Diversity Recommendation via Causal Deconfounding of Co-purchase Relations and Counterfactual Exposure

Jingmao Zhang[1*], Zhiting Zhao[1*], Yunqi Lin[1*], Jianghong Ma[1†], Tianjun Wei[2‡], Haijun Zhang[1], Xiaofeng Zhang[1]

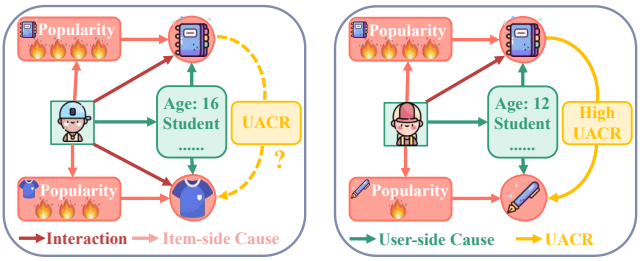
[1]Harbin Institute of Technology(Shenzhen), Shenzhen, China
[2]Nanyang Technological University, Singapore
220810426@stu.hit.edu.cn, 24s151185@stu.hit.edu.cn, 210110519@stu.hit.edu.cn,
majianghong@hit.edu.cn, tjwei2-c@my.cityu.edu.hk, hjzhang@hit.edu.cn, zhangxiaofeng@hit.edu.cn

Abstract

Beyond user-item modeling, item-to-item relationships are increasingly used to enhance recommendation. However, common methods largely rely on co-occurrence, making them prone to item popularity bias and user attributes, which degrades embedding quality and performance. Meanwhile, although diversity is acknowledged as a key aspect of recommendation quality, existing research offers limited attention to it, with a notable lack of causal perspectives and theoretical grounding. To address these challenges, we propose **Cadence**: Diversity Recommendation via **Ca**usal **De**confounding of Co-purchase Relations a**n**d **C**ounterfactual **E**xposure—a **plug-and-play** framework built upon LightGCN as the backbone, primarily designed to enhance recommendation diversity while preserving accuracy. First, we compute the Unbiased Asymmetric Co-purchase Relationship (UACR) between items—excluding item popularity and user attributes—to construct a deconfounded directed item graph, with an aggregation mechanism to refine embeddings. Second, we leverage UACR to identify diverse categories of items that exhibit strong causal relevance to a user's interacted items but have not yet been engaged with. We then simulate their behavior under high-exposure scenarios, thereby significantly enhancing recommendation diversity while preserving relevance. Extensive experiments on real-world datasets demonstrate that our method consistently outperforms state-of-the-art diversity models in both diversity and accuracy, and further validates its effectiveness, transferability, and efficiency over baselines.

Introduction

Traditional recommendation methods primarily focus on modeling user–item interactions (Wu et al. 2024; He et al. 2020; Yu et al. 2022; Wei, Ma, and Chow 2023a), whereas some studies have increasingly considered item–item relationships (Liu et al. 2025; Du et al. 2024; Zhou et al. 2025; Wei, Ma, and Chow 2023b; Wei, Chow, and Ma 2024; Zhang et al. 2025c). This line of work leverages structural item correlations to enhance representation capacity and improve generalization. By analyzing co-occurrence patterns in users' historical interactions, these methods uncover valuable item-level relationships that better inform user interest modeling.



(a) Co-purchase relations are confounded by item popularity and user attributes.

(b) High UACR shows pen relevance, but lower popularity may discourage user interaction.

Figure 1: Example of casual relations in recommendation.

Limitations of Co-occurrence-Based Item Modeling. Conventional methods for modeling item relationships predominantly rely on co-occurrence statistics derived from observational data (Chou and Cheng 2023; Yan et al. 2022; Sugahara, Yamasaki, and Okamoto 2024). However, these correlations are often confounded by factors such as item popularity and user attributes, which obscure true item dependencies. For instance, as shown in Figure 1a, (1) a user may purchase both a book and a jacket due to their popularity, or (2) because, as a 16-year-old student, they are inherently inclined to prefer both, even though the items are not intrinsically related. These spurious signals can mislead recommender systems into forming inaccurate dependencies, ultimately degrading recommendation quality.

To address this, a more principled approach is required—one that can disentangle confounding influences and recover the true causal pathways governing user behavior. Recent advances in causal inference (Luo et al. 2025; Huang et al. 2025) provide a theoretical foundation for such modeling, enabling the simulation of interventions and counterfactual scenarios (Peters, Janzing, and Schölkopf 2017). Within this framework, we introduce the

*These authors contributed equally.
†Corresponding author.
‡Corresponding author.

concept of ***Unbiased Asymmetric Co-purchase Relationships (UACR)***, a class of directional item–item associations that capture the causal influence an item exerts on another, free from spurious statistical correlations.

Limitations of Counterfactual Reasoning in Diversity Modeling. Counterfactual reasoning (Pearl 2009), a core tool in causal inference, models hypothetical scenarios to capture shifts in user behavior (Tang et al. 2025). It has demonstrated strong potential in tasks such as causal representation learning, bias correction, and ranking optimization, and is emerging as a promising foundation for robust and explainable recommender systems. Although recent research has increasingly applied counterfactual methods to improve accuracy (Zhao et al. 2025; Wang et al. 2023; Liu et al. 2022), fairness (Wang et al. 2024; Zhu et al. 2023), and explainability (Tan et al. 2021; Baklanov 2024), its potential for enhancing recommendation diversity remains largely overlooked. In practice, user interests are multifaceted, and optimizing for accuracy often leads to narrow, homogenized recommendation results, diminishing user satisfaction and limiting system effectiveness (Yin et al. 2023; Duricic et al. 2023; Peng et al. 2024). Introducing counterfactual reasoning into diversity modeling provides a principled mechanism for uncovering latent interests that are obscured by item popularity. For instance, as illustrated in Figure 1b, a user may be interested in both a book and a pen due to the high UACR from the book to the pen and the user's identity as a student. Yet the pen, being less popular, may remain unseen.

To address the above limitation, a more principled approach is required—one that simulates counterfactual scenarios to reveal how shifts in item exposure may influence user behavior. Such modeling enables the identification of underrepresented yet relevant items, thereby enhancing diversity. Within this framework, we theoretically establish that **item embedding norms correlate with popularity under the classical recommendation framework**. This insight motivates the use of counterfactual interventions to mitigate exposure bias and uncover latent user interests.

To collectively model UACR and promote diversity, we propose **Cadence**: Diversity Recommendation via **Ca**usal **De**confounding of Co-purchase Relations a**n**d **C**ounterfactual **E**xposure, which is a unified GNN-based framework that integrates causal inference with counterfactual augmentation, and features **plug-and-play** transferability for GNN-based accuracy-oriented recommendation architectures. **First**, we introduce a *UACR-Guided Causal Inference and Graph Refinement (CIGR)* module. It estimates UACR between items based on historical user interactions and constructs a deconfounded intrinsic item dependency graph to capture more accurate semantic associations. A structure-aware item-to-item aggregation mechanism is then applied to refine item representations. **Second**, we present a *UACR-Guided Candidate Selection and Counterfactual Exposure (CSCE)*module. This module combines UACR with a two-stage candidate selection strategy. It incorporates counterfactual interventions to simulate the effects of item popularity shifts on user behavior, thereby improving recommendation coverage and diversity.

The main contributions are summarized as follows:

- To the best of our knowledge, this study is the **first** to leverage counterfactual inference for modeling directional co-purchase dependencies, explicitly aimed at improving the diversity of downstream recommendations.
- We propose a ***plug-and-play*** framework that constructs a deconfounded causal item graph via UACR to enhance item representations, and incorporates a two-stage selection strategy with counterfactual intervention to simulate exposure variations, thereby improving diversity while preserving relevance.
- Extensive experiments on multiple real-world datasets show that our method surpasses state-of-the-art models in both accuracy and diversity, while offering superior transferability and efficiency over baselines.

Methodology

Overview

As illustrated in Figure 2, the proposed model comprises two principal UACR-Guided modules: *Causal Inference and Graph Refinement (CIGR)* and *Candidate Selection and Counterfactual Exposure (CSCE)*. The first module estimates UACR through counterfactual reasoning and propensity-based causal inference, followed by graph refinement to retain only high-confidence item-to-item links. This process enhances the directional accuracy and structural integrity of the item representation space. The second module leverages these causal signals to construct personalized, category-aware candidate sets and further enriches recommendation diversity while preserving relevance by simulating high-exposure conditions through norm-based counterfactual adjustments.

Causal Inference and Graph Refinement

UACR Learning. Conventional approaches model item dependencies via co-occurrence statistics (Zhou et al. 2025; Chou and Cheng 2023; Yan et al. 2022). While effective in capturing associative patterns, these statistical methods are vulnerable to confounding influences such as item popularity and user attributes, often resulting in biased dependency structures. To address this limitation, we introduce a causal inference framework that estimates Unbiased Asymmetric Co-purchase Relationships (UACR) through counterfactual reasoning under the potential outcomes model. This allows for the recovery of directional item dependencies grounded in causal rather than correlational signals.

To model causal dependencies between items, we begin by constructing a bipartite user–item interaction graph $\mathcal{G} = (\mathcal{U} \cup \mathcal{I}, \mathcal{E})$, where $\mathcal{U}$ and $\mathcal{I}$ denote the sets of users and items, respectively, and edges $(u, i) \in \mathcal{E}$ indicate observed interactions. For each user $u \in \mathcal{U}$ and item $i \in \mathcal{I}$, we define a binary treatment variable $T_i(u)$ to indicate whether user u has interacted with item i:

$$T_i(u) = \mathbb{I}((u, i) \in \mathcal{E}), \tag{1}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. Given a treatment item i and a target item j, we then define the potential out-

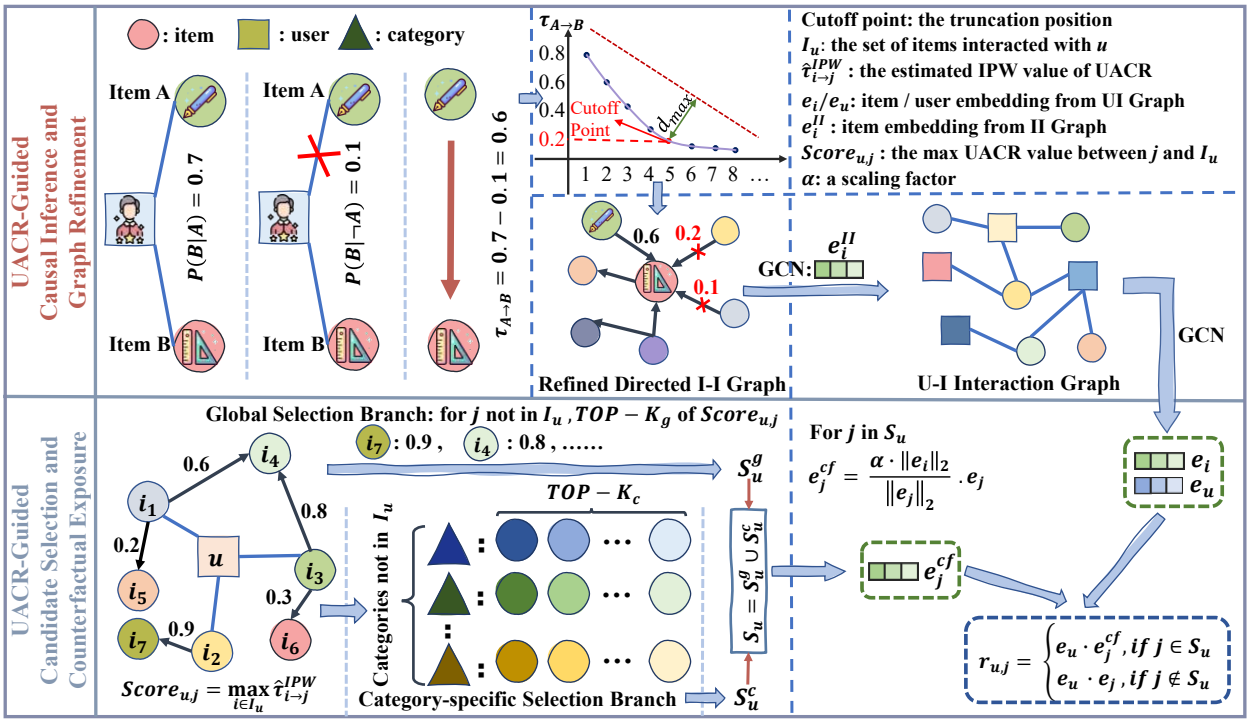


Figure 2: Overall architecture of Cadence. It consists of two UACR-Guided modules: *Causal Inference and Graph Refinement*, which constructs deconfounded item graphs and refines representations; and *Candidate Selection and Counterfactual Exposure*, which simulates popularity interventions to enhance diversity while preserving relevance.

comes for user u as:

$$Y_j^{(i)}(u,1) \equiv Y_j(u \mid T_i(u)=1),$$
$$Y_j^{(i)}(u,0) \equiv Y_j(u \mid T_i(u)=0), \tag{2}$$

where $Y_j^{(i)}(u,1)$ and $Y_j^{(i)}(u,0)$ denote the likelihood of interacting with item j under exposure and non-exposure to item i. The observed behaviour corresponds to the realized potential outcome:

$$Y_j^{(i)}(u) = T_i(u)\cdot Y_j^{(i)}(u,1) + (1-T_i(u))\cdot Y_j^{(i)}(u,0). \tag{3}$$

To quantify the causal influence from item i to item j, we define the average treatment effect (ATE) over the user population:

$$\tau_{i\to j} = \mathbb{E}_u\left[Y_j^{(i)}(u,1) - Y_j^{(i)}(u,0)\right]. \tag{4}$$

As both potential outcomes cannot be observed for the same user, we approximate the ATE through stratified sampling. Users are partitioned into a treatment group $\mathcal{U}_1 = \{u \mid T_i(u) = 1\}$ and a control group $\mathcal{U}_0 = \{u \mid T_i(u) = 0\}$, yielding the empirical estimate:

$$\tau_{i\to j} \approx \mathbb{E}_{u\in\mathcal{U}_1}[Y_j(u)] - \mathbb{E}_{u\in\mathcal{U}_0}[Y_j(u)]. \tag{5}$$

This group-level contrast approximates the causal effect of item i on engagement with item j, while implicitly correcting for item popularity through the control group's baseline preference.

While the treatment–control stratification offers a baseline for causal effect estimation, it implicitly assumes user homogeneity by assigning equal weight to all samples. In reality, users vary in preferences, demographics and behavioural traits, which may affect their exposure to item i and propensity to engage with item j. These confounding factors can lead to biased attribution of causal influence.

To address this challenge, we adopt a propensity score reweighting strategy to adjust for observed user-level confounders. By assigning lower weights to interactions driven by user-specific bias and emphasizing those indicative of true item-to-item influence, this approach improves both the accuracy and interpretability of the causal estimates.

To estimate propensity scores, we encode users and items into fixed-dimensional representations using parameterized embedding functions. For each user u and item i, we define their embedding vectors $\mathbf{x}_u$ and $\mathbf{v}_i$ as:

$$\mathbf{x}_u = f_\theta^{(u)}(\Phi_u), \quad \mathbf{v}_i = f_\theta^{(i)}(F_i), \tag{6}$$

where $f_\theta^{(u)}$ and $f_\theta^{(i)}$ are parameterized encoder functions for users and items, Φ_u denotes the feature set of user u, including both behavioural signals and static attributes, and F_i represents the feature set of item i. While the encoder functions may adopt expressive architectures such as GNNs or pre-trained language models, we employ a minimal configuration with fixed, randomly initialized embeddings $\mathbf{x}_u$ and $\mathbf{v}_i$ to validate the robustness and efficiency of our causal framework independent of representation learning. With $\mathbf{x}_u$

and $\mathbf{v}_i$, we estimate the propensity score using a multi-layer perceptron (MLP) with a sigmoid activation $\sigma(\cdot)$:

$$e_{u,i} = \hat{P}(T_i(u) = 1 \mid \mathbf{x}_u, \mathbf{v}_i) = \sigma\big(\text{MLP}(\mathbf{x}_u, \mathbf{v}_i)\big), \quad (7)$$

where $e_{u,i}$ represents the propensity score, the estimated likelihood that user u interacts with item i given their latent features $\mathbf{x}_u, \mathbf{v}_i$. This score is subsequently used to reweight user samples, mitigating bias in causal effect estimation.

To obtain an unbiased estimate of causal influence, we adopt inverse probability weighting (IPW) and define a Horvitz–Thompson-style estimator (Horvitz and Thompson 1952; Rosenbaum and Rubin 1983) for the treatment effect from item i to item j, which we refer to as the UACR score:

$$\hat{\tau}^{\text{IPW}}_{i\to j} = \underbrace{\frac{\sum_{u\in\mathcal{U}} \omega_u \cdot T_i(u) \cdot Y_j^{(i)}(u,1)}{\sum_{u\in\mathcal{U}} \omega_u \cdot T_i(u)}}_{\text{Treated group}} - \underbrace{\frac{\sum_{u\in\mathcal{U}} \omega_u \cdot (1-T_i(u)) \cdot Y_j^{(i)}(u,0)}{\sum_{u\in\mathcal{U}} \omega_u \cdot (1-T_i(u))}}_{\text{Control group}}, \quad (8)$$

where each user is assigned a sample-specific weight:

$$\omega_u = \frac{T_i(u)}{e_{u,i}} + \frac{1-T_i(u)}{1-e_{u,i}}. \quad (9)$$

This reweighting corrects for covariate imbalance between treatment and control groups, ensuring that the estimated effect captures causal rather than spurious associations driven by user heterogeneity or item popularity. Under standard assumptions, the IPW estimator is asymptotically unbiased and converges in probability to the true ATE:

$$\hat{\tau}^{\text{IPW}}_{i\to j} \xrightarrow{p} \tau_{i\to j}. \quad (10)$$

Accordingly, $\hat{\tau}^{\text{IPW}}_{i\to j}$ can be interpreted as the net causal effect from item i to item j, providing a principled means to isolate item-to-item influence while accounting for confounding factors such as user attributes, item popularity, and random behavioural noise.

UACR-Guided Item Graph Refinement. Following the estimation of UACR scores $\hat{\tau}^{\text{IPW}}_{j\to i}$, we refine the item graph by discarding edges with weak or negative causal influence. For each target item i, let $\mathcal{I}_{\text{co}}$ denote the set of items with observed co-purchase interactions involving i. For each item $j \in \mathcal{I}_{\text{co}}$, we compute the UACR score $\hat{\tau}^{\text{IPW}}_{j\to i}$ and sort the items in descending order by their scores, yielding a ranked sequence of edge weights $\{w_1, w_2, \ldots, w_n\}$.

To determine a principled truncation point, we adopt the geometric farthest-point truncation rule, a variant of the elbow method (Hastie et al. 2009). Specifically, the ordered UACR scores are represented as points (k, w_k) in two-dimensional space. A straight line is drawn connecting the endpoints $(1, w_1)$ and (n, w_n), and the perpendicular distance from each intermediate point to this line is computed. The index k^* corresponding to the maximum deviation identifies the optimal truncation position. The top-k^* incoming items are retained as the pruned neighbor set $\mathcal{N}_i^{\text{in}} \subseteq \{j \mid j \to i\}$. The selected UACR scores are normalized as:

$$\sum_{j\in\mathcal{N}_i^{\text{in}}} \tilde{\tau}^{\text{IPW}}_{j\to i} = 1, \quad \text{where} \quad \tilde{\tau}^{\text{IPW}}_{j\to i} = \frac{\hat{\tau}^{\text{IPW}}_{j\to i}}{\sum_{\ell\in\mathcal{N}_i^{\text{in}}} \hat{\tau}^{\text{IPW}}_{\ell\to i}}. \quad (11)$$

This truncation strategy ensures that the item graph retains only high-confidence causal links, enhancing its structural integrity and interpretability for representation learning.

We adopt LightGCN (He et al. 2020) as the backbone and incorporate a UACR-guided item graph refinement module, resulting in a two-stage embedding learning pipeline. In the first stage, we introduce a UACR-driven item-item aggregation layer, where each item node i aggregates signals from its upstream neighbours $\mathcal{N}_i^{\text{in}}$ via multi-layer graph convolutions. The aggregation is weighted by the normalized causal strengths $\tilde{\tau}^{\text{IPW}}_{j\to i}$ derived from UACR estimation:

$$e_i^{(l+1)} = \frac{\sum_{j\in\mathcal{N}_i^{\text{in}}} \tilde{\tau}^{\text{IPW}}_{j\to i}}{\sqrt{|\mathcal{N}_i^{\text{in}}||\mathcal{N}_j^{\text{out}}|}}\, e_j^{(l)}, \quad e_i^{\text{II}} = \frac{1}{L_{\text{II}}+1}\sum_{l=0}^{L_{\text{II}}} e_i^{(l)}, \quad (12)$$

where $e_i^{(l)}$ denotes the representation of item i at layer l. The formulation combines degree-based normalization with causal edge reweighting to emphasize structurally and semantically relevant neighbors. The final item representation e_i^{II} is computed by averaging representations across $L_{\text{II}}+1$ propagation layers. This design enhances the semantic fidelity of learned item representations by prioritizing causally influential interactions.

Subsequently, we feed the refined item embeddings e_i^{II} into a standard LightGCN framework (He et al. 2020) to perform convolution over the user-item bipartite graph:

$$e_u^{(l+1)} = \sum_{i\in\mathcal{N}_u} \frac{e_i^{\text{II},(l)}}{\sqrt{|\mathcal{N}_u||\mathcal{N}_i|}}, e_i^{(l+1)} = \sum_{u\in\mathcal{N}_i} \frac{e_u^{(l)}}{\sqrt{|\mathcal{N}_i||\mathcal{N}_u|}}, \quad (13)$$

where $\mathcal{N}_u$ and $\mathcal{N}_i$ denote the neighbors of user u and item i in the user-item interaction graph, respectively. The embeddings $e_i^{\text{II},(l)}$ are initialized using the causal refinement output from the previous stage, enabling user embeddings $e_u^{(l)}$ to be informed by structurally and causally relevant item representations. Final representations are obtained by average pooling of embeddings from all layers:

$$e_u = \frac{1}{L+1}\sum_{l=0}^{L} e_u^{(l)}, \quad e_i = \frac{1}{L+1}\sum_{l=0}^{L} e_i^{(l)}. \quad (14)$$

This design preserves the simplicity and efficiency of LightGCN, while significantly enhancing the quality and interpretability of item embeddings through the integration of UACR correction signals. The resulting user and item embeddings are thus optimized for both collaborative structure and causal coherence.

Candidate Selection and Counterfactual Exposure

UACR-Guided Two-Stage Candidate Selection. To enhance diversity while preserving relevance, we introduce a two-stage candidate selection strategy guided by UACR.

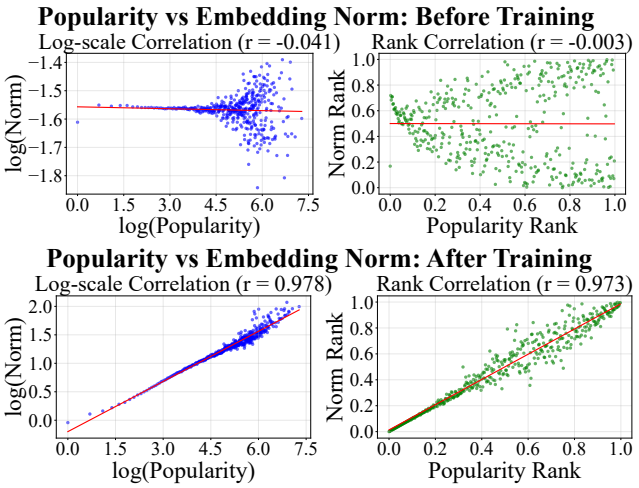


Figure 3: Embedding norm vs. popularity on the TaoBao dataset. Top: before training; bottom: after training. Left: log scale; right: rank correlation.

The proposed framework comprises a global selection branch and a category-specific selection branch in parallel.

For each user u and an unseen item j, we define a UACR-based interaction score as the maximum estimated causal effect from the user's historical items:

$$\text{score}_{u,j} = \max_{i \in I_u} \hat{\tau}^{\text{IPW}}_{i \to j}, \quad I_u = \{i : (u,i) \in \mathcal{E}\}, \tag{15}$$

where I_u denotes the interaction history of user u.

In the *global selection* branch, all items not previously interacted with are ranked in descending order of $\text{score}_{u,j}$, and the top K_g items are globally relevant candidates:

$$S_u^g = \text{Top-}K_g\left(\{j : j \notin I_u\} \text{ ranked by score}_{u,j}\right). \tag{16}$$

In parallel, the *category-specific selection* branch partitions the candidate pool by category. For each category c, the top K_c items, ranked by $\text{score}_{u,j}$, are selected from the set of unseen items:

$$S_u^c = \bigcup_c \text{Top-}K_c\left(\{j \in I^c : j \notin I_u\} \text{ ranked by score}_{u,j}\right), \tag{17}$$

where I^c denotes the set of items belonging to category c.

The final candidate set S_u is formed by unifying the outputs of both selection branches:

$$S_u = S_u^g \cup S_u^c. \tag{18}$$

This two-stage strategy captures items with high causal relevance while ensuring coverage across categories. The resulting set provides a semantically rich and diverse input for downstream ranking and counterfactual enhancement.

Counterfactual Norm-Based Enhancement. We empirically observe a positive correlation between the norm (i.e., Euclidean magnitude) of item embeddings and their popularity, which we formally analyze in the *next subsection* titled "Theoretical Insight: Norm-Popularity Correlation Proof." Although candidate items selected via UACR exhibit strong causal relevance to user preferences, many of them remain under-exposed in historical data, which may obscure their true utility.

To address this, we pose a counterfactual hypothesis: *if these candidate items had received exposure levels comparable to or even higher than items previously interacted with by the user, would the user have engaged with them?* Motivated by this, we introduce a norm-based enhancement strategy to simulate such high-exposure conditions.

Concretely, for a candidate item j and a reference item i from user interaction, we scale the embedding of j to match the norm of i, amplified by a scalar factor $\alpha \geq 1$:

$$\mathbf{e}_j^{cf} \leftarrow \frac{\alpha \cdot \|\mathbf{e}_i\|_2}{\|\mathbf{e}_j\|_2} \cdot \mathbf{e}_j, \quad \alpha \geq 1, \tag{19}$$

where $\|\cdot\|_2$ denotes the L2 norm (for vectors) or spectral norm (for matrices), and $\mathbf{e}_j^{cf}$ is the counterfactually adjusted embedding of item j.

The final user–item relevance score is then computed as:

$$r_{u,j} = \begin{cases} \mathbf{e}_u \cdot \mathbf{e}_j^{cf}, & \text{if } j \in S_u, \\ \mathbf{e}_u \cdot \mathbf{e}_j, & \text{otherwise.} \end{cases} \tag{20}$$

This norm-based counterfactual adjustment enables the ranking model to simulate user behaviour under hypothetical popularity scenarios, thus enriching the recommendation signal with causal insights beyond historical exposure bias.

Theoretical Insight: Norm-Popularity Correlation Proof. In graph-based recommender systems, we theoretically establish that item embedding norms correlate positively with popularity under the LightGCN + BPR framework.

We prove that when items are sampled proportionally to their popularity during training, the expected norm update satisfies:

$$\mathbb{E}[\Delta\|\mathbf{e}_i^{(0)}\|_2] = \eta \cdot \tilde{\kappa}_i \cdot \frac{n_i}{|D|} - \eta\lambda\mathbb{E}[\|\mathbf{e}_i^{(0)}\|_2], \tag{21}$$

where n_i is the interaction count of item i, $|D|$ is total interactions, and $\tilde{\kappa}_i$ captures alignment effects. Using the Azuma–Hoeffding inequality, we show that if popularity difference satisfies $n_i - n_j \gtrsim \mathcal{O}(\sqrt{T})$, then after T training steps, $\|\mathbf{e}_i\|_2 > \|\mathbf{e}_j\|_2$ holds with exponentially high probability. This establishes that **embedding norms grow approximately linearly with popularity under standard training.**

Empirically, on the TaoBao dataset (Fig. 3), norms and popularity show strong linear correlation after convergence (Pearson $r = 0.978$, rank correlation $r = 0.973$), validating our theoretical analysis. Due to space constraints, the complete theoretical verification (including detailed derivations, hyperparameter tuning, and complexity analysis) is provided in the full paper on arXiv (Zhang et al. 2025b).

Experiments

Experimental Settings

Datasets. We evaluate our model on three real-world datasets: Beauty, TaoBao, and Toy. Among them, the Beauty

Dataset	Metric	EDUA	DGCN	DGRec	CPGRec	KG-Diverse	DivGCL	Ours	%Imp
Beauty	R@100	0.2364	0.2395	0.2399	0.2014	0.2104	0.2280	**0.2590**	7.96%
	R@300	0.3585	0.3790	0.3915	0.3293	0.3629	0.3320	**0.4078**	4.16%
	HR@100	0.2373	0.3418	0.3420	0.3763	0.3240	0.2964	**0.4396**	16.8%
	HR@300	0.3562	0.4792	0.5021	0.5318	0.4795	0.4182	**0.5687**	6.94%
	C@100	20.1702	18.2876	19.0557	20.3565	19.4270	18.9745	**23.9280**	17.54%
	C@300	28.3444	26.9694	27.5704	28.2383	26.3016	28.0006	**29.8727**	5.39%
	E@100	3.5733	3.4813	3.4222	3.5078	3.5918	3.3427	**4.0048**	11.50%
	E@300	3.8880	3.7103	3.7369	3.7625	3.7442	3.7408	**4.0122**	3.19%
TaoBao	R@100	0.0301	0.0394	0.0472	0.0404	0.0336	0.0426	**0.0699**	48.09%
	R@300	0.0609	0.0831	0.0951	0.0792	0.0686	0.0750	**0.1119**	17.67%
	HR@100	0.0309	0.2634	0.3026	0.2871	0.2110	0.0394	**0.3727**	23.17%
	HR@300	0.0624	0.4369	0.4817	0.4454	0.3652	0.0727	**0.5103**	5.95%
	C@100	36.7610	38.1183	39.0597	36.1320	39.1295	36.3410	**52.1118**	33.18%
	C@300	82.6850	84.4989	89.1684	81.3092	86.8845	85.8789	**176.2450**	97.65%
	E@100	4.1283	4.3700	4.1307	4.1422	4.4906	4.2672	**4.6401**	3.33%
	E@300	4.9726	4.9730	4.8560	4.9419	4.9761	4.9631	**6.3578**	27.77%
Toy	R@100	0.0373	0.0348	0.0534	0.0634	0.0634	0.0813	**0.1183**	45.51%
	R@300	0.0737	0.0826	0.0987	0.1045	0.1243	0.1296	**0.1919**	48.07%
	HR@100	0.0398	0.0668	0.0987	0.1138	0.1148	0.0782	**0.2052**	78.75%
	HR@300	0.0774	0.1518	0.1762	0.1823	0.2153	0.1288	**0.3121**	44.96%
	C@100	31.9189	42.5791	47.0231	41.4623	39.4143	45.8592	**64.7435**	37.68%
	C@300	72.0048	76.2601	85.9381	78.1188	76.7754	84.3747	**102.2771**	19.01%
	E@100	3.8175	4.8195	4.9407	4.5742	4.3662	4.9633	**5.6773**	14.39%
	E@300	4.6669	5.2774	5.5712	5.1874	4.9913	5.5659	**5.6915**	2.16%

Table 1: Performance comparison of different methods across three datasets (Beauty, TaoBao, Toy).

and TaoBao datasets have been widely adopted in previous studies, such as DGCN (Zheng et al. 2021) and DGRec (Yang et al. 2023). The Toy dataset is derived from the Amazon Review dataset[1] and undergoes the following preprocessing steps: (1) interactions involving items without category information are removed; (2) a 5-core filter is applied to retain only users and items with at least five interactions. Table 2 presents detailed statistics of all datasets.

Dataset	Beauty	TaoBao	Toy
Users	8,159	82,633	190,217
Items	5,862	136,710	73,130
Interactions	98,566	4,230,631	1,657,410
Density	0.2061%	0.0375%	0.1193%
Categories	41	3,108	193
Avg. Size	139.595	43.986	378.912

Table 2: Dataset Statistics Comparison.

Baselines and Evaluation Metrics. We conduct comparative experiments with six representative diversity-aware recommendation models, including EDUA (Liang et al. 2021), DGCN (Zheng et al. 2021), DGRec (Yang et al. 2023), CPGRec (Li et al. 2024), KG-Diverse (Liu et al. 2024), and DivGCL (Gong et al. 2025). For evaluation, we adopt metrics that assess both accuracy and diversity aspects. Specifically, Recall (R@N) and Hit Ratio (HR@N) are used to measure the accuracy of results, while Coverage (C@N) and Entropy (E@N) evaluate the diversity of the results,where entropy is calculated with base-2 logarithm. Here, $N \in \{100, 300\}$ (Zheng et al. 2021; Yang et al. 2023) is the number of items recommended to each user.

[1]https://cseweb.ucsd.edu/~jmcauley/datasets/amazon_v2/

Overall Performance Comparison

Table 1 reports the results, with the best in bold, second-best underlined, and %Imp indicating the improvement over the second-best. Our key findings are:

Diversity superiority. In terms of diversity, our method consistently outperforms all baselines on diversity metrics. For example, on C@100, compared to the second-best results, we achieve improvements of 17.54%, 33.18%, and 37.68% on Beauty, TaoBao, and Toy datasets, respectively. These results show the effectiveness of our counterfactual exposure strategy in discovering diverse candidate items, especially under large-scale sparse data conditions.

Accuracy superiority. In terms of accuracy, our method consistently achieves the best results on accuracy metrics across all three datasets. For example, on R@100, compared to the second-best results, we achieve improvements of 7.96%, 48.09%, and 45.51% on Beauty, TaoBao, and Toy datasets, respectively. These gains highlight the model's strong capability in leveraging UACR to disentangle spurious correlations and preserve true causal item dependencies.

Generalization superiority. Our method shows consistently strong performance across three datasets with diverse characteristics. These datasets vary significantly in scale, density, and number of categories. The consistent improvements across all of them highlight the strong generalization ability of our approach, validating its robustness and adaptability to different recommendation scenarios.

Ablation Study

As shown in Table 3, we investigate the impact of two key UACR-Guided modules in our model: Causal Inference and Graph Refinement (CIGR) and Candidate Selection and Counterfactual Exposure (CSCE). *w/o CIGR* and *w/o CSCE*

Dataset	Beauty				TaoBao				Toy			
	R@100	HR@100	C@100	E@100	R@100	HR@100	C@100	E@100	R@100	HR@100	C@100	E@100
Ours	**0.2590**	**0.4396**	**23.9280**	**4.0048**	**0.0699**	**0.3727**	**52.1118**	**4.6401**	**0.1183**	**0.2052**	**64.7435**	**5.6773**
w/o CIGR	0.2572	0.4350	23.4661	3.9142	0.0683	0.3670	49.2669	4.5821	0.1162	0.2025	64.2619	5.6628
w/o CSCE	0.2587	0.4369	16.8167	3.0879	0.0526	0.3169	33.5962	4.0163	0.1177	0.2005	35.4518	4.1258
w/o Both	0.2471	0.4256	16.5391	3.0475	0.0505	0.3030	34.5216	4.0068	0.1074	0.1853	36.8562	4.2382

Table 3: Ablation study on three datasets.

Dataset	Beauty		TaoBao		Toy	
	Recall @100	Coverage @100	Recall @100	Coverage @100	Recall @100	Coverage @100
DimCL	0.2554	15.4732	0.0539	32.7501	0.1181	33.6089
+Ours	0.2678	20.3906	0.0665	53.8833	0.1436	60.3681
Impro.	+4.86%	+31.78%	+23.38%	+64.53%	+21.59%	+79.62%
DCF	0.2427	16.0417	0.0464	34.5393	0.1027	35.5815
+Ours	0.2534	22.1235	0.0612	56.7396	0.1336	56.7258
Impro.	+4.41%	+37.91%	+31.90%	+64.28%	+30.10%	+59.42%
PPAC	0.2560	15.0843	0.0485	33.9561	0.1122	33.6564
+Ours	0.2629	22.1567	0.0641	55.3692	0.1457	53.2037
Impro.	+2.70%	+46.89%	+32.16%	+63.06%	+29.86%	+58.08%
NCL	0.2535	14.9884	0.0524	33.9166	0.0986	38.4338
+Ours	0.2627	24.1063	0.0680	51.7184	0.1284	74.0611
Impro.	+3.63%	+60.83%	+29.77%	+52.49%	+30.22%	+92.70%

Table 4: Transferability analysis on three datasets.

denote the removal of the *CIGR* and *CSCE* module respectively, and *w/o Both* indicates the removal of both modules (i.e., retaining only the base LightGCN).

Effect of w/o CIGR. Removing CIGR leads to consistent declines in accuracy and diversity across all datasets—for instance, on TaoBao, R@100 drops from 0.0699 to 0.0683 and C@100 from 52.1118 to 49.2669, with similar patterns on Beauty and Toy. These results indicate that: (1) UACR captures true item associations, and CIGR refines the graph accordingly to enhance user and item embedding quality, thus improving accuracy. (2) CIGR offers high-quality embeddings essential for counterfactual exposure; only semantically meaningful and stable embeddings allow norm scaling to simulate exposure effectively, while poor ones may amplify noise and harm diversity and accuracy.

Effect of w/o CSCE. Removing the CSCE module results in performance drops in both accuracy and diversity, with diversity most affected—for instance, on the Toy dataset, C@100 and E@100 drop sharply from 64.7435 to 35.4518 and from 5.6773 to 4.1258. Similar trends appear on Beauty and TaoBao. This indicates that: (1) CSCE uses counterfactual exposure to uncover underexposed category-level interests, significantly boosting diversity; (2) CSCE uses UACR and user interaction history to identify causally relevant candidates, enhancing diversity while preserving relevance.

Transferability Analysis

Given its **plug-and-play nature**, we integrate our model into four representative accuracy-oriented recommendation frameworks: DimCL (Zhang et al. 2025a), DCF (He et al. 2024), PPAC (Ning et al. 2024), and NCL (Lin et al. 2022), covering various research paradigms such as debiasing, denoising, causal modeling, and contrastive learning.

Table 4 presents the performance comparison before and after applying our method (+Ours). These results indicate that (1) our method demonstrates strong transferability and general applicability, delivering notable gains in coverage that reflect its effectiveness in preserving accuracy while substantially enhancing diversity in accuracy-oriented recommendation models. (2) Its effectiveness is closely tied to the scale and density of the dataset. On the smaller and denser Beauty dataset, UACR estimation is less reliable and the original collaborative signals are relatively complete, which limits performance gains. In contrast, the larger and sparser TaoBao and Toy datasets exhibit weaker collaborative signals, where integrating our method yields more substantial improvements—particularly in settings that closely resemble real-world recommendation scenarios.

Conclusion

In this study, we address two core challenges in recommendation systems: item relationship modeling being susceptible to popularity bias and user attribute confounding, and diversity recommendation methods lacking causal perspectives. We propose Cadence, a plug-and-play framework built upon LightGCN backbone. The framework addresses these issues through CIGR that removes confounding factors and refines embeddings, and CSCE that discovers users' potential diverse interests while preserving relevance. Experiments demonstrate that Cadence outperforms state-of-the-art methods in both diversity and accuracy metrics.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (Project No. 62202122 and 62073272), the Guangdong Basic and Applied Basic Research Foundation under Grant No. 2024A1515011949, the Shenzhen Science and Technology Program under Grant No. GXWD20231130110308001, JCYJ20250604145617023, JCYJ20240813104843058, and JCYJ20240813104837050, the Shenzhen Education Science "14th Five-Year Plan" 2023 Annual Project on Artificial Intelligence Special Project under Grant No. rgzn23001, the Guangdong Province Higher Education Research and Reform Project under Grant No. YueJiaoGaoHan(2024) No.9 (1227), and the Guangdong Province General Colleges and Universities Innovation Team Project under No. 2022KCXTD038.

References

Baklanov, M. 2024. CEERS: Counterfactual Evaluations of Explanations in Recommender Systems. In *Proceedings of the 18th ACM Conference on Recommender Systems*, 1323–1329.

Chou, Y. H.; and Cheng, P. J. 2023. Incorporating co-purchase correlation for next-basket recommendation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 3823–3827.

Du, Y.; Wang, Z.; Sun, Z.; Ma, Y.; Liu, H.; and Zhang, J. 2024. Disentangled Multi-interest Representation Learning for Sequential Recommendation. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 677–688.

Duricic, T.; Kowald, D.; Lacic, E.; and Lex, E. 2023. Beyond-accuracy: a review on diversity, serendipity, and fairness in recommender systems based on graph neural networks. *Frontiers in big data*, 6: 1251072.

Gong, W.; Geng, Y.; Zhang, D.; Zhu, Y.; Xu, X.; Xiang, H.; Beheshti, A.; Zhang, X.; and Qi, L. 2025. DivGCL: A Graph Contrastive Learning Model for Diverse Recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 16853–16861.

Hastie, T.; Tibshirani, R.; Friedman, J. H.; and Friedman, J. H. 2009. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

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.

He, Z.; Wang, Y.; Yang, Y.; Sun, P.; Wu, L.; Bai, H.; Gong, J.; Hong, R.; and Zhang, M. 2024. Double correction framework for denoising recommendation. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 1062–1072.

Horvitz, D. G.; and Thompson, D. J. 1952. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260): 663–685.

Huang, Y.; Liang, K.; Huang, Y.; Zeng, X.; Chen, K.; and Zhou, B. 2025. Social Recommendation via Graph-Level Counterfactual Augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 334–342.

Li, X.; Ma, J.; Liu, K.; Feng, S.; Zhang, H.; and Wang, Y. 2024. Category-based and Popularity-guided Video Game Recommendation: A Balance-oriented Framework. In *Proceedings of the ACM Web Conference 2024*, 3734–3744.

Liang, Y.; Qian, T.; Li, Q.; and Yin, H. 2021. Enhancing domain-level and user-level adaptivity in diversified recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 747–756.

Lin, Z.; Tian, C.; Hou, Y.; and Zhao, W. X. 2022. Improving graph collaborative filtering with neighborhood-enriched contrastive learning. In *Proceedings of the ACM web conference 2022*, 2320–2329.

Liu, G.; Yang, F.; Jiao, Y. A.; Garakani, A. B.; Tong, T.; Gao, Y.; and Jiang, M. 2025. Learning attribute as explicit relation for sequential recommendation.

Liu, X.; Yang, L.; Liu, Z.; Yang, M.; Wang, C.; Peng, H.; and Yu, P. S. 2024. Knowledge graph context-enhanced diversified recommendation. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, 462–471.

Liu, Y.; Yen, J.-N.; Yuan, B.; Shi, R.; Yan, P.; and Lin, C.-J. 2022. Practical counterfactual policy learning for top-k recommendations. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 1141–1151.

Luo, H.; Wu, Y.; Chen, Y.; Zhuang, F.; and Wang, D. 2025. CDC: Causal Domain Clustering for Multi-Domain Recommendation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1840–1849.

Ning, W.; Cheng, R.; Yan, X.; Kao, B.; Huo, N.; Haldar, N. A. H.; and Tang, B. 2024. Debiasing recommendation with personal popularity. In *Proceedings of the ACM Web Conference 2024*, 3400–3409.

Pearl, J. 2009. *Causality*. Cambridge university press.

Peng, K.; Raghavan, M.; Pierson, E.; Kleinberg, J.; and Garg, N. 2024. Reconciling the accuracy-diversity tradeoff in recommendations. In *Proceedings of the ACM Web Conference 2024*, 1318–1329.

Peters, J.; Janzing, D.; and Schölkopf, B. 2017. *Elements of causal inference: foundations and learning algorithms*. The MIT press.

Rosenbaum, P. R.; and Rubin, D. B. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1): 41–55.

Sugahara, K.; Yamasaki, C.; and Okamoto, K. 2024. Is it really complementary? revisiting behavior-based labels for complementary recommendation. In *Proceedings of the 18th ACM Conference on Recommender Systems*, 1091–1095.

Tan, J.; Xu, S.; Ge, Y.; Li, Y.; Chen, X.; and Zhang, Y. 2021. Counterfactual explainable recommendation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 1784–1793.

Tang, S.; Lin, S.; Ma, J.; and Zhang, X. 2025. CoDeR: Counterfactual Demand Reasoning for Sequential Recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 12649–12657.

Wang, X.; Li, Q.; Yu, D.; Li, Q.; and Xu, G. 2024. Counterfactual explanation for fairness in recommendation. *ACM Transactions on Information Systems*, 42(4): 1–30.

Wang, X.; Zhou, K.; Tang, X.; Zhao, W. X.; Pan, F.; Cao, Z.; and Wen, J.-R. 2023. Improving conversational recommendation systems via counterfactual data simulation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2398–2408.

Wei, T.; Chow, T. W.; and Ma, J. 2024. FPSR+: Toward Robust, Efficient, and Scalable Collaborative Filtering With

Partition-Aware Item Similarity Modeling. *IEEE Transactions on Knowledge and Data Engineering*, 36(12): 8283–8296.

Wei, T.; Ma, J.; and Chow, T. W. 2023a. Collaborative residual metric learning. In *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval*, 1107–1116.

Wei, T.; Ma, J.; and Chow, T. W. 2023b. Fine-tuning partition-aware item similarities for efficient and scalable recommendation. In *Proceedings of the ACM Web Conference 2023*, 823–832.

Wu, W.; Wang, C.; Shen, D.; Qin, C.; Chen, L.; and Xiong, H. 2024. Afdgcf: Adaptive feature de-correlation graph collaborative filtering for recommendations. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, 1242–1252.

Yan, A.; Dong, C.; Gao, Y.; Fu, J.; Zhao, T.; Sun, Y.; and McAuley, J. 2022. Personalized complementary product recommendation. In *Companion Proceedings of the Web Conference 2022*, 146–151.

Yang, L.; Wang, S.; Tao, Y.; Sun, J.; Liu, X.; Yu, P. S.; and Wang, T. 2023. Dgrec: Graph neural network for recommendation with diversified embedding generation. In *Proceedings of the sixteenth ACM international conference on web search and data mining*, 661–669.

Yin, Q.; Fang, H.; Sun, Z.; and Ong, Y.-S. 2023. Understanding diversity in session-based recommendation. *ACM Transactions on Information Systems*, 42(1): 1–34.

Yu, J.; Yin, H.; Xia, X.; Chen, T.; Cui, L.; and Nguyen, Q. V. H. 2022. Are graph augmentations necessary? simple graph contrastive learning for recommendation. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, 1294–1303.

Zhang, C.; Han, Q.; Tan, Q.; Wang, S.; Zhao, X.; and Chen, R. 2025a. DimCL: Dimension-Aware Augmentation in Contrastive Learning for Recommendation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, 1913–1923.

Zhang, J.; Zhao, Z.; Lin, Y.; Ma, J.; Wei, T.; Zhang, H.; and Zhang, X. 2025b. Diversity Recommendation via Causal Deconfounding of Co-purchase Relations and Counterfactual Exposure. *arXiv preprint arXiv:2512.17733*.

Zhang, J.; Zhao, Z.; Lin, Y.; Ma, J.; Wei, T.; Zhang, H.; and Zhang, X. 2025c. Dual-Phase Playtime-guided Recommendation: Interest Intensity Exploration and Multimodal Random Walks. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 6232–6241.

Zhao, Z.; Ren, Z.; Yang, J.; Yan, Z.; Wang, Z.; Yang, L.; Ren, P.; Chen, Z.; de Rijke, M.; and Xin, X. 2025. Improving sequential recommenders through counterfactual augmentation of system exposure. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1508–1518.

Zheng, Y.; Gao, C.; Chen, L.; Jin, D.; and Li, Y. 2021. DGCN: Diversified recommendation with graph convolutional networks. In *Proceedings of the Web Conference 2021*, 401–412.

Zhou, Y.; Wang, Y.; Cui, Q.; Guan, X.; and Cisternas, F. 2025. Basket-enhanced heterogenous hypergraph for price-sensitive next basket recommendation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Zhu, Y.; Ma, J.; Wu, L.; Guo, Q.; Hong, L.; and Li, J. 2023. Path-specific counterfactual fairness for recommender systems. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 3638–3649.