Enhancing Healthcare Recommendations: A Privacy-Protective and Interpretable Cross-Domain Framework

Xun Liang, Zhiying Li, Hongxun Jiang*

School of Information, Renmin University of China
No. 59 Zhongguancun Street, Beijing, 100872, P.R. China
{xliang, zhiyingli, jianghx}@ruc.edu.cn

Abstract

Cross-domain recommendations in healthcare services differ from traditional ones in electronic commerce due to the need for heightened medical privacy protection for a small group of users, while ensuring the majority, who may lack sufficient medical knowledge, can understand the recommendations. To recommend doctors who provide online consultations to health video viewers and enable multimodal cross-domain recommendations from short video platforms (source domain) to online healthcare communities (target domain), this paper introduces a framework based on the User-Centric Synthetic Data Architect (UCSDA) and Pre-trained Large Language Model (PtLLM). UCSDA employs a user-centric, advanced selection-synthesis mechanism to filter users' cold interaction items and synthesize noise items, reducing privacy leakage risk. PtLLM focuses on necessary patient and doctor IDs during the recommendation decision process to generate explanations. The model's effectiveness and scalability were validated using three public datasets and a healthcare cross-domain recommendation dataset. In addition to traditional evaluation metrics, strong privacy metrics and the unique sentence ratio were used to assess privacy protection and interpretability. We also compared the characteristics of privacy protection and interpretability between e-commerce and healthcare recommendation scenarios.

Code and datasets — https://github.com/zyl-mc/HCR

Introduction

In today's digital landscape, short video platforms have successfully adapted to users' fragmented attention spans through rapid content updates, reshaping the attention economy. As user engagement has grown, advertisers and e-commerce platforms have increasingly integrated these platforms, facilitating smooth transitions from video content to direct purchases. To enhance user experience, cross-domain recommendation (CDR) algorithms have been developed to recommend products based on user behavior, improving both customer satisfaction and creating new marketing opportunities for businesses.

In healthcare, CDR systems differ significantly from traditional e-commerce (Jiang, Mi, and Xu 2024). The target audience includes consumers and patients, most of whom lack professional medical knowledge. As a result, recommendations must be highly explainable, allowing patients to understand and accept them, which in turn boosts the adoption of recommended treatments and improves user satisfaction with healthcare services. Moreover, sensitive patient data, such as medical records, require strict protection, meaning recommendation algorithms must also safeguard this information.

Healthcare recommendation algorithms are still challenging due to the opacity of machine learning models, which hinders understanding the link between patient data and recommendations. Since patients have limited medical knowledge, clear explanations are needed to build trust. Explainability can be improved with templates (Zhang et al. 2014), highlighted images (Chen et al. 2019), and autogenerated text (Li, Zhang, and Chen 2020), the latter being boosted by natural language generation (NLG). GPT models are good at personalization but are complex for healthcare. Researchers are looking into task-adaptive models that keep GPT's structure but make it generate more understandable healthcare explanations, showing potential for better explainability.

Most existing CDR systems assume that plaintext data can be transferred across domains (Chen et al. 2023b). However, in healthcare, where patient privacy is critical, this assumption falls short (Liu et al. 2021). The absence of robust privacy-preserving technologies hinders the practical use of CDRs in this sector, emphasizing the need for privacy-preserving CDRs (PPCDRs) that can balance privacy protection with high-quality recommendations. While some studies (Liu et al. 2021; Chen et al. 2023a) have introduced privacy-preserving techniques, such as differential privacy and federated learning, these methods face two key challenges. First, they often overlook the heterogeneity and multimodal nature of healthcare data, which includes text, images, audio, and video. To provide accurate recommendations, CDR systems must integrate these multimodal features effectively. Second, existing solutions mainly protect privacy during model training and result collection, yet in healthcare, users are particularly concerned about the privacy of their interaction data. Additionally, decentralized CDR frameworks are difficult to implement due to high communication and computational costs. In response, the use of synthetic data has emerged as a promising solution

*Corresponding author.

to ensure privacy protection in healthcare CDRs.

To meet the demand for privacy-preserving and interpretable CDRs in healthcare, this paper proposes a Hybrid Cross-domain Recommendation (HCR) framework with a self-supervised modality-aware information encoder. This encoder captures and integrates behavioral and modality information from various sources, enabling a comprehensive understanding of user behavior and preferences through self-supervised auxiliary tasks. Furthermore, we introduce a privacy-preserving synthesizer that perturbs real user interaction data, preventing the leakage of plaintext from the source domain while preserving user preferences. Finally, we incorporate a pre-trained language model into the HCR framework, employing a two-stage adjustment strategy to sequentially fine-tune prompt and model parameters. This approach ensures alignment and generates patient-friendly textual explanations, enhancing the system's effectiveness and user trust.

Related Work

CDR systems utilize data from dense domains to address sparsity in others, improving user-item modeling and knowledge transfer (Hu, Zhang, and Yang 2018). For example, Chen et al. (2023b) aligned users across domains with transformation modules. Recent studies also integrated user reviews to enhance model performance. However, most CDR models overlook privacy, which is particularly critical in healthcare. Balancing effective recommendations with privacy protection remains a key challenge.

Recently, many privacy-preserving methods have been added to recommendation systems. PriCDR (Chen et al. 2022) adds differential privacy noise to the user-item matrix and shares data directly. However, these often separate privacy from collaborative filtering, leading to less than optimal results. DPSMRec (Liu et al. 2023b) addresses this by integrating semantic and structural information with a differential privacy-enhanced sparse optimal transport algorithm. Federated learning has been used in CDR, with models like FedCT (Liu et al. 2021) and P2FCDR (Chen et al. 2023a) improving user data privacy. But most ignore the protection of recommendation interaction data. PPGenCDR (Liao et al. 2023) fills this gap using GANs and Rényi differential privacy for a robust framework. GANs have potential, but come with challenges like complex training and mode collapse. Current methods focus on recommendation accuracy over privacy measurement. In healthcare, it is vital to introduce metrics for privacy effectiveness, as done in vehicular communication. This paper presents several privacy metrics for accurate and consistent evaluations.

In recommendation systems, items have audio, images, and text features. Multimodal representation learning turns them into vectors for better content understanding, but it is challenging due to data complexity and heterogeneity. Early methods used pretrained neural networks to extract features and combined them with user behavior data (He and McAuley 2016). Recently, GCN-based methods like MMGCN (Wei et al. 2019) and GRCN (Wei et al. 2020) have excelled in multimodal recommendations by creating user-item interaction graphs. However, they often miss the differences in user preferences across modalities. Adaptive methods like SLMRec (Tao et al. 2022) and BM3 (Zhou et al. 2023) use self-supervised learning to improve modality alignment and fusion, but struggle with noise in item representations. Our proposal adds a self-supervised task to learn user modal preferences and performs fine-grained feature aggregation to avoid noise contamination in node embeddings.

The Transformer (Vaswani 2017), initially used in machine translation, has shown effectiveness on natural language tasks, but it requires large models and data. Prompt learning (Liu et al. 2023a) has advanced, allowing pretrained models to handle various tasks with specific prompts, avoiding retraining. This paper aims to integrate pre-trained language models into a privacy-preserving cross-domain framework, focusing on healthcare recommendations. We suggest using GPT-2 with continuous prompts (ID embeddings) and a two-stage tuning method to align prompts and model parameters.

Methodology

Problem Formulation

Consider two domains, $\mathcal{S}$ and $\mathcal{T}$, with the same set $\mathcal{U}$ of $N_{\mathcal{U}}$ users, but different rating matrices $\mathbf{R}_{\mathcal{S}}$ and $\mathbf{R}_{\mathcal{T}}$. $N_{\mathcal{S}}$ and $N_{\mathcal{T}}$ items are in $\mathcal{S}$ and $\mathcal{T}$, respectively. We create unique ID embeddings for users and items in both domains. We use pretrained models to extract embeddings from text, images, and audio. Visual features are extracted from video keyframes using VGG19 (Simonyan and Zisserman 2014), acoustic features from audio tracks with Librosa (McFee et al. 2015), and textual features by fine-tuning XLM-RoBERTa (Conneau, Khandelwal, and Goyal 2019). The modalities are represented as $\mathcal{M} = \{v, a, t\}$ for visual, acoustic, and textual. The goal is to transfer information from the source domain to the target domain while preserving privacy and improving recommendation performance. The proposed HCR model recommends items with explanations. Its framework includes a Self-Supervised Modality-Aware Information Encoder, Privacy-Preserving Synthesizer, and Explanation Generator, as shown in Fig. 1.

Self-Supervised Modality-Aware Information Encoder

Adding modal features helps improve user preference modeling, which then boosts the accuracy of recommendation systems (Zhang et al. 2021). We have designed a module that automatically combines these different types of data to better understand users and make better recommendations. It is in both domains but does not share its learning across domains.

Behavior Embedding. Inspired by the recent success of applying LightGCN (He et al. 2020) for recommendation, we design a content-aware graph convolution operation to encode information about various interactions on user-item graph. The message propagation stage at l-th graph convo-

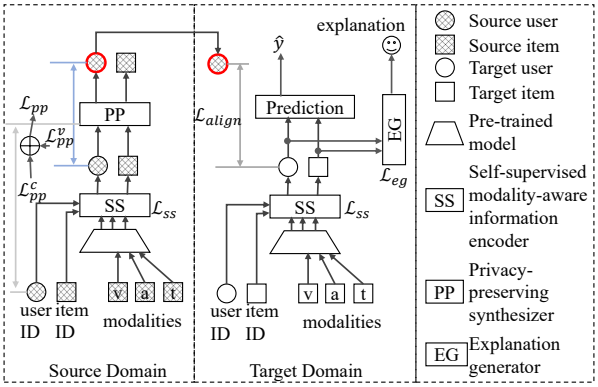

Figure 1: Framework of HCR. Pre-trained models are utilized to extract modality features, and the Self-Supervised Modality-Aware Information Encoder is designed to aggregate multimodal features and behavioral information. The Privacy-Preserving Synthesizer is used to perturb user interaction data, while the Explanation Generator is used to generate textual explanations for a given user-item pair.

lution layer can be formulated as:

$$\mathbf{E}_{id}^{(l)} = \sigma\left(\hat{\mathbf{A}}\mathbf{E}_{id}^{(l-1)}\mathbf{W}^{(l-1)}\right), \tag{1}$$

where σ is the non-linear ReLu function, $\hat{\mathbf{A}}$ is the re-normalization of the adjacency matrix. Then a simplified graph convolutional layer is defined as:

$$\mathbf{E}_{id}^{(l)} = \left(\mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}\right)\mathbf{E}_{id}^{(l-1)}. \tag{2}$$

The representations of the l-th layer encode the l-order neighbors' information. By incorporating residual connections into the users' and items' initial embeddings in a GCN, we can mitigate the over-smoothing problem and get the final representations $\overline{\mathbf{E}}_{id}$:

$$\begin{aligned}\overline{\mathbf{E}}_{u,id} &= \frac{1}{L+1}\sum_{l=0}^{L}\mathbf{E}_{u,id}^{l} \\ \overline{\mathbf{E}}_{i,id} &= \frac{1}{L+1}\sum_{l=0}^{L}\mathbf{E}_{i,id}^{l} + \mathbf{E}_{i,id}^{0}.\end{aligned} \tag{3}$$

Modality Embedding. Studies have stopped modal noise in graph neural networks using indirect modal info, but this harms recommendation systems. We fix this by adding an item ID embedding matrix and a user-item interaction matrix to adjust the modal embeddings. We start by introducing the item's initial ID embedding matrix $\mathbf{E}_{\mathcal{I},id}$ and use it to transform the modality matrix $\mathbf{E}_{\mathcal{I},m}$ into a modal-specific representation $\overline{\mathbf{E}}_{\mathcal{I},m}$ through an element-wise multiplication with a sigmoid activated transformation:

$$\overline{\mathbf{E}}_{\mathcal{I},m} = \mathbf{E}_{\mathcal{I},id} \odot \sigma\left(\mathbf{W}_1\mathbf{E}_{\mathcal{I},m} + \mathbf{b}_1\right), \tag{4}$$

where $\mathbf{W}_1$ and $\mathbf{b}_1$ are learnable weights. Then, we multiply the user-item interaction matrix $\mathbf{R}_{\mathcal{I}}$ with the transformed matrix $\overline{\mathbf{E}}_{\mathcal{I},m}$ to get the fused user modality vector $\mathbf{E}_{\mathcal{U},m}$:

$$\overline{\mathbf{E}}_{\mathcal{U},m} = \mathbf{W}_2\dot{\mathbf{E}}_{\mathcal{I},m}\left(\mathbf{R}_{\mathcal{I}}\right)^T + \mathbf{b}_2. \tag{5}$$

Finally, by concatenating user and item single-modal representations, we obtain the modal-specific representations of all nodes $\overline{\mathbf{E}}_m \in \mathbb{R}^{d\times(N_U+N_{\mathcal{I}})}$.

Information Fusion. We consider visual, acoustic, textual, and ID embeddings as inputs for the self-supervised task. Using a multi-task strategy, the graph-based recommender is the main task, supported by the self-supervised task. We apply spatial transformations to ID and modality embeddings before fusion, aligning them in modality-specific spaces:

$$\mathfrak{T}_m(\mathbf{X}) : S(\mathbf{X}) \to S(m), \tag{6}$$

where $\mathfrak{T}_m(\mathbf{X})$ projects the input feature $\mathbf{X}$ from its original feature space to a specific modal space, $\mathfrak{T}_m(\mathbf{X}) = \sigma\left(\mathbf{W}_m\mathbf{X} + \mathbf{b}_m\right)$, $\mathbf{W}_m \in \mathbb{R}^{d\times d}$, $\mathbf{b}_m \in \mathbb{R}^d$, σ is the sigmoid nonlinearity. This maintains a distinct space for each modality, with ID embeddings guiding the alignment of visual, acoustic, and textual embeddings into a common space. The final fusion embeddings are an adaptive fusion of the transformed embeddings:

$$\mathbf{E}_{fuse} = \frac{1}{|\mathcal{M}|}\begin{bmatrix}\mathfrak{T}_v\left(\overline{\mathbf{E}}_{id}\right) \odot \mathfrak{T}_v\left(\overline{\mathbf{E}}_v\right) \\ + \mathfrak{T}_a\left(\mathfrak{T}_v\left(\overline{\mathbf{E}}_{id}\right)\right) \odot \mathfrak{T}_a\left(\overline{\mathbf{E}}_a\right) \\ + \mathfrak{T}_t\left(\mathfrak{T}_a\left(\mathfrak{T}_v\left(\overline{\mathbf{E}}_{id}\right)\right)\right) \odot \mathfrak{T}_t\left(\overline{\mathbf{E}}_t\right)\end{bmatrix}. \tag{7}$$

User preferences are embedded in behavioral features, so we use this information to enhance the fused embeddings. We also create a self-supervised learning task to delve into user preferences, aiming to maximize mutual information between behavioral and multimodal features (Kemertas et al. 2020). This task follows the InfoNCE loss (Oord, Li, and Vinyals 2018) and is formulated to measure the cosine similarity between ID and fused embeddings, normalized by a temperature hyperparameter:

$$\mathcal{L}_{ss} = -\log\frac{\exp\left(\mathrm{sim}\left(\bar{e}_{i,id}, e_{i,fuse}\right)/\tau_{ss}\right)}{\sum_{j\in[N]}\exp\left(\mathrm{sim}\left(\bar{e}_{j,id}, e_{j,fuse}\right)/\tau_{ss}\right)}. \tag{8}$$

Privacy-Preserving Synthesizer

$I_{\mathcal{S},u}$ denotes the set of items in $\mathcal{S}$ that user u has interacted with. Note that the interactions can be either implicit (e.g., click) or explicit (e.g., rating). Given the historical data $I_{\mathcal{S},u}$ of user u in source domain $\mathcal{S}$, our goal is to generate the synthetic data under users' privacy preference, i.e. Virtual items $V_{\mathcal{S},u}$, to replace a certain percentage of the original items. Note that synthetic interaction data contains few or no sensitive information.

Cold Items Selection. Historical interaction items vary in their impact on user preferences. Less influential items, or cold items, can be replaced with synthetic items for privacy protection. The replacement ratio, k, determines how many items are swapped. We use an attention mechanism to assess each item's contribution to a user's preferences: The attention weight a_{ui} for item i is calculated using a ReLU-activated hidden layer and a smoothing exponent β:

$$\begin{aligned}v_{ui} &= \mathbf{h}^T\,\mathrm{ReLU}\left(\mathbf{W}_3\left([\mathbf{e}_{u,fuse} : \mathbf{e}_{i,fuse}]\right) + \mathbf{b}_3\right), \\ a_{ui} &= \frac{\exp\left(v_{ui}\right)}{\left[\sum_{j'\in I_u}\exp\left(v_{ui'}\right)\right]^{\beta}}.\end{aligned} \tag{9}$$

The user's preferences $\mathbf{p}_u$ are represented by the weighted average of the fused embeddings of items they've purchased:

$$\mathbf{p}_u = \frac{1}{|I_{\mathcal{S},u}|} \sum_{i \in I_{\mathcal{S},u}} a_{ui} \mathbf{e}_{i,fuse}. \tag{10}$$

We expect the user preference representation to be similar to the user's ID embedding. To learn the attention weights, we use the ID embedding $\overline{\mathbf{e}}_{u,id}$ as supervision and define an L2 regularizer:

$$\mathcal{L}_{pp}^{c} = \sum_{u \in \mathcal{U}} \left\| f\left(\mathbf{p}_u, \theta\right) - \overline{\mathbf{e}}_{u,id} \right\|^2, \tag{11}$$

where an MLP with dropout is used as the transformation function $f(\cdot)$ to map $\mathbf{p}_u$ to the same space as $\overline{\mathbf{e}}_{u,id}$. θ represents the trainable parameters of the MLP. During training, we minimize $\mathcal{L}_{pp}^{c}$ to determine the contribution of each item. The items with the lowest attention weights, representing the top k percent, are identified as cold items $I_{\mathcal{S},u}^{c}$.

Virtual Items Generation. The synthesizer creates a virtual item to attract user u's attention, considering user preferences, privacy sensitivity, and cold item characteristics. It combines the user vector $\mathbf{e}_{u,fuse}$, a cold item vector $\mathbf{e}_{i,fuse}(i \in I_{\mathcal{S},u}^{c})$, and a privacy parameter γ_u, then projects this into a latent space: $\boldsymbol{z}_{ui} = \mathbf{W}_4 \left[\mathbf{e}_{u,fuse}; \mathbf{e}_{i,fuse}; \gamma_u\right] + \mathbf{b}_4$, where $\boldsymbol{z}_{ui}$ is the latent feature of the output.

The similarity between the latent feature and all item embeddings is calculated, and the probability distribution over all items is estimated using softmax: $\boldsymbol{h}_{ui} = \boldsymbol{z}_{ui} \mathbf{E}_{\mathcal{S},fuse}$, but is non-differential. To address this, Gumbel-Softmax (Havrylov and Titov 2017) is used for a differentiable approximation. After generating a virtual interaction item, the user's privacy-protected representation is updated: $\mathbf{e}_u^{pp} = \overline{\mathbf{e}}_{u,id} + \mathbf{e}_{u,fuse} + \mathbf{e}_i$.

To prevent privacy leakage, a privacy regularizer constrains the similarity between the privacy-preserving and source user embeddings:

$$\mathcal{L}_{pp}^{v} = \sum_{(u,i)} \left[\mathrm{sim}\left(\mathbf{e}_u^{pp}, \mathbf{e}_u^{src}\right) - \gamma_u\right]_+. \tag{12}$$

Here, $[z]_+ = \max(z, 0)$ represents the hinge loss. The sensitivity γ_u serves as a threshold, allowing a certain level of similarity between the source and privacy-preserving user embeddings. The final loss function, combining $\mathcal{L}_{pp}^{c}$ and $\mathcal{L}_{pp}^{v}$, optimizes the synthesizer, balancing user preference and privacy:

$$\mathcal{L}_{pp} = \lambda_{pp} \mathcal{L}_{pp}^{c} + (1 - \lambda_{pp}) \mathcal{L}_{pp}^{v}. \tag{13}$$

The model aims to produce preferred virtual items and create perturbed user vectors for privacy-protected use in the target domain.

Explanation Generator

We use the pre-trained language model GPT-2 to generate explanations for recommendations by treating user and item embeddings as continuous prompts. This approach avoids the need for conversion and preserves important information. During training, we represent the input sequence as a concatenation of user-item embeddings, explanation words, and special token embeddings for users and items. The sequence is then processed by GPT-2, and a linear layer with softmax is used for next-word prediction. The vector $\mathbf{c}_t$ represents the probability distribution over the vocabulary. The negative log-likelihood (NLL) is used as the loss function to compute the average of user-item pairs in the training set:

$$\mathcal{L}_{eg} = \frac{1}{|\mathcal{T}|} \sum_{(u,i) \in \mathcal{T}} \frac{1}{|E_{u,i}|} \sum_{t=1}^{|E_{u,i}|} -\log c_{2+t}^{e_t}, \tag{14}$$

where $c_t^{e_t}$ is offset by two positions (i.e., user embedding and item embedding) because the explanation is placed at the end of the sequence.

For inference, we aim to generate the explanation word sequence with the highest log-likelihood. We use greedy decoding to select the word with the highest probability at each step. In the fine-tuning strategy, we introduce additional prompt parameters for the pre-trained language model. To bridge the gap between these parameters and the pre-trained language model parameters, we use a fixed-prompt LM tuning method. This involves first optimizing the prompt parameters, then fine-tuning all parameters together.

Prediction

Based on the ID embeddings enhanced by user-item interaction behavior graph and the fused multi-modal embeddings, we form the final representations of users and items in target domain:

$$\begin{aligned} \mathbf{e}_u^{\mathcal{T}} &= \overline{\mathbf{e}}_{u,id} + \mathbf{e}_{u,fuse}, \\ \mathbf{e}_i^{\mathcal{T}} &= \overline{\mathbf{e}}_{i,id} + \mathbf{e}_{i,fuse}. \end{aligned} \tag{15}$$

The inner product is adopted to predict the likelihood of interaction between user u and item i: $\hat{y} = (\mathbf{e}_u^{\mathcal{T}})^\top \mathbf{e}_i^{\mathcal{T}}$.

Optimization

During the phase of model training, we adopt binary cross entropy (BCE) loss as the basic optimization task for recommendation prediction in target domain:

$$\mathcal{L}_{\text{pred}} = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)], \tag{16}$$

where N is the number of samples, y_i is the i-th ground truth label. Additionally, we bridge the knowledge of source domain to the target domain by user alignment module, which minimizes the user differences in both domains ($\mathcal{S}$ and $\mathcal{T}$):

$$\mathcal{L}_{\text{align}} = \sum_{u=1}^{N_{\mathcal{U}}} \left\| \mathbf{e}_u^{pp} - \mathbf{e}_u^{\mathcal{T}} \right\|_F^2. \tag{17}$$

Overall, the optimization of HCR is to minimize:

$$\mathcal{L} = \mathcal{L}_{\text{pred}} + \lambda_a \mathcal{L}_{\text{align}} + \lambda_{ss} \mathcal{L}_{ss} + \lambda_{eg} \mathcal{L}_{eg} + \lambda \|\boldsymbol{\Theta}\|_2^2 \tag{18}$$

where $\boldsymbol{\Theta}$ is the set of model parameters; λ_a, λ_{ss}, λ_{eg} and λ are hyperparameters to adjust the balance between alignment task, self-supervised task, explanation generation task and the effect of L_2 regularization, respectively. HCR can model the distribution of private data in source domain and enhance the recommendation performance in the target domain.

Datasets	Users	Items	Ratings	Density
Amazon Arts	49,201	81,983	556,942	0.01%
Amazon Fashion		153,014	852,016	0.01%
Amazon Software	4,026	8,976	49,266	0.14%
Amazon Game		13,691	46,125	0.08%
Healthcare Video	22,824	50,256	322,159	0.03%
Healthcare Doctor		75,680	215,912	0.01%

Table 1: Statistics of datasets.

Experiments

This section aims to answer the following research questions through experiments and case studies.

Q1: Can the proposed HCR model achieve better performance and enhanced privacy compared to existing SOTA models of plaintext CDR, multimodal CDR, and advanced PPCDR?

Q2: Can the SS (Self-Supervised modality-aware encoder) and PP (Privacy Protection) submodules within HCR enhance its performance?

Q3: In the field of healthcare services, does the explanations provided by HCR significantly enhance patients' understanding of the recommendations?

Q4: How do various hyperparameters affect the performance of HCR?

Experimental Settings

Datasets. This paper first compares the performance of our proposed HCR method with advanced benchmark algorithms through extensive experiments on a large-scale public Amazon dataset (Cao et al. 2022). We focus on four subdomains: Arts, Fashion, Software, and Games.

We also collected a dataset for online medical consultation (OMC) services, combining data from a short video-sharing platform and an online health platform. The dataset includes two domains: health-related videos by certified professionals (source domain) and physicians on the health platform (target domain). User comments on the videos are labeled as positive (1) or negative (0), while doctor ratings range from 1 to 5. We then reversed the source and target domains and identified a shared user set by matching comments with patient queries. Dataset statistics are shown in Tab. 1. All data used is publicly available and collected compliantly. Viewer IDs are anonymized, depersonalized codes are used, and private information is manually excluded. The data aims to advance multimodal medical recommendation research while adhering to ethical standards.

Baselines. This paper compares the proposed HCR model with the following three baseline categories:

(1) CDR models.

- **CoNet**(Hu, Zhang, and Yang 2018) is a cross-connection unit to enable dual knowledge transfer across domains.
- **CCTL**(Zhang et al. 2023) a SOTA CDR model using a representation enhancement network to preserve domain-specific features.

(2) Multimodal CDR models.

- **SEMI**(Lei et al. 2021) presents a sequential multi-modal network for e-commerce micro-video recommendations.
- **PMMRec**(Li et al. 2024) uses multi-modal content for cross-domain and cross-platform recommendations.

(3) PPCDR models.

- **PriCDR**(Chen et al. 2022) publishes the perturbed interaction data using DP to the target domain.
- **DPSMrec**(Liu et al. 2023b) is a differentially private model for PPCDR, using semantic and structural data.
- **PPGenCDR**(Liao et al. 2023) is the SOTA PPCDR model using a GAN-based framework.

Additionally, we set up two variants of HCR: Ours w/o SS and Ours w/o PP.

- **Ours w/o SS** removes the self-supervised modality-aware encoder, instead averaging modality embeddings from a pre-trained model and adding them to item ID embeddings. These multimodal item embeddings, along with user ID embeddings, are processed by LightGCN for message passing on the user-item graph. However, it lacks independent modality learning and fusion, which may introduce noise into the graph.
- **Ours w/o PP** removes the privacy-preserving synthesizer and directly shares the user embeddings obtained from graph convolution with the target domain for alignment.

Evaluation Protocols. This study's experimental results are evaluated across three key dimensions: recommendation accuracy, privacy protection, and interpretability. We evaluate recommendation performance with hit ratio (HR) and normalized discounted cumulative gain (NDCG), cutting off the ranked list at 5 (Liu et al. 2023b).

We assess privacy-preserving techniques in recommender systems by evaluating user embedding vectors for privacy risks. We consider a user's information sources: embedding vector $\mathbf{e}_u^{\text{src}}$, item embedding matrix $\mathbf{E}_I^{\text{src}} \in \mathbb{R}^{d \times N_S}$, and interaction vector $\boldsymbol{r}_u \in \mathbb{R}^{N_S}$. An attacker with access to the item ID matrix $E_{I,\text{id}}$, privacy-protected embedding e_u^{pp}, and partial interaction data $\boldsymbol{r}_u^{\text{pp}} = \boldsymbol{r}_u \odot [1, 1, 0, \ldots, 0]$ may use the NeuMF model to infer user embedding $\mathbf{e}_u^{\text{atk}}$ and item matrix $\mathbf{E}_I^{\text{atk}}$. We measure privacy leakage using Normalized Entropy (priv_{NE}) and Conditional Privacy Loss (priv_{CPL}), where higher priv_{NE} indicates stronger protection and higher priv_{CPL} indicates greater leakage.

To evaluate explanation performance, we use the Unique Sentence Ratio (USR) (Li, Zhang, and Chen 2020), defined as $USR = \frac{|\mathcal{E}|}{N}$, where $\mathcal{E}$ is the set of unique sentences generated and N is the number of test samples. A low USR indicates many identical explanations, suggesting poor diversity, while a high USR reflects better contextual adaptation and diverse explanations.

Implementation Details. Our HCR model is implemented in PyTorch with key parameters tuned for optimal performance. We use the Adam optimizer with a learning rate of 0.001 and a regularization coefficient of 0.0001. Early stopping and validation follow LightGCN's approach. For the self-supervised task, we set the temperature $\tau_{ss} = 0.5$. In

the privacy-preserving synthesizer, we set a fixed $\lambda_{pp} = 0.4$. For explanations, we use a pretrained GPT-2 model from huggingface with Byte Pair Encoding (BPE) to handle rare words, setting the length to 20 BPE tokens and embedding size to 768.

Performance Comparison (Q1)

We report the average comparison results from 5 runs on the Healthcare and Amazon datasets in the Tab. 2, where A → B means transferring knowledge of domain A to domain B.

PPCDR Algorithms Performance: HCR outperforms advanced privacy-preserving cross-domain benchmarks, achieving nearly 30% better privacy protection on the Healthcare dataset. This demonstrates its robust privacy-preserving capabilities while considering user preferences, making it ideal for privacy-sensitive domains like healthcare. However, its recommendation performance still falls short compared to multimodal cross-domain benchmarks, highlighting the importance of effective modality use in preference modeling.

Superiority of Multimodal CDR Algorithms: Multimodal algorithms generally achieve better recommendation accuracy by effectively using modality information for comprehensive user preference modeling and cross-domain knowledge transfer. However, modality noise poses challenges. SEMI struggles with noise filtering, and PMMRec's contrastive learning may overlook behavior-driven fusion. HCR overcomes this by using a shallow graph neural network for high-order modality extraction and adaptive fusion through behavior-driven guidance, integrating modality preferences while minimizing noise.

Comparison with Multimodal CDR Algorithms: Compared to the optimal benchmark PMMRec, HCR significantly enhances privacy metrics through Gumbel-Softmax sampling and item replacement. But this kind of shifting may pay some cost, leading to recommendation accuracy loss on certain datasets such as Amazon Arts→Fashion. PMMRec excels at aligning modality features for detailed item representations but struggles with complementary modalities. HCR, with its modality-aware encoder and self-supervised ID embeddings, better integrates diverse modalities and user preferences.

Ablation Study (Q2)

Removing the SS module (Ours w/o SS) significantly reduces recommendation performance, emphasizing the importance of modality information in cross-domain tasks. The proposed self-supervised encoder enhances user behavior modeling and knowledge transfer, capturing modality-specific details while avoiding noise without compromising privacy protection.

Conversely, removing the privacy module (Ours w/o PP) boosts recommendation performance but increases information leakage, underscoring the need for strong privacy mechanisms. HCR effectively balances privacy with performance by carefully managing information interference.

Figure 2: Empirical investigation and visual examples.

Empirical Study of Explanation (Q3)

We assessed the impact of interpretability on medical and health service recommendations by comparing explanations in the e-commerce and healthcare domains. We enhanced the PMMRec model with the HCR model's explanation feature, creating PMMRec+. Fifty users were split into two groups, U1 and U2, each keeping five historical items. Group U1 used HCR, and Group U2 used PMMRec+, with group assignments randomized to prevent bias.

Experiments were conducted on two datasets: (Amazon) Software → Game and (Healthcare) Video → Doctor, with 500 participants split into A/B groups. Group A received only recommendations, while Group B received recommendations with explanations. Participants rated the relevance of recommendations on a scale of 1 to 5, unaware of the recommendation sources.

We compared the average scores of HCR and PMMRec+ across different scenarios and demonstrated the effect of recommendation explanations (top 3 explanations for space limits) through an example in Fig 2. Key findings include:

- Without explanations, both models scored similarly, with PMMRec+ slightly ahead. Adding explanations improved scores for both models, enhancing user understanding and acceptance.
- Explanations had a greater impact in healthcare than in e-commerce, where images sufficed for user engagement.
- PMMRec+ performed better in comprehension even without explanations, likely due to its strong visual interpretation.
- Healthcare recommendations were harder for users to grasp, highlighting the need for explanations, which significantly boosted doctor recommendation acceptance.

In summary, text explanations improved user understanding and engagement, especially in healthcare, where interpretability is crucial.

Sensitivity Analysis (Q4)

We evaluated the privacy-preserving synthesizer by measuring recommendation precision and privacy protection in various settings, illustrated as Fig. 3. As sensitivity γ decreases and replacement ratio k increases, recommendation performance worsens while privacy protection improves, show-

	(Healthcare) Video → Doctor				(Healthcare) Doctor → Video				(Amazon) Arts → Fashion				(Amazon) Software → Game			
	HR	NDCG	NE↑	CPL↓	HR	NDCG	NE↑	CPL↓	HR	NDCG	NE↑	CPL↓	HR	NDCG	NE↑	CPL↓
CoNet	.2762	.1817	.0988	.3511	.2817	.1865	.0971	.3640	.3194	.2065	.0896	.3077	.3034	.1984	.0902	.2660
CCTL	.3126	.2115	.1270	.2359	.3256	.2176	.1278	.2342	.3702	.2478	.1265	.2301	.3508	.2315	.1211	.2877
PriCDR	.4142	.3061	.5490	.1510	.4231	.3168	.5486	.1518	.5133	.4065	.5511	.1530	.4558	.3370	.5385	.2174
DPSMRec	.4356	.3276	.5502	.1456	.4447	.3390	.5503	.1462	.5437	.4418	.5591	.1489	.4789	.3607	.5396	.2102
PPGenCDR	.4403	.3275	.5569	.1450	.4487	.3390	.5570	.1567	.5501	.4410	.5651	.1481	.4836	.3610	.5433	.2081
SEMI	.4392	.3389	.2216	.2166	.4485	.3514	.2191	.2147	.5428	.4580	.2231	.2056	.4831	.3764	.2201	.2580
PMMRec	.4449	.3526	.2803	.1875	.4538	.3510	.2777	.1890	.5517	.4615	.2855	.1886	.4887	.3803	.2744	.2543
Ours w/o SS	.4345	.3379	.7211	.1109	.4480	.3501	.7140	.1156	.5489	.4402	.7056	.1189	.4832	.3771	.6712	.1298
Ours w/o PP	**.4465**	**.3574**	.2928	.1981	**.4551**	**.3557**	.3044	.1915	**.5523**	**.4620**	.3056	.1961	**.4902**	**.3824**	.2659	.1843
Ours	.4453	.3560	**.7220**	**.1033**	.4542	.3543	**.7199**	**.1102**	.5508	.4601	**.7128**	**.1052**	.4890	.3809	**.6852**	**.1204**
Impro.(%)	0.09	0.96	157.58	44.91	0.09	0.94	159.24	41.69	-0.16	-0.3	149.67	44.22	0.06	0.16	149.71	52.65

Table 2: The recommendation performance and privacy protection of various recommendation models on cross-domain datasets.

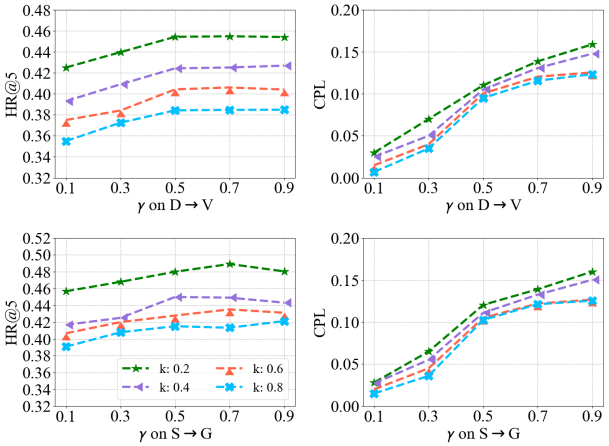

Figure 3: Performance comparison of HCR on different privacy settings. k and γ represent replacement ratio and sensitivity, respectively.

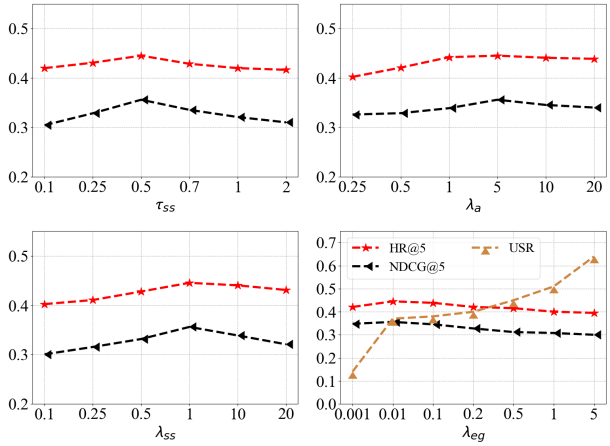

Figure 4: Results on (Healthcare) Video → Doctor.

ing the synthesizer's customizable privacy features. Despite high privacy settings, the data still reflects user preferences. However, at high sensitivity ($\gamma = 0.9$), performance suffers due to reduced data diversity.

Optimal sensitivity varies by scenario: $\gamma = 0.5$ for (Healthcare) Video → Doctor and $\gamma = 0.7$ for (Amazon) Software → Game. Sensitivity impacts healthcare recommendations less than gaming, likely due to dataset complexity. Similarly, the replacement ratio k affects the Amazon dataset less due to denser interactions. The HCR model effectively balances privacy and performance across scenarios, maintaining strong privacy while ensuring competitive accuracy.

Fig. 4 illustrates hyperparameter effects on the (Healthcare) Video → Doctor scenario. HCR performs best with specific settings for alignment (λ_a), temperature (τ_{ss}), SS (λ_{ss}), and explanation generation (λ_{eg}). Performance is stable across λ_a values, reflecting good use of source domain knowledge. Bell-shaped curves for τ_{ss} and λ_{ss} emphasize the need for careful hyperparameter tuning. Higher λ_{eg} values boost interpretability but lower accuracy, suggesting a trade-off that can be optimized.

Conclusion and Future Work

This paper introduces an interpretable PPCDR framework tailored for healthcare. By integrating the UCSDA and a PtLLM, we effectively address the dual challenges of protecting sensitive data and providing patient-friendly explanations. Its performance, validated across multiple datasets using both traditional and healthcare-specific metrics, highlights its scalability and robustness.

Although HCR offers a significant advancement in PPCDRs, limitations remain. The reliance on public datasets suggests a need for testing in more diverse real-world healthcare settings, where cross-domain data sharing and interoperability are complicated. Future work should also explore further improvements in privacy techniques and multimodal data integration to optimize recommendation quality.

In conclusion, our framework contributes to the growing demand for interpretable and privacy-conscious recommendations in healthcare. Further research is needed to refine these systems, ensuring they continue to meet the evolving demands of this sensitive domain.

Acknowledgments

The Fundamental Research Funds for the Central Universities and the Research Funds of Renmin University of China (23XNL017). Supported by the Fundamental Research Funds for the Central Universities, and the Research Funds of Renmin University of China (23XNH147).

References

Cao, J.; Sheng, J.; Cong, X.; Liu, T.; and Wang, B. 2022. Cross-domain recommendation to cold-start users via variational information bottleneck. In *ICDE*, 2209–2223.

Chen, C.; Wu, H.; Su, J.; Lyu, L.; Zheng, X.; and Wang, L. 2022. Differential private knowledge transfer for privacy-preserving cross-domain recommendation. In *WWW*, 1455–1465.

Chen, G.; Zhang, X.; Su, Y.; Lai, Y.; Xiang, J.; Zhang, J.; and Zheng, Y. 2023a. Win-win: a privacy-preserving federated framework for dual-target cross-domain recommendation. In *AAAI*, volume 37, 4149–4156.

Chen, X.; Chen, H.; Xu, H.; Zhang, Y.; Cao, Y.; Qin, Z.; and Zha, H. 2019. Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation. In *ACM SIGIR*, 765–774.

Chen, X.; Zhang, Y.; Tsang, I. W.; Pan, Y.; and Su, J. 2023b. Toward equivalent transformation of user preferences in cross domain recommendation. *ACM Transactions on Information Systems*, 41(1): 1–31.

Conneau, A.; Khandelwal, K.; and Goyal, e. a. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Havrylov, S.; and Titov, I. 2017. Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. *Advances in neural information processing systems*, 30.

He, R.; and McAuley, J. 2016. VBPR: visual bayesian personalized ranking from implicit feedback. In *AAAI*, volume 30.

He, X.; Deng, K.; Wang, X.; Li, Y.; Zhang, Y.; and Wang, M. 2020. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *ACM SIGIR*, 639–648.

Hu, G.; Zhang, Y.; and Yang, Q. 2018. Conet: Collaborative cross networks for cross-domain recommendation. In *ACM CIKM*, 667–676.

Jiang, H.; Mi, Z.; and Xu, W. 2024. Online Medical Consultation Service–Oriented Recommendations: Systematic Review. *Journal of Medical Internet Research*, 26: e46073.

Kemertas, M.; Pishdad, L.; Derpanis, K. G.; and Fazly, A. 2020. Rankmi: A mutual information maximizing ranking loss. In *CVPR*, 14362–14371.

Lei, C.; Liu, Y.; Zhang, L.; Wang, G.; Tang, H.; Li, H.; and Miao, C. 2021. Semi: A sequential multi-modal information transfer network for e-commerce micro-video recommendations. In *ACM SIGKDD*, 3161–3171.

Li, L.; Zhang, Y.; and Chen, L. 2020. Generate neural template explanations for recommendation. In *ACM CIKM*, 755–764.

Li, Y.; Du, H.; Ni, Y.; Zhao, P.; Guo, Q.; Yuan, F.; and Zhou, X. 2024. Multi-modality is all you need for transferable recommender systems. In *ICDE*, 5008–5021.

Liao, X.; Liu, W.; Zheng, X.; Yao, B.; and Chen, C. 2023. Ppgencdr: A stable and robust framework for privacy-preserving cross-domain recommendation. In *AAAI*, volume 37, 4453–4461.

Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; and Neubig, G. 2023a. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9): 1–35.

Liu, S.; Xu, S.; Yu, W.; Fu, Z.; Zhang, Y.; and Marian, A. 2021. FedCT: Federated collaborative transfer for recommendation. In *ACM SIGIR*, 716–725.

Liu, W.; Zheng, X.; Chen, C.; Hu, M.; Liao, X.; Wang, F.; Tan, Y.; Meng, D.; and Wang, J. 2023b. Differentially private sparse mapping for privacy-preserving cross domain recommendation. In *ACM MM*, 6243–6252.

McFee, B.; Raffel, C.; Liang, D.; Ellis, D. P.; McVicar, M.; Battenberg, E.; and Nieto, O. 2015. librosa: Audio and music signal analysis in python. In *SciPy*, 18–24.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Tao, Z.; Liu, X.; Xia, Y.; Wang, X.; Yang, L.; Huang, X.; and Chua, T.-S. 2022. Self-supervised learning for multimedia recommendation. *IEEE Transactions on Multimedia*, 25: 5107–5116.

Vaswani, A. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Wei, Y.; Wang, X.; Nie, L.; He, X.; and Chua, T.-S. 2020. Graph-refined convolutional network for multimedia recommendation with implicit feedback. In *ACM MM*, 3541–3549.

Wei, Y.; Wang, X.; Nie, L.; He, X.; Hong, R.; and Chua, T.-S. 2019. MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video. In *ACM MM*, 1437–1445.

Zhang, J.; Zhu, Y.; Liu, Q.; Wu, S.; Wang, S.; and Wang, L. 2021. Mining latent structures for multimedia recommendation. In *ACM MM*, 3872–3880.

Zhang, W.; Zhang, P.; Zhang, B.; Wang, X.; and Wang, D. 2023. A collaborative transfer learning framework for cross-domain recommendation. In *ACM SIGKDD*, 5576–5585.

Zhang, Y.; Lai, G.; Zhang, M.; Zhang, Y.; Liu, Y.; and Ma, S. 2014. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *ACM SIGIR*, 83–92.

Zhou, X.; Zhou, H.; Liu, Y.; Zeng, Z.; Miao, C.; Wang, P.; You, Y.; and Jiang, F. 2023. Bootstrap latent representations for multi-modal recommendation. In *WWW*, 845–854.