58±0.32	**8.06±0.27**	7.17±0.20	7.47±0.39	7.78±0.39	7.96±0.61	7.92±0.38	**8.17±0.27***	1.36
	Cold-start	NDCG	3.16±0.19	3.15±0.30	2.97±0.38	3.63±0.25	3.51±0.38	3.73±0.37	3.80±0.33	3.83±0.48	**4.00±0.35**	**4.39±0.21***	9.75
		HR	6.26±0.39	6.11±0.42	6.03±0.45	7.40±0.39	7.18±0.51	7.18±0.74	7.56±0.56	**7.94±0.85**	7.56±0.51	**8.63±0.19***	8.69
	All	NDCG	3.98±0.19	3.89±0.06	3.88±0.14	4.40±0.17	4.13±0.14	4.36±0.21	**4.52±0.15**	4.49±0.29	4.46±0.19	**5.79±0.29***	3.23
		HR	7.59±0.39	7.33±0.14	7.36±0.27	8.54±0.32	7.77±0.34	8.30±0.38	**8.68±0.22**	8.54±0.50	8.66±0.29	**10.61±0.22***	2.35
Elec	Tailed	NDCG	6.96±0.23	6.74±0.25	6.78±0.29	7.63±0.19	7.09±0.24	7.78±0.13	7.64±0.10	7.60±0.30	**7.77±0.11**	**8.00±0.10***	2.96
		HR	11.65±0.48	11.39±0.47	11.49±0.53	**12.83±0.41**	11.66±0.56	13.05±0.28	12.45±0.25	12.56±0.39	12.66±0.28	**13.49±0.19***	5.14
	Cold-start	NDCG	9.35±0.33	9.22±0.19	9.16±0.19	9.73±0.65	9.59±0.35	9.85±0.30	**9.90±0.25**	9.76±0.48	9.73±0.40	**10.26±0.42***	3.64
		HR	14.71±0.49	14.76±0.60	14.94±0.51	15.76±1.24	15.00±1.00	**15.94±0.43**	15.24±0.60	15.35±0.90	15.53±0.34	**17.12±0.39***	7.40
	All	NDCG	8.20±0.22	8.06±0.27	8.08±0.34	8.58±0.16	7.84±0.08	9.00±0.12	8.95±0.13	8.84±0.19	**9.04±0.04**	**9.33±0.09***	3.22
		HR	13.08±0.43	12.81±0.45	12.86±0.59	13.60±0.50	12.31±0.19	**14.46±0.31**	14.03±0.21	13.96±0.26	14.08±0.26	**15.28±0.26***	5.64

"*" denotes statistically significant improvements ($p < 0.05$), as determined by a paired t-test comparison with the second best result.

4.1 DATASETS

Following previous works [Cao et al., 2022c, Xu et al., 2023c, Cao et al., 2022b, Xu et al., 2023b], we conducted offline experiments on three widely used CDSR datasets from Amazon across six domains: "Cloth-Sport", and "Phone-Elec", and "Game-Video". All behavioral sequences were collected in chronological order, and the data were split into 80% training, 10% validation, and 10% testing. To simulate real-world recommendation scenarios, we included non-overlapping users and controlled the overlapping ratio ($\mathcal{K}_o$) across domains. Detailed dataset information and statistics are provided in Appendix B.

4.2 EXPERIMENT SETTING

Evaluation Protocol. To assess our approach across different user types, we retain all overlapping ($\mathcal{K}_o = 100\%$) and non-overlapping users in the training set, while randomly selecting 20% of overlapping users from the test set and treating them as cold-start users. We evaluate performance on long-tailed users (interaction sequences shorter than the bottom 80% average), cold-start users, and all users [Ma et al., 2019, Cao et al., 2022c,a]. For fair comparison [Krichene and Rendle, 2020, Zhao et al., 2020], we construct a ranking candidate set by sampling 999 negative items per user along with one positive ground-truth item. Performance is measured by NDCG@10 [Järvelin and Kekäläinen, 2002] and HR@10, where higher values indicate better performance. Implementation details are provided in Appendix D.

Compared Methods. To verify the effectiveness of our model, we compare i^2VAE with the following SOTA baselines which can be divided into three branches including: (1) single-domain recommendation methods (SDR), i.e., Multi-VAE [Liang et al., 2018], SVAE [Sachdeva et al., 2019] and SASRec [Kang and McAuley, 2018]. (2) cross-domain sequential recommendation methods (CDSR), i.e., Pi-Net [Ma et al., 2019], DASL [Li et al., 2021] and C^2DSR [Cao et al., 2022a]. (3) cross-domain recommendation methods (CDR), i.e., DisenCDR [Cao et al., 2022b], SA-VAE [Salah et al., 2021] and CDRIB [Cao et al., 2022c]. Details are in Appendix C.

4.3 COMPARISON RESULTS (RQ1)

We compare the performance of i^2VAE with state-of-the-art baselines across multiple real-world CDSR datasets, as shown in Table 1. While the extent of baseline improvements varies across datasets and user types, i^2VAE consistently outperforms previous methods for long-tailed, cold-start, and all users. In general, CDR/CDSR methods outperform

Table 2: Ablation study results (%) on the "Cloth-Sport" dataset. "w/o" indicates the removal of the module.

Domain	User Types	Metric	Model Variants			i^{2}VAE
			w/o PSG	w/o CD-R&DS-R	w/o DN-R	
Cloth	Tailed	NDCG	2.41±0.14	2.38±0.09	2.39±0.14	**2.52 ±0.07**
		HR	4.59±0.12	4.49±0.21	4.45±0.21	**4.73 ±0.16**
	Cold-Start	NDCG	3.21±0.22	3.23±0.12	3.30±0.17	**3.45 ±0.35**
		HR	6.21±0.23	6.27±0.34	6.46±0.58	**6.77 ±0.51**
	All	NDCG	2.54±0.12	2.54±0.11	2.54±0.12	**2.59 ±0.11**
		HR	4.92±0.20	4.84±0.26	4.89±0.23	**5.00 ±0.17**
Sport	Tailed	NDCG	3.35±0.12	3.28±0.11	3.27±0.07	**3.36 ±0.10**
		HR	6.51±0.24	6.39±0.16	6.33±0.18	**6.66 ±0.17**
	Cold-Start	NDCG	5.26±0.26	5.18±0.30	5.19±0.28	**5.43 ±0.29**
		HR	10.24±0.46	9.90±0.34	10.10±0.43	**10.30 ±0.46**
	All	NDCG	4.36±0.05	4.31±0.16	4.32±0.15	**4.40 ±0.10**
		HR	8.44±0.14	8.41±0.25	8.44±0.24	**8.53 ±0.14**

SDR models by capturing cross-domain interests, which is particularly beneficial in sparse data settings. Among the most challenging groups, long-tailed users see only limited improvements from most baselines, as their sparse interaction histories constrain learning capacity. In contrast, i^{2}VAE surpasses the second-best model by 0.43% to 5.00% by synthesizing pseudo-sequences and applying denoising regularization to filter out irrelevant information, thereby enhancing recommendation quality. Similarly, for cold-start users, i^{2}VAE achieves significant performance gains of 1.17% to 15.20% over the second-best model by leveraging $r^y(z_t^y|y)$ and $r^x(z_t^x|x)$ to learn cross-domain representations for both overlapping and non-overlapping users while mitigating negative transfer caused by domain-specific information, further refining inference quality. These results highlight i^{2}VAE's robustness across different user groups and its ability to effectively model interest representations in long-tailed and cold-start scenarios.

4.4 ABLATION STUDY (RQ2)

We assess the importance of each module by examining their impact on performance. Analysis shows that replacing PSG-generated pseudo-sequences with random ones decreases performance across all user types, confirming that even imperfect recall models enhance user interest understanding. Removing informative and disentangle regularizers (CD-R & DS-R) similarly reduces performance, with the strongest impact on cold-start users, highlighting the importance of structured representations and non-overlapping user training. Finally, removing DN-R degrades performance, particularly for long-tailed users, demonstrating the need to denoise pseudo sequences while leveraging augmented interests.

4.5 MODEL ANALYSIS (RQ3)

To verify the performance of i^{2}VAE in CDSR scenarios with varying data densities, we conduct studies by varying the data density D_s in {25%, 50%, 75%, 100%}. As the density decreases, the actual user-item interaction records in the training and testing sets are down-sampled to test the robustness of i^{2}VAE and the second-best SA-VAE on sparser datasets. We re-run experiments on the 'Cloth-Sport' dataset with other settings as in Sections 3.2.1 and 3.2.3. The results are presented in Table 3. All experiments are conducted five times with different random seeds, and average values are reported. As expected, both models' performance decreases with lower data density due to the increased challenge in interest learning. SA-VAE shows a significant performance decline because it relies heavily on rich user interactions within each domain. In contrast, i^{2}VAE generally achieves better recommendation results with less performance degradation. This is mainly due to the PSG and the denoise regularizer supplementing the sparse interaction data. We also designed experiments to test our model's performance with fewer cross-domain overlapping users, and the results can be found in Appendix F.1. Moreover, the investigate results of the parameter sensitivity of sequence length T and the harmonic factors λ_a and λ_c can be found in Appendix F.2. These additional experiments confirm that i^{2}VAE remains effective across different levels of overlapping users and exhibits stable performance under hyperparameter variations.

5 RELATED WORK

Cross-Domain Sequential Recommendation [Ma et al., 2019, Sun et al., 2021, Xu et al., 2023c] aims to improve sequential recommendation (SR) performance by utilizing user behavior sequences from multiple related domains. PiNet[Ma et al., 2019] and PSJNet [Sun et al., 2021] design gating mechanism to learn and transfer cross-domain information on overlapping users. The attentive learning-based model DASL [Li et al., 2021] uses dual attentive learning to transfer the user's latent interests bidirectionally across two domains. Similarly, DA-GCN[Guo et al., 2021] and MIFN[Ma et al., 2022] build user-item bipartite graphs to facilitate cross-domain information transferring on overlapping users. Moreover, C^{2}DSR [Cao et al., 2022a] employs GNNs as sequential attentive encoder to learn the collaborative signals and utilize contrastive learning to align single- and cross-domain user representations. However, these methods heavily depend on overlapping users, limiting their effectiveness for long-tailed and cold-start users.

Cross-Domain Recommendation [Zhu et al., 2020a, Zhang et al., 2018, Zhao et al., 2019, Zhu et al., 2020b, Ouyang et al., 2020, Salah et al., 2021] leverages multi-domain behavior patterns to address data sparsity and cold-start issues in single-domain recommendation. Recent studies focus on transfer learning [Hu et al., 2018, Liu et al., 2020], using transfer modules to map and fuse representations across domains, or modeling domain-shared information [Cao et al., 2022a,b]. DisenCDR [Cao et al., 2022b] assumes fully overlapping users and employs mutual-information-based regularizers to disentangle domain-specific and domain-shared

Table 3: Experiment results (%) on "Cloth-Sport" dataset with different density (D_s).

Domain	User	Metric	$D_s = 25\%$		$D_s = 50\%$		$D_s = 75\%$		$D_s = 100\%$	
			SA-VAE	i^{2}VAE	SA-VAE	i^{2}VAE	SA-VAE	i^{2}VAE	SA-VAE	i^{2}VAE
Cloth	Tailed	NDCG	1.39	**1.51**	1.40	**1.53**	1.40	**1.46**	2.40	**2.52**
		HR	2.75	**2.96**	2.69	**2.88**	2.85	**2.95**	4.43	**4.73**
	Cold-Start	NDCG	0.84	**0.85**	0.90	**1.12**	1.41	**1.72**	3.25	**3.45**
		HR	1.94	**1.82**	2.26	**2.26**	2.82	**3.32**	5.89	**6.77**
	All	NDCG	1.38	**1.49**	1.39	**1.53**	1.42	**1.57**	2.52	**2.59**
		HR	2.74	**2.95**	2.71	**2.92**	2.95	**3.27**	4.81	**5.00**
Sport	Tailed	NDCG	2.35	**2.53**	2.47	**2.72**	2.80	**3.26**	3.29	**3.36**
		HR	4.27	**4.62**	4.41	**5.09**	4.86	**5.66**	6.31	**6.66**
	Cold-Start	NDCG	0.71	**0.76**	0.99	**1.06**	1.88	**2.14**	5.05	**5.43**
		HR	1.62	**1.67**	2.02	**2.02**	3.57	**3.84**	9.63	**10.30**
	All	NDCG	2.49	**2.70**	3.03	**3.28**	3.18	**3.70**	4.31	**4.40**
		HR	4.46	**4.84**	5.47	**6.16**	5.62	**6.60**	8.06	**8.53**

interests, but fails to handle cold-start users effectively. SA-VAE [Salah et al., 2021] pre-trains a VAE on the source domain and aligns latent variables between source and target domain VAEs, but rely fully on the overlappping users. CDRIB [Cao et al., 2022c] applies the information bottleneck principle and contrastive learning for overlapping users but lacks generalizability for long-tailed users and compresses the rich inner-domain interest data, limiting its robustness compared to our variational regularizers.

6 CONCLUSION

In this paper, we propose i^{2}VAE to enhance the performance of long-tailed and cold-start users. Our model introduces interest-enhancing regularizers, which enable the learning of distinct inner- and cross-domain interests and extract relevant information from pseudo-sequences to enrich users' sparse interaction. Empirical experiments demonstrate that i^{2}VAE achieves SOTA performance across all user types.

References

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *International conference on machine learning*, pages 531–540. PMLR, 2018.

Anthony J Bell. The co-information lattice. In *Proceedings of the fifth international workshop on independent component analysis and blind signal separation: ICA*, volume 2003. Citeseer, 2003.

Jiangxia Cao, Xin Cong, Jiawei Sheng, Tingwen Liu, and Bin Wang. Contrastive cross-domain sequential recommendation. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 138–147, 2022a.

Jiangxia Cao, Xixun Lin, Xin Cong, Jing Ya, Tingwen Liu, and Bin Wang. Disencdr: Learning disentangled representations for cross-domain recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 267–277, 2022b.

Jiangxia Cao, Jiawei Sheng, Xin Cong, Tingwen Liu, and Bin Wang. Cross-domain recommendation to cold-start users via variational information bottleneck. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, pages 2209–2223. IEEE, 2022c.

Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

Thomas M Cover, Joy A Thomas, et al. Entropy, relative entropy and mutual information. *Elements of information theory*, 2(1):12–13, 1991.

Lei Guo, Li Tang, Tong Chen, Lei Zhu, Quoc Viet Hung Nguyen, and Hongzhi Yin. Da-gcn: A domain-aware attentive graph convolution network for shared-account cross-domain sequential recommendation. *arXiv preprint arXiv:2105.03300*, 2021.

Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 639–648, 2020.

Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv:1511.06939*, 2015.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

Guangneng Hu, Yu Zhang, and Qiang Yang. Conet: Collaborative cross networks for cross-domain recommendation. In *Proceedings of the 27th ACM international conference on information and knowledge management*, pages 667–676, 2018.

Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002.

SeongKu Kang, Junyoung Hwang, Dongha Lee, and Hwanjo Yu. Semi-supervised learning for cross-domain recommendation to cold-start users. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 1563–1572, 2019.

Wang-Cheng Kang and Julian McAuley. Self-attentive sequential recommendation. In *2018 IEEE international conference on data mining (ICDM)*, pages 197–206. IEEE, 2018.

Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Walid Krichene and Steffen Rendle. On sampled metrics for item recommendation. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1748–1757, 2020.

Pan Li, Zhichao Jiang, Maofei Que, Yao Hu, and Alexander Tuzhilin. Dual attentive sequential learning for cross-domain click-through rate prediction. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 3172–3180, 2021.

Dawen Liang, Rahul G Krishnan, Matthew D Hoffman, and Tony Jebara. Variational autoencoders for collaborative filtering. In *Proceedings of the 2018 world wide web conference*, pages 689–698, 2018.

Xixun Lin, Jia Wu, Chuan Zhou, Shirui Pan, Yanan Cao, and Bin Wang. Task-adaptive neural process for user cold-start recommendation. In *Proceedings of the Web Conference 2021*, pages 1306–1316, 2021.

Meng Liu, Jianjun Li, Guohui Li, and Peng Pan. Cross domain recommendation via bi-directional transfer graph collaborative filtering networks. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 885–894, 2020.

Muyang Ma, Pengjie Ren, Yujie Lin, Zhumin Chen, Jun Ma, and Maarten de Rijke. π-net: A parallel information-sharing network for shared-account cross-domain sequential recommendations. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pages 685–694, 2019.

Muyang Ma, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Lifan Zhao, Peiyu Liu, Jun Ma, and Maarten de Rijke. Mixed information flow for cross-domain sequential recommendations. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(4):1–32, 2022.

Tong Man, Huawei Shen, Xiaolong Jin, and Xueqi Cheng. Cross-domain recommendation: An embedding and mapping approach. In *IJCAI*, volume 17, pages 2464–2470, 2017.

William McGill. Multivariate information transmission. *Transactions of the IRE Professional Group on Information Theory*, 4(4):93–111, 1954.

Wentao Ouyang, Xiuwu Zhang, Lei Zhao, Jinmei Luo, Yu Zhang, Heng Zou, Zhaojie Liu, and Yanlong Du. Minet: Mixed interest network for cross-domain click-through rate prediction. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 2669–2676, 2020.

Liam Paninski. Estimation of entropy and mutual information. *Neural computation*, 15(6):1191–1253, 2003.

Chung Park, Taesan Kim, Taekyoon Choi, Junui Hong, Yelim Yu, Mincheol Cho, Kyunam Lee, Sungil Ryu, Hyungjun Yoon, Minsung Choi, et al. Cracking the code of negative transfer: A cooperative game theoretic approach for cross-domain sequential recommendation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 2024–2033, 2023.

Alfréd Rényi. On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1: contributions to the theory of statistics*, volume 4, pages 547–562. University of California Press, 1961.

Noveen Sachdeva, Giuseppe Manco, Ettore Ritacco, and Vikram Pudi. Sequential variational autoencoders for collaborative filtering. In *Proceedings of the twelfth ACM international conference on web search and data mining*, pages 600–608, 2019.

Aghiles Salah, Thanh Binh Tran, and Hady Lauw. Towards source-aligned variational models for cross-domain recommendation. In *Proceedings of the 15th ACM Conference on Recommender Systems*, pages 176–186, 2021.

Jiarui Shi and Quanmin Wang. Cross-domain variational autoencoder for recommender systems. In *2019 IEEE 11th International Conference on Advanced Infocomm Technology (ICAIT)*, pages 67–72. IEEE, 2019.

Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 1441–1450, 2019.

Wenchao Sun, Muyang Ma, Pengjie Ren, Yujie Lin, Zhumin Chen, Zhaochun Ren, Jun Ma, and Maarten De Rijke. Parallel split-join networks for shared account cross-domain sequential recommendations. *IEEE Transactions on Knowledge and Data Engineering*, 35(4):4106–4123, 2021.

Jiaxi Tang and Ke Wang. Personalized top-n sequential recommendation via convolutional sequence embedding. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 565–573, 2018.

Jianling Wang, Kaize Ding, Liangjie Hong, Huan Liu, and James Caverlee. Next-item recommendation with sequential hypergraphs. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pages 1101–1110, 2020.

Tianxin Wei and Jingrui He. Comprehensive fair meta-learned recommender system. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1989–1999, 2022.

Tianxin Wei, Bowen Jin, Ruirui Li, Hansi Zeng, Zhengyang Wang, Jianhui Sun, Qingyu Yin, Hanqing Lu, Suhang Wang, Jingrui He, et al. Towards unified multi-modal personalization: Large vision-language models for generative recommendation and beyond. *arXiv preprint arXiv:2403.10667*, 2024.

Wujiang Xu, Shaoshuai Li, Mingming Ha, Xiaobo Guo, Qiongxu Ma, Xiaolei Liu, Linxun Chen, and Zhenfeng Zhu. Neural node matching for multi-target cross domain recommendation. *arXiv preprint arXiv:2302.05919*, 2023a.

Wujiang Xu, Xuying Ning, Wenfang Lin, Mingming Ha, Qiongxu Ma, Linxun Chen, Bing Han, and Minnan Luo. Towards open-world cross-domain sequential recommendation: A model-agnostic contrastive denoising approach. *arXiv preprint arXiv:2311.04760*, 2023b.

Wujiang Xu, Qitian Wu, Runzhong Wang, Mingming Ha, Qiongxu Ma, Linxun Chen, Bing Han, and Junchi Yan. Rethinking cross-domain sequential recommendation under open-world assumptions. *arXiv preprint arXiv:2311.04590*, 2023c.

Wujiang Xu, Zujie Liang, Jiaojiao Han, Xuying Ning, Wenfang Lin, Linxun Chen, Feng Wei, and Yongfeng Zhang. Slmrec: empowering small language models for sequential recommendation. *arXiv e-prints*, pages arXiv–2405, 2024a.

Wujiang Xu, Xuying Ning, Wenfang Lin, Mingming Ha, Qiongxu Ma, Qianqiao Liang, Xuewen Tao, Linxun Chen, Bing Han, and Minnan Luo. Towards open-world cross-domain sequential recommendation: A model-agnostic contrastive denoising approach. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 161–179. Springer, 2024b.

Qian Zhang, Dianshuang Wu, Jie Lu, and Guangquan Zhang. Cross-domain recommendation with probabilistic knowledge transfer. In *International Conference on Neural Information Processing*, pages 208–219. Springer, 2018.

Wei Zhang, Pengye Zhang, Bo Zhang, Xingxing Wang, and Dong Wang. A collaborative transfer learning framework for cross-domain recommendation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5576–5585, 2023.

Cheng Zhao, Chenliang Li, and Cong Fu. Cross-domain recommendation via preference propagation graphnet. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 2165–2168, 2019.

Wayne Xin Zhao, Junhua Chen, Pengfei Wang, Qi Gu, and Ji-Rong Wen. Revisiting alternative experimental settings for evaluating top-n item recommendation algorithms. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2329–2332, 2020.

Feng Zhu, Yan Wang, Chaochao Chen, Guanfeng Liu, Mehmet Orgun, and Jia Wu. A deep framework for cross-domain and cross-system recommendations. *arXiv preprint arXiv:2009.06215*, 2020a.

Feng Zhu, Yan Wang, Chaochao Chen, Guanfeng Liu, and Xiaolin Zheng. A graphical and attentional framework for dual-target cross-domain recommendation. In *IJCAI*, pages 3001–3008, 2020b.

Yongchun Zhu, Kaikai Ge, Fuzhen Zhuang, Ruobing Xie, Dongbo Xi, Xu Zhang, Leyu Lin, and Qing He. Transfer-meta framework for cross-domain recommendation to cold-start users. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 1813–1817, 2021.

i^{2}VAE: Interest Information Augmentation with Variational Regularizers for Cross-Domain Sequential Recommendation (Supplementary Material)

Xuying Ning[§1] **Wujiang Xu**[*¶2] **Tianxin Wei**[1] **Xiaolei Liu**[3]

[1]University of Illinois Urbana-Champaign, Urbana, IL, USA

[2]Rutgers University, New Brunswick, NJ, USA

[3]Independent Researcher

A THEORETICAL DERIVATION

A.1 PROOF OF PROPOSITION 1

Proposition 2. *Let $\boldsymbol{Z^X} \in \mathbb{R}^d$ represent the domain-specific interest in domain X, and $\boldsymbol{Z_t^Y} \in \mathbb{R}^d$ denote the cross-domain transferable interest from domain Y to X. To effectively disentangle domain-specific and cross-domain interests, we aim to minimize their mutual information, $I(\boldsymbol{Z^X}; \boldsymbol{Z_t^Y})$, which is equivalent to:*

$$\max\Big\{ -I(\boldsymbol{X}; \boldsymbol{Z^X}) - I(\boldsymbol{X}; \boldsymbol{Z_t^Y}) + I(\boldsymbol{X}; \boldsymbol{Z^X}, \boldsymbol{Z_t^Y})\Big\}.$$

Proof. The mutual information between cross-domain representations $I(\boldsymbol{Z^X}; \boldsymbol{Z_t^Y})$ can be decomposed into three components using the chain rule of mutual information as follows. This decomposition provides key insights about information flow across domains.

$$I(\boldsymbol{Z^X}; \boldsymbol{Z_t^Y}) = I(\boldsymbol{Z^X}; \boldsymbol{X}) - I(\boldsymbol{Z^X}; \boldsymbol{X}|\boldsymbol{Z_t^Y}) + I(\boldsymbol{Z^X}; \boldsymbol{Z_t^Y}|\boldsymbol{X})$$

Due to our disentangling assumption, $\boldsymbol{Z^X}$ only represents the domain-specific interest, $q(z^x|x) = q(z^x|x, z_t^y)$ holds. Therefore, the last term above, $I(\boldsymbol{Z^X}; \boldsymbol{Z_t^Y}|\boldsymbol{X})$, vanishes:

$$\begin{aligned} I(\boldsymbol{Z^X}; \boldsymbol{Z_t^Y}|\boldsymbol{X}) &= H(\boldsymbol{Z^X}|\boldsymbol{X}) - H(\boldsymbol{Z^X}|\boldsymbol{X}, \boldsymbol{Z_t^Y}) \\ &= H(\boldsymbol{Z^X}|\boldsymbol{X}) - H(\boldsymbol{Z^X}|\boldsymbol{X}) = 0. \end{aligned}$$

This results in the following derivation of $-I(\boldsymbol{Z^X}; \boldsymbol{Z_t^Y})$, which maintains equality based on the chain rule of mutual information[Cover, 1999]:

$$\begin{aligned} -I(\boldsymbol{Z^X}; \boldsymbol{Z_t^Y}) &= -I(\boldsymbol{X}; \boldsymbol{Z^X}) + I(\boldsymbol{Z^X}; \boldsymbol{X}|\boldsymbol{Z_t^Y}) \\ &= -I(\boldsymbol{X}; \boldsymbol{Z^X}) - I(\boldsymbol{X}; \boldsymbol{Z_t^Y}) + I(\boldsymbol{X}; \boldsymbol{Z^X}, \boldsymbol{Z_t^Y}). \end{aligned}$$

□

This equation illustrates that $\boldsymbol{Z^X}$ and $\boldsymbol{Z_t^Y}$ should collectively provide valuable information for domain X (the third term), while the individual amount of information of $\boldsymbol{Z^X}$ and $\boldsymbol{Z_t^Y}$ are penalized (the first and second terms) to avoid redundancy and ensure non-overlapping information. The proof is completed.

[*]Equal contribution.

[†]Correspondence to: Wujiang Xu <wujiang.xu@rutgers.edu>.

[§]Equal contribution.

[¶]Correspondence to: Wujiang Xu <wujiang.xu@rutgers.edu>.

A.2 DETAILED DERIVATION FOR $\mathcal{L}_{I(\boldsymbol{X};\boldsymbol{Z}^{\boldsymbol{X}},\boldsymbol{Z}_{\boldsymbol{t}}^{\boldsymbol{Y}})}$

We rewrite $I(\boldsymbol{X};\boldsymbol{Z}^{\boldsymbol{X}},\boldsymbol{Z}_{\boldsymbol{t}}^{\boldsymbol{Y}})$ in the form of an expectation and derive its lower bound using the generative distribution $p_\theta(x \mid z^x, z_t^y)$.

$$\begin{aligned}
I(\boldsymbol{X};\boldsymbol{Z}^{\boldsymbol{X}},\boldsymbol{Z}_{\boldsymbol{t}}^{\boldsymbol{Y}}) &= \mathbb{E}_{q(z^x,z_t^y|x)p_D(x)}\left[\log\frac{q(x|z^x,z_t^y)}{p_D(x)}\right] \\
&= H(\boldsymbol{X}) + \mathbb{E}_{q(z^x,z_t^y|x)p_D(x)}\left[\log q(x|z^x,z_t^y)\right] + \mathbb{E}_{q(z^x,z_t^y|x)p_D(x)}\left[\log p(x|z^x,z_t^y) - \log p(x|z^x,z_t^y)\right] \\
&= H(\boldsymbol{X}) + \mathbb{E}_{q(z^x,z_t^y|x)p_D(x)}\left[\log p(x|z^x,z_t^y)\right] + \mathbb{E}_{q(z^x,z_t^y)}\left[D_{KL}\left(q(x|z^x,z_t^y)\|p(x|z^x,z_t^y)\right)\right] \\
&\geq H(\boldsymbol{X}) + \mathbb{E}_{q(z^x,z_t^y|x)p_D(x)}\left[\log p(x|z^x,z_t^y)\right].
\end{aligned} \tag{13}$$

The final inequality in Eq. (A.2) is derived from the non-negativity of the KL divergence. Next, we expand and rewrite the second term in Eq. (A.2) using the integral form of the expectation as follows:

$$\begin{aligned}
\mathbb{E}_{q(z^x,z_t^y|x)p_D(x)}\left[\log p(x|z^x,z_t^y)\right] &= \int q(z^x,z_t^y|x)p_D(x)\log p(x|z^x,z_t^y)\,dx\,dz^x\,dz_t^y \\
&= \int p_D(x)\left(\int q(z^x,z_t^y|x,y)p_D(y|x)\,dy\right)\log p(x|z^x,z_t^y)\,dx\,dz^x\,dz_t^y \\
&= \int p_D(x)q(z^x|x)\left(\int q(z_t^y|x,y)p_D(y|x)dy\right)\log p(x|z^x,z_t^y)dxdz^xdz_t^y \\
&= \int p_D(x,y)q(z^x|x)q(z_t^y|x,y)\log p(x|z^x,z_t^y)\,dx\,dy\,dz^x\,dz_t^y \\
&= \mathbb{E}_{p_D(x,y)q(z^x|x)q(z_t^y|x,y)}\left[\log p(x|z^x,z_t^y)\right].
\end{aligned} \tag{14}$$

We then substitute the derived variant in Eq. (14) back into Eq. (A.2) to obtain the lower bound of $I(\boldsymbol{X};\boldsymbol{Z}^{\boldsymbol{X}},\boldsymbol{Z}_{\boldsymbol{t}}^{\boldsymbol{Y}})$:

$$I(\boldsymbol{X};\boldsymbol{Z}^{\boldsymbol{X}},\boldsymbol{Z}_{\boldsymbol{t}}^{\boldsymbol{Y}}) \geq H(\boldsymbol{X}) + \mathbb{E}_{p_D(x,y)q(z^x|x)q(z_t^y|x,y)}\left[\log p(x|z^x,z_t^y)\right].$$

A.3 DETAILED DERIVATION FOR $\mathcal{L}_{-I(\boldsymbol{X};\boldsymbol{Z}_{\boldsymbol{t}}^{\boldsymbol{Y}}|\boldsymbol{Y})}$

We expand $-I(\boldsymbol{X};\boldsymbol{Z}_{\boldsymbol{t}}^{\boldsymbol{Y}}|\boldsymbol{Y})$ into the form of an expectation. Since $q(z_t^y|y)$ cannot be directly obtained within our framework, we utilize $r^y(z_t^y|y)$ to approximate it.

$$\begin{aligned}
-I(\boldsymbol{X};\boldsymbol{Z}_{\boldsymbol{t}}^{\boldsymbol{Y}}|\boldsymbol{Y}) &= -\mathbb{E}_{p_D(x,y)q(z_t^y|x,y)}\left[\log\frac{q(z_t^y|x,y)}{q(z_t^y|y)}\right] \\
&= -\mathbb{E}_{p_D(x,y)q(z_t^y|x,y)}\left[\log\left\{\frac{q(z_t^y|x,y)}{r^y(z_t^y|y)}\cdot\frac{r^y(z_t^y|y)}{q(z_t^y|y)}\right\}\right] \\
&= -\mathbb{E}_{p_D(x,y)}\left[D_{KL}\left(q(z_t^y|x,y)\|r^y(z_t^y|y)\right)\right] + \mathbb{E}_{p_D(x,y)}\left[D_{KL}\left(q(z_t^y|y)\|r^y(z_t^y|y)\right)\right] \\
&\geq -\mathbb{E}_{p_D(x,y)}\left[D_{KL}\left(q(z_t^y|x,y)\|r^y(z_t^y|y)\right)\right].
\end{aligned} \tag{15}$$

The final inequality in Eq. (15) is derived from the non-negativity of the KL divergence.

A.4 DETAILED DERIVATION FOR $\mathcal{L}_{-I(\boldsymbol{X};\boldsymbol{Z}^{\boldsymbol{X}})}$

We expand and rewrite $I(\boldsymbol{Z}^{\boldsymbol{X}};\boldsymbol{Z}_{\boldsymbol{a}}^{\boldsymbol{X}})$ as follows to extract the optimizable component:

$$\begin{aligned}
I(\boldsymbol{Z}^{\boldsymbol{X}};\boldsymbol{Z}_{\boldsymbol{a}}^{\boldsymbol{X}}) &= \mathbb{E}_{q(z^x,z_a^x|x)}\left[\log\frac{q(z_a^x,z^x|x)}{q(z^x|x)q(z_a^x|x)}\right] \\
&= \mathbb{E}_{q(z^x|x)q(z_a^x|z^x,x)}\left[\log\frac{q(z_a^x|x)}{q(z^x|x)} + \log\frac{q(z^x|z_a^x,x)}{q(z_a^x|x)}\right] \\
&= -\mathbb{E}_{q(z_a^x|z^x,x)}\left[D_{KL}\left(q(z^x|x)\|q(z_a^x|x)\right)\right] + \mathbb{E}_{q(z^x,z_a^x|x)}\left[\log\frac{q(z^x|z_a^x,x)}{q(z_a^x|x)}\right] \\
&= -\mathbb{E}_{q(z_a^x|z^x,x)}\left[D_{KL}\left(q(z^x|x)\|q(z_a^x|x)\right)\right] + \epsilon,
\end{aligned}$$

which is decomposed into two terms, one of which is an optimizable KL divergence. The first term in our framework is optimizable, while the remaining intractable term is denoted as ϵ. We primarily focus on optimizing the first term.

B DATASET

We conducted experiments on the Amazon 14 dataset[1] across six cross-domain pairs: "Cloth-Sport", "Phone-Elec", "Game-Video", following previous works [Cao et al., 2022c, Xu et al., 2023c, Cao et al., 2022b, Xu et al., 2023b]. The dataset statistics are summarized in Table 4. All behavioral sequences were collected in chronological order. To ensure a realistic evaluation setting, we included both non-overlapping and overlapping users, adjusting the overlapping ratio ($\mathcal{K}_o$) to control the number of shared users across domains. Furthermore, to evaluate performance on cold-start users, a certain proportion of overlapping users ($\mathcal{K}_{cs}$) were randomly designated as cold-start users for validation and testing phases, in which we randomly remove the sequence from one domain of the selected overlapping users while retaining the last user-item interaction as the ground truth.

Table 4: Statistics on three Amazon's CDSR datasets.

Dataset	$\lvert\mathcal{U}\rvert$	$\lvert\mathcal{V}\rvert$	$\lvert\mathcal{E}\rvert$	#O	$\lvert\mathcal{S}\rvert$	Density
Cloth	41,454	17,939	175,552	9,721	4.50	0.024%
Sport	27,209	12,654	159,098		6.10	0.046%
Phone	27,320	9,478	140,886	20,342	5.36	0.054%
Elec	107,580	40,446	758,374		8.00	0.017%
Game	24,929	12,314	146,639	2,171	6.23	0.048%
Video	19,347	8,746	139,236		7.66	0.082%

#O: the number of overlapping users across domains.

C COMPARING METHODS

In this section, we provide a detailed introduction to our compared baselines. We also include the official code used for their implementation, and for methods without official code, we have reproduced the implementations based on the descriptions in the papers.

Single-domain recommendation methods:

• Multi-VAE[2][Liang et al., 2018]: Multi-VAE is a variational autoencoder (VAE)-based model that enhances traditional linear factorization methods through Bayesian inference. It addresses data sparsity by learning latent preference distributions, making it more effective than traditional matrix factorization.

• SVAE[3][Sachdeva et al., 2019]: Sequential Variational Autoencoder (SVAE) integrates RNNs into a VAE framework to capture the temporal dynamicss of user behavior, balancing short-term and long-term preferences. SVAE is particularly effective for time-dependent user interactions, such as video or music recommendations.

• SASRec[4][Kang and McAuley, 2018]: SASRec uses self-attention mechanisms to model long-range dependencies in user behavior sequences. Unlike RNNs, it efficiently handles long sequences by balancing model complexity with capturing subtle user preferences.

Cross-domain sequential recommendation methods:

• Pi-Net[5][Ma et al., 2019]: Pi-Net generates shared user embeddings using a gating mechanism to differentiate behaviors across domains. It excels in scenarios requiring integrated information from multiple domains, such as combining video and music user behavior.

• DASL[6][Li et al., 2021]: DASL employs a dual-attention mechanism to enhance cross-domain recommendation accuracy by focusing on user behaviors in both source and target domains. It effectively captures cross-domain behavior patterns, making it suitable for scenarios where precise cross-domain behavior modeling is needed.

[1]Available at `http://jmcauley.ucsd.edu/data/amazon/index_2014.html`
[2]`https://github.com/dawenl/vae_cf`
[3]`https://github.com/noveens/svae_cf`
[4]`https://github.com/pmixer/SASRec.pytorch`
[5]`https://github.com/mamuyang/PINet`
[6]`https://github.com/lpworld/DASL`

• C^2DSR[7][Cao et al., 2022a]: C^2DSR leverages graphical attention encoders and contrastive learning to jointly model intra- and cross-domain preferences. This method is particularly effective in addressing data sparsity and cold-start issues by capturing global user interests.

Cross-domain recommendation methods:

• EMCDR [Man et al., 2017]: EMCDR uses MLPs to learn domain-specific representations and maps them across domains using overlapping user information. It's especially useful when user behavior in one domain needs to be mapped to another, though it relies heavily on overlapping users.

• SA-VAE [Salah et al., 2021]: SA-VAE aligns latent spaces between source and target domains using VAE, exploring both rigid and soft alignment strategies. It is effective for cross-domain recommendations in cold-start and sparse data scenarios by leveraging shared features across domains.

• CDRIB[8][Cao et al., 2022c]: CDRIB applies the information bottleneck principle to extract domain-shared features, improving recommendation effectiveness and addressing cold-start challenges.

D IMPLEMENTATION DETAILS

D.1 HYPERPARAMETER SETTINGS

To ensure fair evaluation, we standardize hyperparameters across all methods. Model-specific hyperparameters for each baseline are set according to their original papers or official code implementations. Across all models, we set the embedding dimension to $d = 128$, the batch size to 512, and train for 100 epochs using Adam, with the learning rate chosen from $\{3 \times 10^{-4}, \ldots, 8 \times 10^{-4}\}$. The historical behavior length T is set to 20, and in our method, the pseudo-sequence length T' is set to 40. The hyperparameters λ_c and λ_a are selected from $\{5 \times 10^{-4}, \ldots, 5 \times 10^{-3}\}$.

To ensure reliable results, each method is run five times with different random seeds, and the best model is selected based on the highest NDCG@10 performance on the validation set via grid search. For the PSG module in I^2VAE, we follow LightGCN's default settings[9], using 20% of the training set as validation data and selecting the best checkpoint after 100 epochs for pseudo-sequence recall.

D.2 IMPLEMENTATION OF DIFFERENT ENCODERS IN I^2VAE

Our variational encoders parameterize the posterior distributions using MLPs. For domain-specific posteriors, the mean and standard deviation are computed as: $\mu^x = \text{MLP}^\mu(\boldsymbol{S^X}), \quad \sigma^x = \text{MLP}^\sigma(\boldsymbol{S^X})$. A similar approach is used for pseudo-sequence posteriors, taking inputs $\boldsymbol{S_a^X}$ and $\boldsymbol{S_a^Y}$. For cross-domain posteriors, we employ multi-head attention. Specifically, the unaggregated user interest representations—i.e., those that have not undergone mean pooling—$\boldsymbol{S^{X'}} = \{\mathbf{h}_{S_1^X}, \cdots, \mathbf{h}_{S_T^X}\}$ serve as the query, while $\boldsymbol{S^{Y'}} = \{\mathbf{h}_{S_1^Y}, \cdots, \mathbf{h}_{S_T^Y}\}$ is used as both the key and value for obtaining the cross-domain transferable interest from domain Y to domain X. The mean vector μ_t^y is computed as:

$$\mu_t^y = \text{MLP}(\text{Mean-Pooling}(\text{Attention}(\boldsymbol{S^{X'}}, \boldsymbol{S^{Y'}}, \boldsymbol{S^{Y'}}))),$$

with σ_t^y derived similarly. The same symmetric process is applied to compute the posteriors of the cross-domain interest representation $\boldsymbol{Z_t^X}$.

[7] https://github.com/cjx96/C2DSR
[8] https://github.com/cjx96/CDRIB
[9] https://github.com/gusye1234/LightGCN-PyTorch

E PERFORMANCE COMPARISON RESULTS

We compare the performance of i^2VAE with state-of-the-art baselines across three real-world CDSR datasets: "Cloth-Sport", "Phone-Elec", and "Game-Video", as shown in the following table. The baselines are categorized into three groups: Single-domain recommendation (SDR) models, CDR-sequential models, and CDR models. We conduct evaluations across three user types—long-tailed, cold-start, and all users—and report recommendation performance in both domains. The results demonstrate that i^2VAE consistently outperforms baselines across different user groups and domains, highlighting its effectiveness in capturing user interests and improving recommendation quality.

Table 5: Experimental Results (%) across different types of users, including long-tailed(tailed), cold-start, and all users, on three CDSR datasets. We highlight the methods with the **best** and **second-best** average performances.

Datasets	User Types	Metric	SDR			CDR-sequential			CDR			Ours	↑(%)
			Multi-VAE	SVAE	SASRec	DASL	PiNet	C^2DSR	DisenCDR	SA-VAE	CDRIB	i^2VAE	
Cloth	Tailed	NDCG	2.31±0.08	2.07±0.16	2.09±0.20	2.28±0.17	2.11±0.17	2.29±0.09	2.20±0.06	**2.40±0.14**	2.27±0.10	**2.52±0.07***	5.00
		HR	4.15±0.12	3.99±0.34	3.99±0.31	**4.46±0.24**	4.15±0.36	4.38±0.32	4.21±0.15	4.43±0.22	4.21±0.21	**4.73±0.16***	6.05
	Cold-start	NDCG	3.23±0.29	**3.41±0.35**	2.86±0.50	3.28±0.18	3.04±0.66	3.08±0.65	2.95±0.29	3.25±0.24	3.00±0.39	**3.45±0.35***	1.17
		HR	6.08±0.38	**6.39±0.43**	5.64±0.86	6.27±0.40	5.58±1.24	5.96±0.95	5.58±0.23	5.89±0.46	5.52±0.58	**6.77±0.51***	5.95
	All	NDCG	2.20±0.10	2.18±0.10	2.11±0.15	2.43±0.09	2.18±0.10	2.40±0.07	2.30±0.05	**2.52±0.10**	2.39±0.06	**2.59±0.11***	2.78
		HR	4.25±0.12	4.28±0.28	4.15±0.30	**4.82±0.11**	4.24±0.18	4.63±0.18	4.48±0.18	4.81±0.19	4.56±0.09	**5.00±0.17***	3.73
Sport	Tailed	NDCG	3.02±0.14	2.90±0.16	2.80±0.21	**3.31±0.36**	2.96±0.15	3.06±0.15	3.14±0.10	3.29±0.15	3.12±0.11	**3.36±0.10***	1.51
		HR	6.06±0.50	5.76±0.34	5.72±0.22	6.29±0.57	5.89±0.25	6.03±0.22	5.94±0.19	**6.31±0.42**	5.81±0.10	**6.66±0.17***	5.55
	Cold-start	NDCG	4.33±0.49	4.19±0.24	3.92±0.40	4.48±0.42	4.24±0.31	4.65±0.53	4.93±0.18	**5.05±0.29**	4.95±0.25	**5.43±0.29***	7.52
		HR	8.89±0.66	8.62±0.73	7.88±0.62	8.15±0.94	8.28±0.27	9.36±0.54	8.54±0.59	**9.63±0.73**	9.09±0.43	**10.30±0.46***	6.96
	All	NDCG	3.79±0.07	3.89±0.15	3.70±0.18	3.93±0.14	3.68±0.08	4.13±0.17	4.21±0.13	**4.31±0.13**	4.27±0.09	**4.40±0.10***	2.20
		HR	7.23±0.27	7.27±0.29	7.00±0.17	7.45±0.32	6.90±0.33	7.84±0.33	7.90±0.22	**8.06±0.26**	7.88±0.33	**8.53±0.14***	5.80
Phone	Tailed	NDCG	3.58±0.13	3.51±0.11	3.41±0.13	**4.10±0.15**	3.68±0.17	3.81±0.17	4.01±0.14	4.06±0.31	4.01±0.23	**4.23±0.21***	3.17
		HR	6.90±0.34	6.74±0.20	6.58±0.32	**8.06±0.27**	7.17±0.20	7.47±0.39	7.78±0.39	7.96±0.61	7.92±0.38	**8.17±0.27***	1.36
	Cold-start	NDCG	3.16±0.19	3.15±0.30	2.97±0.38	3.63±0.25	3.51±0.38	3.73±0.37	3.80±0.33	3.83±0.48	**4.00±0.35**	**4.39±0.21***	9.75
		HR	6.26±0.39	6.11±0.42	6.03±0.45	7.40±0.39	7.18±0.51	7.18±0.74	7.56±0.56	**7.94±0.85**	7.56±0.51	**8.63±0.19***	8.69
	All	NDCG	3.98±0.19	3.89±0.06	3.88±0.14	4.40±0.17	4.13±0.14	4.36±0.21	**4.52±0.15**	4.49±0.29	4.46±0.19	**5.79±0.29***	3.23
		HR	7.59±0.39	7.33±0.14	7.36±0.27	8.54±0.32	7.77±0.34	8.30±0.38	**8.68±0.22**	8.54±0.50	8.66±0.29	**10.61±0.22***	2.35
Elec	Tailed	NDCG	6.96±0.23	6.74±0.25	6.78±0.29	7.63±0.19	7.09±0.24	7.78±0.13	7.64±0.10	7.60±0.30	**7.77±0.11**	**8.00±0.10***	2.96
		HR	11.65±0.48	11.39±0.47	11.49±0.53	**12.83±0.41**	11.66±0.56	13.05±0.28	12.45±0.25	12.56±0.39	12.66±0.28	**13.49±0.19***	5.14
	Cold-start	NDCG	9.35±0.33	9.22±0.19	9.16±0.19	9.73±0.65	9.59±0.35	9.85±0.30	**9.90±0.25**	9.76±0.48	9.73±0.40	**10.26±0.42***	3.64
		HR	14.71±0.49	14.76±0.60	14.94±0.51	15.76±1.24	15.00±1.00	**15.94±0.43**	15.24±0.60	15.35±0.90	15.53±0.34	**17.12±0.39***	7.40
	All	NDCG	8.20±0.22	8.06±0.27	8.08±0.34	8.58±0.16	7.84±0.08	9.00±0.12	8.95±0.13	8.84±0.19	**9.04±0.04**	**9.33±0.09***	3.22
		HR	13.08±0.43	12.81±0.45	12.86±0.59	13.60±0.50	12.31±0.19	**14.46±0.31**	14.03±0.21	13.96±0.26	14.08±0.26	**15.28±0.26***	5.64
Game	Tailed	NDCG	5.08±0.14	5.17±0.08	5.12±0.19	4.64±0.25	4.64±0.25	5.04±0.08	5.25±0.22	**5.31±0.15**	5.23±0.14	**5.51±0.12***	3.77
		HR	9.48±0.38	9.46±0.14	9.39±0.34	8.45±0.38	8.45±0.38	9.34±0.17	9.73±0.35	**9.95±0.39**	9.68±0.35	**10.09±0.18***	1.41
	Cold-start	NDCG	9.13±0.64	**9.97±1.37**	8.61±1.92	7.46±0.93	7.46±0.93	9.18±1.57	8.91±1.44	9.69±1.09	7.72±1.63	**10.92±0.70***	9.53
		HR	20.00±3.95	19.15±1.90	16.60±2.48	15.32±1.59	15.32±1.59	16.17±2.89	16.60±2.48	**19.57±1.59**	15.74±2.89	**22.55±2.17***	15.20
	All	NDCG	5.39±0.15	5.40±0.09	**5.57±0.10**	4.82±0.21	4.82±0.21	5.53±0.11	5.49±0.14	5.51±0.09	5.30±0.05	**5.79±0.12***	4.01
		HR	9.95±0.30	9.94±0.18	10.20±0.30	8.87±0.22	8.87±0.22	10.28±0.23	10.28±0.23	**10.32±0.04**	9.92±0.19	**10.61±0.15***	2.77
Video	Tailed	NDCG	6.79±0.11	6.86±0.10	6.82±0.18	6.32±0.18	6.32±0.18	6.75±0.20	6.89±0.16	6.88±0.10	**7.05±0.08**	**7.08±0.14***	0.43
		HR	12.43±0.19	12.50±0.18	12.26±0.39	11.46±0.36	11.46±0.36	12.34±0.35	12.57±0.29	12.57±0.25	**12.68±0.22**	**12.95±0.26***	2.13
	Cold-start	NDCG	7.40±2.13	8.19±1.50	7.93±1.39	8.10±3.21	8.10±3.21	**8.73±1.17**	8.57±1.31	8.54±2.13	8.26±1.71	**9.62±1.58***	10.10
		HR	18.67±2.67	21.78±3.27	20.00±2.43	17.78±4.66	17.78±4.66	21.78±2.59	**22.22±2.81**	19.11±3.61	20.44±2.18	**22.67±2.18***	2.03
	All	NDCG	6.49±0.09	6.59±0.10	6.50±0.11	5.87±0.15	5.87±0.15	6.46±0.19	6.04±0.19	6.58±0.16	**6.67±0.11**	**6.79±0.09***	1.80
		HR	12.10±0.16	12.18±0.14	11.97±0.31	10.92±0.22	10.92±0.22	12.14±0.25	11.25±0.30	12.32±0.29	**12.33±0.16**	**12.67±0.18***	2.76

"*" denotes statistically significant improvements ($p < 0.05$), as determined by a paired t-test comparison with the second best result.

F MODEL ANALYSIS

F.1 ANALYSIS OF OVERLAPPING RATIO

We designed experiments to test our model's performance with fewer cross-domain overlapping users by randomly downsampling the proportion of overlapping users ($\mathcal{K}_o$) in the training set while keeping the test set unchanged. We compared our model with SA-VAE at $\mathcal{K}_o$ values of {25%, 50%, 75%, 100%}. This experiment simulates real-world scenarios, requiring the model to maintain stable performance with few overlapping users. Results are shown in Table 6.

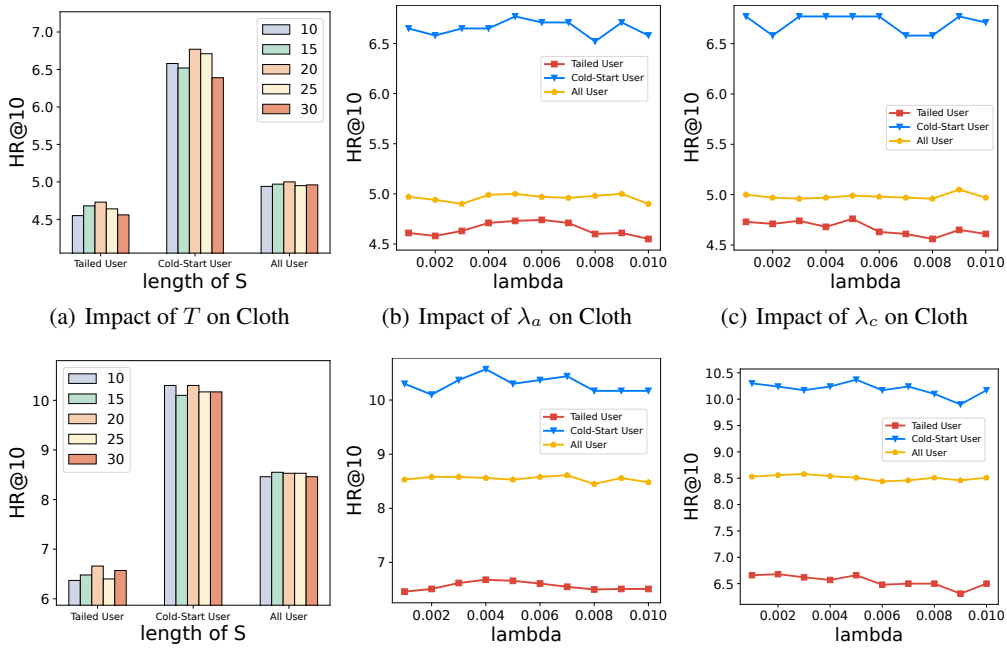
As the number of overlapping users decreases, both SA-VAE and our model's performance decline due to the increased difficulty in capturing cross-domain signals. However, our model consistently outperforms SA-VAE at every $\mathcal{K}_o$ level. This is because our model does not rely solely on overlapping users. While capturing cross-domain information becomes harder, our denoised pseudo-sequence effectively captures intra-domain interests. Additionally, the cross-domain and disentangle regularizers allow better inference on non-overlapping users for cross-domain signals based on behaviors in one domain.

Table 6: Experiment results (%) on "Cloth-Sport" dataset with the varying overlapping ratio ($\mathcal{K}_o$).

Domain	User	Metric	$\mathcal{K}_o = 25\%$		$\mathcal{K}_o = 50\%$		$\mathcal{K}_o = 75\%$		$\mathcal{K}_o = 100\%$	
			SA-VAE	i^{2}VAE	SA-VAE	i^{2}VAE	SA-VAE	i^{2}VAE	SA-VAE	i^{2}VAE
Cloth	Tailed	NDCG	0.93	**1.32**	1.69	**1.87**	2.25	**2.30**	2.40	**2.52**
		HR	2.02	**2.72**	3.24	**3.70**	4.26	**4.50**	4.43	**4.73**
	Cold-Start	NDCG	0.95	**1.62**	2.19	**2.52**	3.26	**3.27**	3.25	**3.45**
		HR	2.07	**3.39**	4.83	**5.20**	6.27	**6.21**	5.89	**6.77**
	All	NDCG	0.92	**1.31**	1.71	**1.93**	2.29	**2.42**	2.52	**2.59**
		HR	1.94	**2.71**	3.43	**3.97**	4.40	**4.75**	4.81	**5.00**
Sport	Tailed	NDCG	1.15	**1.92**	2.31	**2.77**	3.07	**3.36**	3.29	**3.36**
		HR	2.48	**4.06**	4.59	**5.60**	5.82	**6.62**	6.31	**6.66**
	Cold-Start	NDCG	1.58	**2.93**	3.57	**4.48**	4.80	**5.27**	5.05	**5.43**
		HR	3.37	**6.13**	6.80	**8.82**	9.29	**10.51**	9.63	**10.30**
	All	NDCG	1.45	**2.65**	3.08	**3.85**	4.03	**4.28**	4.31	**4.40**
		HR	2.97	**5.35**	5.88	**7.53**	7.65	**8.22**	8.06	**8.53**

F.2 ANALYSIS OF PARAMETER SENSITIVITY

In this section, we investigate the parameter sensitivity of sequence length T and the harmonic factors λ_a and λ_c. Figures 3(a) and (d) show the impact of sequence length T on model performance (HR@10) in the Cloth and Sport domains, with T varying in {10, 15, 20, 25, 30}. The model performs best when $T = 20$. Increasing T from 10 to 20 improves performance due to richer historical information, but performance decreases for $T > 20$ due to padding items. Figures 3(b) and (e) show the impact of λ_a on model performance with λ_a varying in {0.001, 0.002, ..., 0.010} and λ_c fixed at 0.001. The model performs best when λ_a is around 0.004 or 0.005, indicating effective denoising without depressing the classification loss. Similarly, Figures 3(c) and (f) show the impact of λ_c on model performance with λ_a fixed at 0.005. The model achieves superior results when λ_c is set to 0.001 or 0.005.



(a) Impact of T on Cloth (b) Impact of λ_a on Cloth (c) Impact of λ_c on Cloth

(d) Impact of T on Sport (e) Impact of λ_a on Sport (f) Impact of λ_c on Sport

Figure 3: Experiment results (%) of parameter sensitivity on the "Cloth" and "Sport" domains, respectively.