Breaking Down Market Barriers: Distilled Prompt-Tuning Approach for Cross-Market Recommendation

Leqi Zhang[1], Wayne Lu[1], Haiyang Zhang[1], Elliott Wen[2], Zhixuan Liang[3], Jia Wang[1] *

[1]Xi'an Jiaotong-Liverpool University, Suzhou, China
[2]University of Auckland, Auckland, New Zealand
[3]The Hong Kong Polytechnic University, Hong Kong SAR, China
leqi.zhang23@student.xjtlu.edu.cn, lu.wayne0603@gmail.com,
{jia.wang02, haiyang.zhang}@xjtlu.edu.cn,
elliott.wen@auckland.ac.nz, zhixuan.liang@connect.polyu.hk

Abstract

Cross-market recommendation (CMR) faces severe challenges from distribution shifts between data-rich source markets and sparse target markets. Existing methods rely on a pre-training and fine-tuning paradigm for knowledge transfer, yet suffer from two key limitations: i) the objective gap between pre-training and full-parameter fine-tuning causes loss of generalized knowledge from source markets; ii) the high computational costs of extensive fine-tuning hinder scalability. To this end, we propose *DCMPT*, a novel **D**istilled **C**ross-**M**arket **P**rompt-**T**uning approach. *DCMPT* reframes the problem under a more efficient pre-training and prompt-tuning paradigm. Instead of full fine-tuning, we adapt a pre-trained universal backbone by freezing its weights and injecting a minimal set of learnable prompts to form a "student" model. To effectively optimize these prompts on sparse data, we introduce a novel teacher-student architecture: a specialized "teacher" model, trained exclusively on the target market, provides dense, market-specific supervision. This guidance is delivered via a dual distillation strategy designed to transfer global ranking patterns and adapt to local consumer tastes. Extensive experiments on real-world market datasets demonstrate that *DCMPT* significantly outperforms state-of-the-art methods, achieving superior target market performance with substantial parameter-efficiency.

Introduction

With the rapid expansion of globalization and e-commerce, online shopping platforms are increasingly tasked with adapting their recommendation systems to diverse international markets. Traditional recommendation models, while effective in the same scenarios (Wang et al. 2018) or homogeneous markets (Lu and Yin 2025; Menglin et al. 2024; Cui et al. 2025), falter in cross-market scenarios where user behaviors, cultural norms, and item attributes vary significantly (Lin et al. 2024; Bonab et al. 2021). For instance, a user's preference for "holiday gifts" in Mexico might be shaped by local traditions like Day of the Dead, whereas in Germany, it could be driven by Christmas customs. These cultural discrepancies lead to substantial distribution shifts (Chen et al. 2023; Peng et al. 2025), enhancing data sparsity and cold-start problems when attempting to transfer interaction patterns across markets.

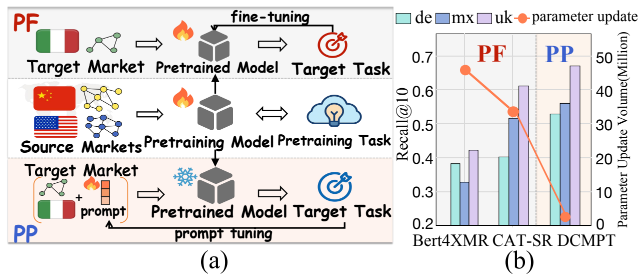


Figure 1: (a). PF vs PP models in Cross-Market Recommendation. (b). Comparison between PF and our PP approach in CMR. We evaluate these methods on three target markets(Germany, Mexico, the UK), measuring recall@10 and parameter update volume.

Efforts to address these challenges have included cross-domain recommendation (CDR) methods, which aim to bridge domain discrepancies through techniques like representation mapping (Li et al. 2024) or multi-task learning (Yan et al. 2022). However, CDR typically assumes some overlap in users or stable preference patterns across domains—assumptions that do not hold in cross-market settings where markets often lack shared users and exhibit stark differences in item popularity and cultural contexts (Du et al. 2024; Bonab et al. 2021). To better tackle these issues, cross-market recommendation (CMR) has emerged, employing a pretraining-and-fine-tuning (PF) paradigm. In this approach, models are pretrained on data from source markets and then fine-tuned for specific target markets (Wang et al. 2024). For example, FOREC pretrains a neural matrix factorization model on source markets and fine-tunes it using market-specific multi-layer perceptrons(MLPs) (Bonab et al. 2021).

Despite considerable progress, we argue that existing *pre-training and finetuning* paradigm suffers from sub-optimal performance in CMR due to three limitations: **First, they face a high risk of negative transfer.** The pre-training objective captures general patterns from source markets, which may misalign with the fine-tuning goal of adapting to target-specific preferences, leading to knowledge forget-

*Corresponding author.

ting or overfitting to spurious correlations (Gao et al. 2025; Wang et al. 2019). **Second, the fine-tuning process is computationally prohibitive.** Fine-tuning large models for each market incurs high computational and storage costs, limiting scalability across many markets. (Wang et al. 2025). **Third, PF paradigm leads to static, isolated models.** Maintaining a separate fine-tuned model for each market hinders cross-market knowledge sharing and increases deployment complexity. Fig. 1(a) illustrates differences between PF and our proposed *pretraining-then-prompt-tuning* (PP) paradigm. To empirically illustrate these challenges, we compare three PF-based methods with our PP approach across three target markets, evaluating recommendation performance and the volume of parameter updates required, as shown in Fig. 1(b). The results highlight that PF methods demand extensive parameter updates, increasing computational costs, and suffer from negative transfer due to distribution mismatches between source and target markets.

To overcome these limitations, we propose *DCMPT*, a novel **D**istilled **C**ross-**M**arket **P**rompt-**T**uning approach. *DCMPT* replaces the costly fine-tuning step with a more efficient **pre-training and prompt-tuning (PP)** paradigm. Instead of updating the entire model, we freeze the powerful pre-trained backbone and inject a small set of learnable **prompts**—lightweight vectors added directly to user/item embeddings—to steer the model's behavior for the target market. To optimize these prompts with target data that provides weak supervision, we propose a knowledge distillation architecture where a target-specific "teacher" model provides market-aware guidance for training the "student" prompts. Specifically, we propose two novel distillation strategies to ensure the prompts are optimized effectively: i) Weighted Ranking Distillation (WRD) guides prompt learning with tailored weights to capture global ranking patterns, and ii) Adaptive Market-aware Ranking Decoupled Distillation (AMRDD) provides nuanced and dynamic guidance to ensure the prompts align with target-specific preferences.

Related Work

Cross-Market Recommendation Cross-market recommendation (CMR) leverages existing user data and products from the source market to address challenges such as data sparsity or cold-start in the target market, thereby enhancing the recommendation accuracy and performance in the target market (Ozturk et al. 2024). There has been limited work on cross-market recommendation, with the most relevant approaches being FOREC(Bonab et al. 2021), Bert4XMR(Hu et al. 2024), and CAT-SR (Wang et al. 2024). They adopt a *pretraining and fine-tuning* paradigm to enhance robustness against sparsity, noise, market bias via distinct techniques—meta-learning, contrastive learning, and transferable attention, improving generalization. For instance, FOREC utilizes meta-learning to pre-train a NeuMF model and fine-tunes it with an MLP to improve performance in the target market. However, these approaches require heavy fine-tuning, incurring high computational costs and offering limited adaptability.

Prompt Learning Prompt-based learning adapts pre-trained models to downstream tasks with minimal parameter updates, gaining traction in NLP (Gao, Fisch, and Chen 2020) and graph-based recommendation (Geng et al. 2022). Recent works employ prompt embeddings to guide inter-domain knowledge transfer (Zhang et al. 2024a), contrastive objectives to refine prompts (Yi, Ounis, and Macdonald 2023), or generative models for dynamic prompt generation (Zhang et al. 2024b). Recent studies have begun to explore prompt-based learning in recommendation systems, particularly within graph-structured models(Geng et al. 2022). Some work integrates prompt embeddings or nodes into heterogeneous graphs to guide inter-domain knowledge transfer (Zhang et al. 2024a; Hao et al. 2024). However, applying prompt-tuning directly to CMR remains non-trivial due to sparse supervision and significant market heterogeneity. To enhance prompt learning in CMR, we design a distillation-enhanced prompt-tuning approach with two distillation strategies.

Problem Formulation

In this paper, we focus on addressing the ranking-based recommendation task in a cross-market setting. Let $\mathcal{M} = \{M_1, \ldots, M_a\}$ be a set of a markets. Let $\mathcal{U}$ and $\mathcal{I}$ be the global sets of users and items, respectively. We make two standard assumptions: user sets are disjoint across markets (i.e., $U_i \cap U_j = \emptyset$ for $i \neq j$), and item sets are partially overlapping, where an item shared across markets maintains a single, unified representation. Each market $M_k \in \mathcal{M}$ is represented as a bipartite graph of interactions $\mathcal{G}_k = (U_k, I_k, \mathcal{E}_k)$, where $U_k \subseteq \mathcal{U}$ and $I_k \subseteq \mathcal{I}$ are the market-specific user and item sets, and $\mathcal{E}_k$ represents the observed interactions. We designate a single market as the target market, M_T, with its graph $\mathcal{G}_T$, and the remaining $a-1$ as source markets $\{M_{S_i}\}$ with graphs $\{\mathcal{G}_{S_i}\}$.

Problem Definition. Given the set of all market interaction graphs $\{\mathcal{G}_{S_1}, ..., \mathcal{G}_{S_{a-1}}, \mathcal{G}_T\}$, our objective is to learn a model $\mathcal{F}$ that, for each user $u \in U_T$, can generate an accurate and personalized ranked list of items. The core challenge is to do so effectively and efficiently, mitigating the data sparsity of the target market M_T.

Methodology

Contemporary CMR systems predominantly employ a *pre-training and fine-tuning* (PF) paradigm, where models pre-trained on multi-market data undergo full parameter updates during target adaptation. While effective in knowledge transfer, this approach faces two critical limitations: (1) Distribution misalignment between source and target markets induces negative transfer, and (2) Extensive parameter updates during fine-tuning create an objective gap that discards valuable cross-market knowledge. By freezing the pre-trained backbone and only tuning a small set of prompts ψ, *pre-training and prompt-tuning* (PP) offers a parameter-efficient solution that mitigates the objective gap (Tong et al. 2025). To further capture market-specific knowledge, as illustrated in Fig. 2, our DCMPT operates in two stages: (1) Pre-training, where a backbone learns from combined

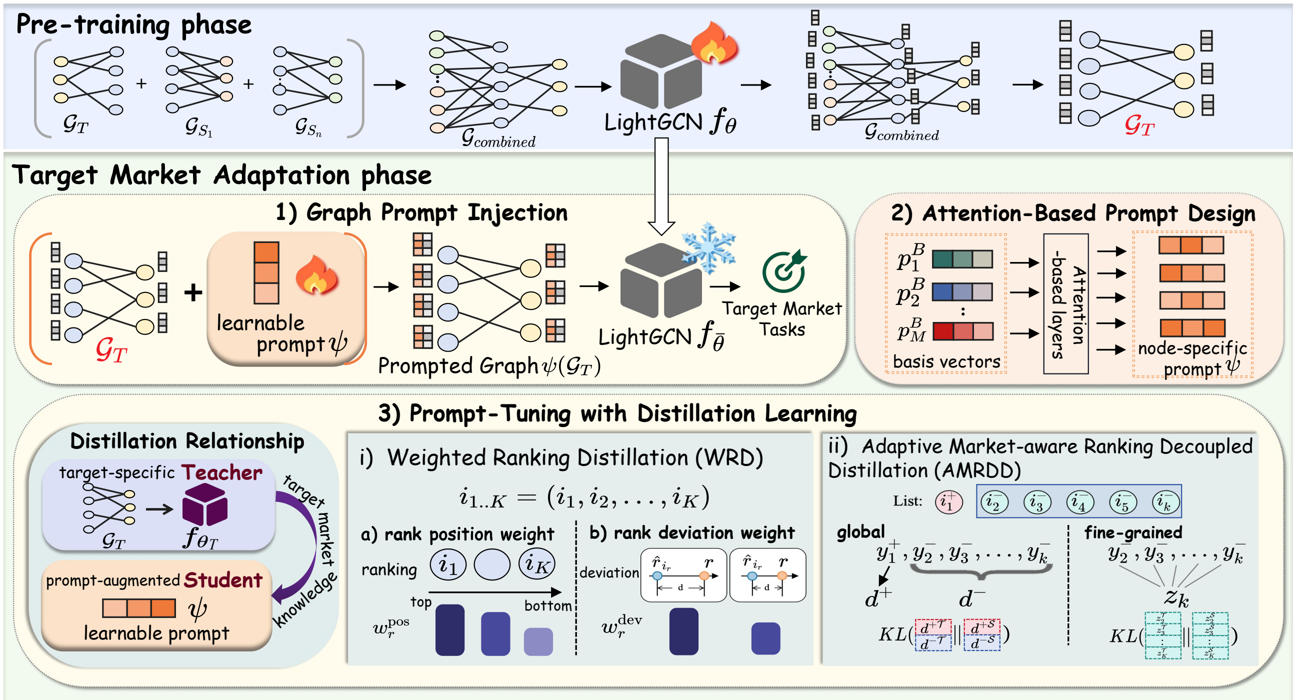


Figure 2: The overview of our method DCMPT, which consists of pretraining phase and target market adaptation phase designed to enable efficient cross-market recommendation. In the adaptation phase, learnable prompts are injected into the frozen backbone model and optimized through joint training loss to tailor it for the target market.

multi-market data, and (2) Target market adaptation, it introduces a target-specialized teacher model $\mathcal{T}$ trained exclusively on target market data, guiding prompt learning (student) through dual-task optimization of $\hat{\mathbf{y}}_{\text{stu}} \Leftrightarrow \mathbf{y}$ and $\hat{\mathbf{y}}_{\text{stu}} \Leftrightarrow \hat{\mathbf{y}}_{\mathcal{T}}$.

$$\text{PF:} \quad \underbrace{\mathcal{F}(\mathcal{X}_T; \Theta = \mathcal{F}_{\text{pre}}(\mathcal{X}_{S\cup T}))}_{\text{Recommendation Procedure}} \underset{\text{Output}}{\overset{\text{Gradients}}{\leftrightarrows}} \underbrace{(\hat{\mathbf{y}} \Leftrightarrow \mathbf{y})}_{\text{Loss Calculation}} \tag{1}$$

$$\text{PP:} \underbrace{\mathcal{F}(\mathcal{X}_T + \psi; \bar{\Theta} = \mathcal{F}_{\text{pre}}(\mathcal{X}_{S\cup T}))}_{\text{Recommendation Procedure (frozen } \bar{\Theta})} \underset{\text{Output}}{\overset{\text{Gradients}}{\leftrightarrows}} \underbrace{(\hat{\mathbf{y}} \Leftrightarrow \mathbf{y})}_{\text{Loss Calculation}} \tag{2}$$

$$\text{Ours:} \underbrace{\mathcal{F}_{\text{stu}}(\mathcal{X}_T + \psi; \; \bar{\Theta} = \mathcal{F}_{\text{pre}}(\mathcal{X}_{S\cup T}))}_{\text{Recommendation Procedure (frozen } \bar{\Theta})} \underset{\text{Output}}{\overset{\text{Gradients}}{\leftrightarrows}} \begin{cases} \hat{\mathbf{y}}_{\text{stu}} \Leftrightarrow \mathbf{y} \\ \hat{\mathbf{y}}_{\text{stu}} \Leftrightarrow \hat{\mathbf{y}}_{\mathcal{T}} \end{cases} \tag{3}$$

Pre-training Phase

The pretraining phase learns generalized user-item interaction patterns. It combines graphs from source markets and the target market. We pretrain a LightGCN model on a combined graph $\mathcal{G}_{comb}$. This graph integrates knowledge across multiple markets.

Graph Concatenation and Node Alignment We construct $\mathcal{G}_{comb}$ by merging the target market graph $\mathcal{G}_T$ with source market graphs $\{\mathcal{G}_{S_i}\}_{i=1}^{a-1}$. Shared items across markets are aligned as identical nodes. This ensures structural consistency. Their interactions are aggregated. Alignment uses a hash-based method with $O(N+E)$ complexity. This approach scales efficiently to billion-node graphs. The combined graph is defined as:

$$\mathcal{G}_{\text{comb}} = \left(U_T \cup \bigcup_{i=1}^{a-1} U_{S_i}, I_T \cup \bigcup_{i=1}^{a-1} I_{S_i}, \mathcal{E}_T \cup \bigcup_{i=1}^{a-1} \mathcal{E}_{S_i} \right)$$

LightGCN Pre-training On this global graph, we pretrain a LightGCN model to learn initial embeddings. The objective is to capture universal collaborative signals present in the combined data:

$$\theta^* = \arg\min_{\theta} \sum_{(u,i)\in\mathcal{E}_{\text{comb}}} \mathcal{L}\left(f_\theta(u, i \mid \mathcal{G}_{\text{comb}}), y_{u,i}\right), \tag{4}$$

where f_θ is the LightGCN scoring function with parameters θ, and $y_{u,i}$ indicates an observed interaction. After this phase, the learned parameters θ^* are **frozen**, denoted as $\bar{\theta}$, to serve as a robust, generalized backbone for adaptation.

Target Market Adaptation Phase

In this phase, we adapt a pre-trained Graph Neural Network (GNN) to a specific target market M_T. The GNN, with parameters θ frozen after training on the combined graph $\mathcal{G}_{\text{comb}}$ of multiple markets, serves as a robust backbone capturing generalized user-item interaction patterns. For adaptation, we first isolate the subgraph corresponding to the target market, $\mathcal{G}_T = (\mathcal{U}_T, \mathcal{I}_T, \mathcal{E}_T)$, from the global graph $\mathcal{G}_{\text{comb}}$. We then introduce learnable prompts, lightweight parameter vectors, added to the node feature matrix of $\mathcal{G}_T$. These prompts modify the input to the GNN, creating a student model tailored to the target market. Importantly, only the prompt parameters ψ are tuned during adaptation, keeping the backbone parameters θ frozen. This parameter-efficient approach mitigates the objective gap and reduces the risk of negative transfer, which can occur in traditional fine-tuning due to distribution mismatches.

Graph Prompt Injection Specifically, for the target market graph $\mathcal{G}_T = (A_T, X_T)$, we inject prompts by modifying the node feature matrix to $X_T' = X_T + \psi$, where ψ is a matrix of learnable prompt vectors. This approach is grounded in prior work (Fang et al. 2024), which demonstrates that such feature-based prompts can effectively approximate structural changes in the graph (i.e., $f(A_T, X_T + \psi) \approx f(A_T', X_T)$). Consequently, prompt-tuning enables the frozen GNN to emulate a structure adapted to $\mathcal{M}_T$, without requiring costly retraining.

Attention-Based Prompt Design To efficiently generate node-specific prompts, we employ an attention-based mechanism. We maintain a small set of M learnable basis prompts $\mathcal{P}^B = \{p_1^B, \ldots, p_M^B\}$, each in $\mathbb{R}^D$, where D is the feature dimension. For each node n, the prompt $\psi_n \in \mathbb{R}^D$ is computed as a weighted sum of these bases:

$$\psi_n = \sum_{k=1}^{M} \alpha_{n,k} p_k^B, \quad \text{where} \quad \alpha_{n,k} = \frac{\exp(\mathbf{a}_k^\top \mathbf{e}_n)}{\sum_{l=1}^{M} \exp(\mathbf{a}_l^\top \mathbf{e}_n)},$$

with learnable attention vectors $\mathbf{a}_k \in \mathbb{R}^D$ and the node's input feature vector $\mathbf{e}_n = X_T[n]$. This design ensures scalability and computational efficiency.

However, optimizing these prompts using only the sparse interaction data from M_T via a standard Bayesian Personalized Ranking (BPR) loss often results in weak and noisy supervision, limiting their effectiveness in capturing target-specific preferences. To address this, we introduce a knowledge distillation approach that leverages a target-market-specific teacher model to provide rich, market-aware supervisory signals, guiding the prompt learning process. We detail this approach in the following subsection.

Prompt-Tuning with Distillation Learning

Our distillation approach addresses sparse supervision in target markets through dual knowledge transfer from a market-specific teacher ($\mathcal{T}$) to a prompt-augmented student ($\mathcal{S}$). The teacher model $\mathcal{T} = f_{\theta_T}(\mathcal{G}_T)$ is trained end-to-end on M_T, while the student combines frozen backbone parameters $\bar{\theta}$ from $\mathcal{G}_{comb}$ with learnable prompts ψ through $\mathcal{S} = f_{\bar{\theta}}(\psi(\mathcal{G}_T))$. The objective is to train the prompts of the student model so that its output closely mimics that of the teacher model. We implement this through two novel distillation strategies:

Weighted Ranking Distillation (WRD) To address the unique challenges of ranking tasks, WRD employs tailored weight mechanisms designed to reduce distributional bias between source and target markets. We select the top K items ranked by the teacher model as reference points to guide the student model's learning. The WRD loss is formulated as:

$$\mathcal{L}_{\text{WRD}} = -\sum_{r=1}^{K} w_r \cdot \log\left(\sigma\left(\hat{y}_{i_r}\right)\right), \tag{5}$$

where $\hat{y}_{i_r}$ represents the student model's predicted score for the item i_r, given by:

$$\hat{y}_{i_r} = f_{\bar{\theta}}(u, i_r \mid \psi(\mathcal{G}_T)). \tag{6}$$

In this setup, i_r denotes the r-th ranked item according to the teacher model, and w_r is a composite weight consisting of two parts: the rank position weight and the rank deviation weight.

- **Rank Position Weight**: The rank position weight w_r^{pos} is designed based on the assumption that items ranked higher by the teacher model are more important for recommendation quality in the target market. To reflect this, we use a softmax function to compute the weights:

$$w_r^{\text{pos}} = \frac{\exp(-r/\lambda)}{\sum_{k=1}^{K} \exp(-k/\lambda)}, \tag{7}$$

  where r is the rank of item i_r in the teacher's predictions, and λ is a temperature parameter that controls the weight distribution.
- **Rank Deviation Weight**: The rank deviation weight w_r^{dev} is introduced to account for the discrepancy between the teacher's and student's rankings for each item. Specifically, if the student model's predicted rank $\hat{r}_{i_r}$ for item i_r differs significantly from the teacher's rank r, a higher weight is assigned to emphasize the correction of this discrepancy. The rank deviation weight is defined as:

$$w_r^{\text{dev}} = \sigma\left(\mu \cdot \left(\hat{r}_{i_r} - r\right)\right), \tag{8}$$

  where σ is the sigmoid function, $\hat{r}_{i_r}$ is the student's predicted rank for i_r, and μ is a hyperparameter that adjusts the sensitivity of the weight to ranking differences.

The final weight w_r for each item i_r is obtained by combining the rank position weight and the rank deviation weight. Specifically, it is calculated as:

$$w_r = \frac{w_r^{\text{pos}} \cdot w_r^{\text{dev}}}{\sum_{j=1}^{K} w_j^{\text{pos}} \cdot w_j^{\text{dev}}}. \tag{9}$$

Adaptive Market-aware Ranking Decoupled Distillation (AMRDD) In CMR tasks, different source markets may provide varying levels of useful signals, with some introducing noise. Moreover, target markets often suffer from data sparsity, which limits the ability to learn robust behavioral patterns and model reliable user preferences. To address these issues, we introduce *AMRDD* which leverages both positive and negative samples to enrich the training data by extracting implicit preferences from unobserved items. Additionally, AMRDD decouples user preferences into global and fine-grained components, allowing for more nuanced modeling. A key feature of AMRDD is its dynamic adjustment of the influence of each source market, ensuring that more relevant markets contribute more significantly to the distillation process.

Decoupling Ranking Preferences We design the *AMRDD* loss into the disentangled parts to help the student model align with target-specific teacher. For each user u in a given market, we construct a ranked list of items $\mathbf{i} = [i_1^+, i_2^-, \ldots, i_K^-]$. This list contains one **positive item** i_1^+, with which the user has interacted, and $K-1$ **negative items** $\{i_k^-\}_{k=2}^K$, sampled randomly from items the user has not interacted with. Both the teacher and student models are scoring functions, $s^{\mathcal{T}}(u,i)$ and $s^{\mathcal{S}}(u,i)$, that produce a logit (a raw prediction score) for a given user-item pair. For the constructed item list $\mathbf{i}$, the models generate logit vectors, denoted as $\mathbf{y}^{\mathcal{T}}$ and $\mathbf{y}^{\mathcal{S}}$ respectively, where $\mathbf{y} = [y_1^+, y_2^-, \ldots, y_K^-]$. The core of AMRDD is a distillation loss that aligns the student's predicted ranking distribution with the teacher's. We decouple this alignment into two distinct components.

a) Global Preference Alignment. First, we capture the coarse-grained user preference for observed items over unobserved ones. To achieve this, we define a **global preference distribution** $d = [d^+, d^-] \in \mathbb{R}^2$. This distribution represents the aggregated probability mass of the positive item versus all negative items. The components are calculated via a softmax over the logits in $\mathbf{y}$:

$$d^+ = \frac{\exp(y_1^+)}{\sum_{k=1}^K \exp(y_k)}, \quad d^- = \frac{\sum_{k=2}^K \exp(y_k^-)}{\sum_{k=1}^K \exp(y_k)}. \tag{10}$$

The global alignment loss $\mathcal{L}_{\text{global}}$ then minimizes the KL divergence between the teacher's global preference $d^{\mathcal{T}}$ and the student's $d^{\mathcal{S}}$:

$$\mathrm{KL}(d^{\mathcal{T}} \| d^{\mathcal{S}}) = d^{+\mathcal{T}} \log\left(\frac{d^{+\mathcal{T}}}{d^{+\mathcal{S}}}\right) + d^{-\mathcal{T}} \log\left(\frac{d^{-\mathcal{T}}}{d^{-\mathcal{S}}}\right). \tag{11}$$

b) Fine-grained Preference Alignment. Second, we distill the nuanced ranking preferences *among the unobserved items*. An absence of interaction does not always mean dislike, and the teacher model may have learned valuable implicit preferences. We define a **fine-grained preference distribution** $z = [z_2, z_3, \ldots, z_K] \in \mathbb{R}^{K-1}$ by applying a softmax over the logits of only the negative items:

$$z_k = \frac{\exp(y_k^-)}{\sum_{j=2}^K \exp(y_j^-)}, \quad \text{for } k = 2, \ldots, K. \tag{12}$$

The fine-grained alignment loss $\mathcal{L}_{\text{fine}}$ transfers the teacher's ranking knowledge $z^{\mathcal{T}}$ to the student $z^{\mathcal{S}}$:

$$\mathrm{KL}(z^{\mathcal{T}} \| z^{\mathcal{S}}) = \sum_{k=2}^K z_k^{\mathcal{T}} \log\left(\frac{z_k^{\mathcal{T}}}{z_k^{\mathcal{S}}}\right). \tag{13}$$

The Final AMRDD Objective We combine the global and fine-grained losses to form the full distillation objective. To dynamically control the contribution of each source market, we introduce an adaptive weight α_m. This weight is calculated based on the similarity between the student's global preference for market m ($d_m^{+\mathcal{S}}$) and the teacher's global preference ($d^{+\mathcal{T}}$). Source markets that are more similar to the target are assigned higher weights:

$$\alpha_m = \frac{\exp(-|d_m^{+\mathcal{S}} - d^{+\mathcal{T}}|)}{\sum_{k \in \mathcal{M}_s} \exp(-|d_k^{+\mathcal{S}} - d^{+\mathcal{T}}|)}. \tag{14}$$

The final AMRDD objective is the weighted sum of losses across all source markets, which we aim to minimize:

$$\mathcal{L}_{\text{AMRDD}} = \sum_{m \in \mathcal{M}_s} \alpha_m \Big[\mathrm{KL}(d^{\mathcal{T}} \| d_m^{\mathcal{S}}) + d^{-\mathcal{T}} \mathrm{KL}(z^{\mathcal{T}} \| z_m^{\mathcal{S}})\Big]. \tag{15}$$

Here, $d_m^{\mathcal{S}}$ and $z_m^{\mathcal{S}}$ are the preference distributions computed by the student model using a batch of data from source market m. Also, the fine-grained loss is weighted by $d^{-\mathcal{T}}$, the teacher's confidence in the unobserved set. This ensures we only focus on fine-grained ranking when the teacher model indicates a substantial preference signal exists among the negative items.

Overall Loss Function

To optimize the prompt learning in the adaptation phase, the final loss function integrates three components: BPR loss, WRD loss, and AMRDD loss, defined as follows:

$$\mathcal{L}_{Total} = \alpha \mathcal{L}_{BPR} + \beta \mathcal{L}_{WRD} + \gamma \mathcal{L}_{AMRDD}, \tag{16}$$

where α, β, and γ are hyperparameters used to balance the contributions of three losses.

Experiment

Our experimental evaluation aims to answer six questions:

- **RQ1 (Effectiveness)**: Does our method outperform competitive single-market, CDR, and CMR baselines?
- **RQ2 (Robustness)**: Can our model generalize well across markets with varying data densities (data-scarce)?
- **RQ3 (Ablation)**: What are the individual contributions of our approach's components: multi-market pretraining, graph prompts, and knowledge distillation?
- **RQ4 (Sensitivity)**: How does model performance vary with respect to key hyperparameters?
- **RQ5 (Effiency)**: How efficient is our method compared to other CMR methods?
- **RQ6 (Embedding)**: How does distillation learning affect the learned feature embeddings?

Experiment Setup

Dataset Following previous work (Bonab et al. 2021; Hu et al. 2024; Wang et al. 2024), we conduct experiments on the XMarket dataset, which contains user-item interaction data from multiple countries. We focus on the electronics category to capture diverse cultures, user preferences, and market sizes. The data is split into training, validation, and test sets using an 80/10/10 ratio.

Baselines To comprehensively validate the proposed method, we compare it with mainstream recommendation models, including methods for single-market, cross-domain (CDR), and cross-market (CMR) recommendation.

- **Single-Market Recommendation:** LightGCN (He et al. 2020) and NMF (He et al. 2017).
- **Cross-Domain Recommendation:** DTCDR (Zhu et al. 2019), NATR (Gao et al. 2019) and DCCDR (Zhang et al. 2023).

	Metrics	NDCG@10							Recall@10						
	Method	**de**	**jp**	**in**	**fr**	**ca**	**mx**	**uk**	**de**	**jp**	**in**	**fr**	**ca**	**mx**	**uk**
single	LightGCN	0.1389	0.1450	0.1583	0.0971	0.0280	0.0915	0.1331	0.1958	0.2258	0.2394	0.1730	0.0560	0.1575	0.2091
	NMF	0.1010	0.1801	0.1768	0.1093	0.0858	0.0539	0.1691	0.1684	0.2453	0.2533	0.1445	0.1567	0.1175	0.2579
CDR	DTCDR	0.1620	0.1827	0.2391	0.1829	0.1201	0.1328	0.2309	0.2287	0.2534	0.2784	0.2256	0.1830	0.1969	0.2841
	NATR	0.1587	0.2138	0.2264	0.1653	0.1055	0.1243	0.2190	0.2139	0.2803	0.2721	0.2075	0.1732	0.1930	0.2751
	DCCDR	0.1651	0.2225	0.2043	0.1882	0.1164	0.1379	0.2181	0.2348	0.2867	0.2436	0.2312	0.1795	0.2041	0.2625
CMR	MAML	0.2588	0.3826	0.5667	0.3739	0.1591	0.2786	0.4530	0.3353	0.4616	0.6478	0.5157	0.2090	0.4412	0.6169
	FOREC	0.3341	0.4243	0.5713	0.4534	0.2188	0.3358	0.4465	0.4470	0.4927	0.6386	0.5684	0.3466	0.4668	0.6048
	Bert4XMR	0.3345	0.4064	0.5630	0.2511	0.1969	0.1868	0.3499	0.3864	0.4689	0.6589	0.3458	0.3167	0.3084	0.4164
	CAT-SR	0.3134	0.4931	0.5673	0.4959	0.2258	0.3798	0.4768	0.4019	0.5546	0.6681	0.5917	0.3779	0.5321	0.6238
	DCMPT	**0.3898***	**0.5658***	**0.5850***	**0.5158***	**0.2640***	**0.4190***	**0.5024***	**0.5365***	**0.6279***	**0.6751***	**0.6176***	**0.4057***	**0.5702***	**0.6739***
Stat.	p-value	$4.6e^{-3}$	$1.5e^{-4}$	$2.5e^{-3}$	$4.7e^{-5}$	$2.1e^{-3}$	$3.1e^{-4}$	$5.5e^{-4}$	$1.2e^{-5}$	$9.8e^{-4}$	$1.1e^{-3}$	$8.5e^{-5}$	$3.3e^{-5}$	$9.6e^{-3}$	$3.6e^{-4}$

Table 1: Performance comparisons. Bold indicates the best result; * denotes statistically significant improvement ($p < 0.05$) over the runner-up (paired t-test); Underline indicates the second-best result.

- **Cross-Market Recommendation:** MAML (Finn, Abbeel, and Levine 2017), FOREC (Bonab et al. 2021), Bert4XMR (Hu et al. 2024), and CAT-SR (Wang et al. 2024).

Hyper-parameters and Environment In the experiments, we tune and optimize several hyperparameters across the models to ensure fair comparison. The Adam optimizer is used with learning rates selected from $\{10^{-1}, 5 \times 10^{-2}, 10^{-2}, 5 \times 10^{-3}, 10^{-3}, 10^{-4}\}$, with 10^{-3} for model training and 10^{-4} for prompt tuning. Embedding dimensions are set to 64, with three GCN layers for message passing. Each batch samples 50 user-item pairs (batch size = 16), and L_2 regularization (1×10^{-6}) with early stopping (patience = 10) is applied to prevent overfitting. All experiments are run with fixed random seeds for reproducibility.

Experimental Results and Discussion

Recommendation Performance Evaluations (RQ1) Table 1 demonstrates that our proposed DCMPT consistently and significantly outperforms all baselines across seven diverse markets on NDCG@10 and Recall@10 metrics. While specialized cross-market models (e.g., FOREC, CAT-SR) expectedly surpass single-market and cross-domain methods, DCMPT establishes a new state-of-the-art by a considerable margin. These results validate that the proposed distilled prompt-tuning paradigm effectively transfers generalized knowledge while enabling precise, parameter-efficient adaptation. DCMPT successfully aligns the model with unique market characteristics, proving its superiority for real-world cross-market recommendation.

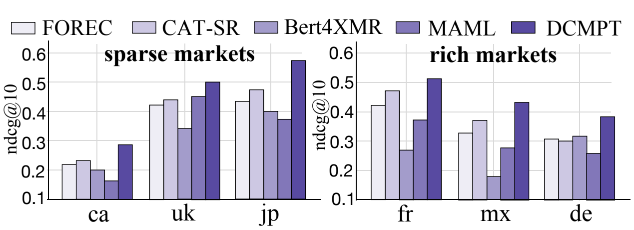


Figure 3: Performance comparison across sparse and rich markets.

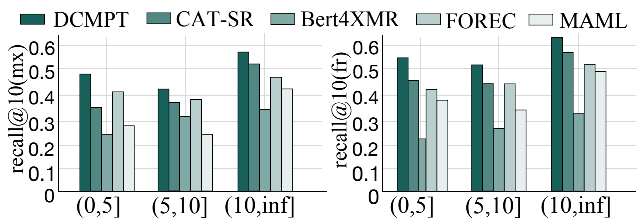


Figure 4: Cold-start performance over diferent item groups.

Robustness under Data-Sparse and Rich Markets (RQ2) To investigate the robustness across varying data densities, we partition the target markets into data-scarce markets (***ca***, ***uk***, ***jp***) and data-rich markets (***fr***, ***mx***, ***in***) based on interaction statistics. As shown in Fig. 3:

- **Data-sparse markets:** DCMPT consistently outperforms all baselines in sparse markets and exhibits more stable performance under low-resource conditions. In contrast, baseline methods such as CAT-SR and Bert4XMR show noticeable performance instability across sparse markets, indicating limited robustness in low-resource scenarios.

- **Data-rich markets:** DCMPT also achieves the best performance across three data-rich markets. For instance, it outperforms the strongest baseline by +7.2% Recall@10 on the ***mx*** market. These results show that DCMPT scales well in data-rich settings while maintaining strong generalization.

Data Sparsity Analysis (RQ2) To further evaluate the robustness of DCMPT under data sparsity compared to other models, we partition items into different groups based on their number of positive interactions in the training set. As shown in Fig. 4, DCMPT consistently achieves superior performance across all item groups on both the ***mx*** and ***fr*** markets. This demonstrates the effectiveness of our method in alleviating data sparsity, especially for cold-start scenarios. In contrast, traditional fine-tuning-based methods show a more substantial drop in performance on sparse items, indicating limited transfer capability under extreme sparsity.

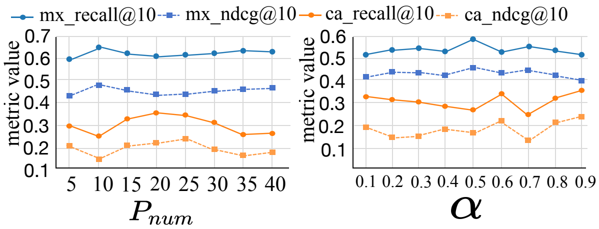


Figure 5: Parameter Sensitivity Analysis.

Ablation Study (RQ3) We conduct an ablation study with three components: (1) DCMPT w/o Pretrain, which removes multi-market pre-training; (2) DCMPT w/o Prompt, which removes the prompt tuning mechanism; and (3) DCMPT w/o Distillation, which removes the distillation losses. As shown in Table 2, model performance collapses without pre-training, highlighting its foundational role. Removing the prompts also causes a sharp decline, demonstrating their critical role in target-specific adaptation. The performance drop from removing distillation is more subtle but consistent, indicating its importance for refining and guiding the prompts effectively. Furthermore, both WRD and AMRDD strategies are crucial for optimal performance, underscoring the necessity of all components in DCMPT.

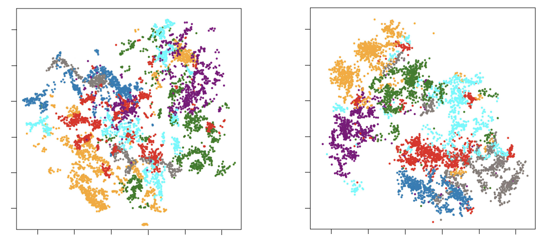
Variant	de	jp	in	fr	ca	mx	uk
DCMPT	0.390	0.566	0.585	0.516	0.264	0.419	0.502
w/o pretrain	0.139	0.165	0.178	0.107	0.028	0.092	0.133
w/o prompt	0.252	0.214	0.378	0.207	0.123	0.293	0.320
w/o distillation	0.336	0.440	0.498	0.435	0.186	0.310	0.392
w/o WRD	0.345	0.468	0.521	0.458	0.211	0.340	0.430
w/o AMRDD	0.353	0.487	0.530	0.464	0.218	0.345	0.421

Table 2: Ablation study on NDCG@10.

Parameter Analysis (RQ4) Fig. 5 reveals two critical insights on ***mx/ca*** markets.

- $\boldsymbol{P_{num}}$. The optimal prompt count reveals a trade-off between model capacity and data density. Data-rich markets like '***mx***' require fewer prompts, while sparse markets like '***ca***' leverage a higher capacity ($P_{\text{num}} = 20$). Performance degrades when exceeding these values, emphasizing the need for adaptive prompt sizing.
- $\boldsymbol{\alpha}$. With β and γ held constant, we examined the impact of varying α on model performance. Volatile markets (***mx***) prefer balanced task/distillation focus ($\alpha = 0.5$, +9.1% vs $\alpha = 0.1$), while stable markets (***ca***) need distillation dominance ($\alpha = 0.9$).

Training Efficiency (RQ5) To evaluate the training efficiency, we compare our prompt-tuning-based DCMPT with CMR methods under the *pretraining and finetuning*(PF) paradigm on the ***uk*** market. Table 3 shows that, models under the PF paradigm require substantially more trainable parameters and longer training time due to full backbone optimization. In contrast, DCMPT significantly reduces parameter overhead while achieving superior recommendation performance, demonstrating the efficiency and effectiveness of lightweight prompt tuning in cross-market adaptation.



(a) before distillation (b) after distillation

Figure 6: t-SNE feature visualization before and after prompt distillation.

Paradigm	Method	Params	Time/Epoch	Training Time	Recall@10
Finetuning	FOREC	22M	21.0s	25 min	0.6048
	Bert4XMR	47M	173.1s	62 min	0.4164
	CAT-SR	34M	42.2s	37 min	0.6238
Prompt-tuning	DCMPT	46K	8.6s	9 min	0.6739

Table 3: Comparison of training efficiency across PF and PP paradigms (on the uk market).

Impact of Distillation on Feature Embeddings (RQ6) Fig 6 shows the t-SNE visualization of the target feature embeddings before and after prompt distillation learning. Each color denotes a different market. Before distillation, the embeddings are entangled with unclear boundaries across markets, indicating insufficient market-specific adaptation. After distillation, the representations exhibit more compact intra-market clustering and clearer inter-market separation. This result suggests that our distillation strategies effectively guides the prompts to capture market-specific information.

Conclusion

This paper presents a novel cross-market recommendation method, DCMPT, which combines prompt tuning and distillation learning with graph neural networks to address the issue of negative transfer in CMR. Leveraging the prompt-tuning from NLP, our approach effectively captures market-specific preferences in the target market and reduces the need for extensive fine-tuning typically required in traditional CMR tasks. To further improve the model's adaptability, we introduce two distillation learning strategies. Extensive experiments demonstrate that DCMPT significantly outperforms existing baseline methods on key metrics such as NDCG and Recall. Future work may extend our approach to multilingual or multimodal CMR scenarios.

Acknowledgements

This research was supported in part by the National Natural Science Foundation of China (Grant No. 72401232), the Natural Science Foundation of the Jiangsu Higher Educational Institution of China (Grant No. 23KJB520037), and the XJTLU Research Development Fund under RDF-21-01-053.

References

Bonab, H.; Aliannejadi, M.; Vardasbi, A.; Kanoulas, E.; and Allan, J. 2021. Cross-market product recommendation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 110–119.

Chen, Y.; Yao, Y.; Chan, W. K. V.; Xiao, L.; Zhang, K.; Zhang, L.; and Ye, Y. 2023. CDR-Adapter: Learning Adapters to Dig Out More Transferring Ability for Cross-Domain Recommendation Models. *arXiv preprint arXiv:2311.02398*.

Cui, X.; Lu, W.; Tong, Y.; Li, Y.; and Zhao, Z. 2025. Diffusion-based multi-modal synergy interest network for click-through rate prediction. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 581–591.

Du, M.; Zhao, Z.; Zhang, X.; Xu, Y.; Wang, X.; and Yang, Q. 2024. Towards Robust Cross-Domain Recommendation with Joint Identifiability of User Preference. *arXiv preprint arXiv:2411.17361*.

Fang, T.; Zhang, Y.; Yang, Y.; Wang, C.; and Chen, L. 2024. Universal prompt tuning for graph neural networks. *Advances in Neural Information Processing Systems*, 36.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning*, 1126–1135.

Gao, C.; Chen, X.; Feng, F.; Zhao, K.; He, X.; Li, Y.; and Jin, D. 2019. Cross-domain recommendation without sharing user-relevant data. In *The world wide web conference*, 491–502.

Gao, J.; Wu, H.; Cheung, Y.-m.; Cao, J.; Yu, H.; and Zhang, Y. 2025. Mitigating Forgetting in Adapting Pre-trained Language Models to Text Processing Tasks via Consistency Alignment. In *Proceedings of the ACM on Web Conference 2025*, 3492–3504.

Gao, T.; Fisch, A.; and Chen, D. 2020. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*.

Geng, S.; Liu, S.; Fu, Z.; Ge, Y.; and Zhang, Y. 2022. Recommendation as language processing (rlp): A unified pretrain, personalized prompt & predict paradigm (p5). In *Proceedings of the 16th ACM Conference on Recommender Systems*, 299–315.

Hao, B.; Yang, C.; Guo, L.; Yu, J.; and Yin, H. 2024. Motif-based prompt learning for universal cross-domain recommendation. In *Proceedings of the 17th ACM international conference on web search and data mining*, 257–265.

He, X.; Deng, K.; Wang, X.; Li, Y.; Zhang, Y.; and Wang, M. 2020. 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*, 639–648. ACM.

He, X.; Liao, L.; Zhang, H.; Nie, L.; Hu, X.; and Chua, T.-S. 2017. Neural collaborative filtering. In *Proceedings of the 26th International Conference on World Wide Web (WWW)*, 173–182. International World Wide Web Conferences Steering Committee.

Hu, Z.; Nakagawa, S.; Cai, S.-M.; Ren, F.; and Deng, J. 2024. Enhancing cross-market recommendations by addressing negative transfer and leveraging item co-occurrences. *Information Systems*, 124: 102388.

Li, X.; Sheng, J.; Cao, J.; Zhang, W.; Li, Q.; and Liu, T. 2024. Cdrnp: Cross-domain recommendation to cold-start users via neural process. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, 378–386.

Lin, Z.; Ding, H.; Hoang, N. T.; Kveton, B.; Deoras, A.; and Wang, H. 2024. Pre-trained recommender systems: A causal debiasing perspective. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, 424–433.

Lu, W.; and Yin, L. 2025. DMMD4SR: Diffusion Model-based Multi-level Multimodal Denoising for Sequential Recommendation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 6363–6372.

Menglin, K.; Wang, J.; Pan, Y.; Zhang, H.; and Hou, M. 2024. C^2DR: Robust Cross-Domain Recommendation based on Causal Disentanglement. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, 341–349.

Ozturk, S.; Ercan, A. B.; Tugay, R.; and Gunduz Oguducu, S. 2024. Enhancing Cross-Market Recommendation System with Graph Isomorphism Networks: A Novel Approach to Personalized User Experience. In *Proceedings of the 2024 9th International Conference on Machine Learning Technologies*, 251–257.

Peng, T.; Yuan, H.; Zhang, Y.; Li, Y.; Dai, P.; Wang, Q.; Wang, S.; and Wu, W. 2025. Tagrec: Temporal-aware graph contrastive learning with theoretical augmentation for sequential recommendation. *IEEE Transactions on Knowledge and Data Engineering*.

Tong, Y.; Lu, W.; Cui, X.; Mao, Y.; and Zhao, Z. 2025. DAPT: Domain-Aware Prompt-Tuning for Multimodal Fake News Detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 7902–7911.

Wang, C.; Fan, Z.; Yang, L.; Yang, M.; Liu, X.; Liu, Z.; and Yu, P. 2024. Pre-Training with Transferable Attention for Addressing Market Shifts in Cross-Market Sequential Recommendation. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, to appear. Barcelona, Spain: ACM.

Wang, H.; Wang, J.; Wang, J.; Zhao, M.; Zhang, W.; Zhang, F.; Xie, X.; and Guo, M. 2018. Graphgan: Graph representation learning with generative adversarial nets. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Wang, L.; Chen, S.; Jiang, L.; Pan, S.; Cai, R.; Yang, S.; and Yang, F. 2025. Parameter-efficient fine-tuning in large language models: a survey of methodologies. *Artificial Intelligence Review*, 58(8): 227.

Wang, Z.; Dai, Z.; Póczos, B.; and Carbonell, J. 2019. Characterizing and avoiding negative transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11293–11302.

Yan, D.; Zhao, Y.; Yang, Z.; Jin, Y.; and Zhang, Y. 2022. FedCDR: Privacy-preserving federated cross-domain recommendation. *Digital Communications and Networks*, 8(4): 552–560.

Yi, Z.; Ounis, I.; and Macdonald, C. 2023. Contrastive graph prompt-tuning for cross-domain recommendation. *ACM Transactions on Information Systems*, 42(2): 1–28.

Zhang, H.; Shen, C.; Sun, X.; Tan, J.; Rong, Y.; Piao, C.; Cheng, H.; and Yi, L. 2024a. Adaptive Coordinators and Prompts on Heterogeneous Graphs for Cross-Domain Recommendations. *arXiv preprint arXiv:2410.11719*.

Zhang, P.; Yan, Y.; Zhang, X.; Kang, L.; Li, C.; Huang, F.; Wang, S.; and Kim, S. 2024b. Gpt4rec: Graph prompt tuning for streaming recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1774–1784.

Zhang, R.; Zang, T.; Zhu, Y.; Wang, C.; Wang, K.; and Yu, J. 2023. Disentangled contrastive learning for cross-domain recommendation. In *International Conference on Database Systems for Advanced Applications*, 163–178. Springer.

Zhu, F.; Chen, C.; Wang, Y.; Liu, G.; and Zheng, X. 2019. Dtcdr: A framework for dual-target cross-domain recommendation. In *Proceedings of the 28th ACM international conference on information and knowledge management*, 1533–1542.