Hierarchical Graph Information Bottleneck for Multi-Behavior Recommendation

Hengyu Zhang*
The Chinese University of Hong Kong
Hong Kong SAR, Hong Kong
hyzhang@se.cuhk.edu.hk

Chunxu Shen*
WeChat, Tencent
Shenzhen, China
lineshen@tencent.com

Xiangguo Sun
The Chinese University of Hong Kong
Hong Kong SAR, Hong Kong
xiangguosun@cuhk.edu.hk

Jie Tan
The Chinese University of Hong Kong
Hong Kong SAR, Hong Kong
jtan@se.cuhk.edu.hk

Yanchao Tan
Fuzhou University
Fuzhou, China
yctan@fzu.edu.cn

Yu Rong
The Chinese University of Hong Kong
Hong Kong SAR, Hong Kong
yu.rong@hotmail.com

Hong Cheng†
The Chinese University of Hong Kong
Hong Kong SAR, Hong Kong
hcheng@se.cuhk.edu.hk

Lingling Yi†
WeChat, Tencent
Shenzhen, China
chrisyi@tencent.com

Abstract

In real-world recommendation scenarios, users typically engage with platforms through multiple types of behavioral interactions. Multi-behavior recommendation algorithms aim to leverage various auxiliary user behaviors to enhance prediction for target behaviors of primary interest (e.g., buy), thereby overcoming performance limitations caused by data sparsity in target behavior records. Current state-of-the-art approaches typically employ hierarchical design following either cascading (e.g., view→cart→buy) or parallel (unified→behavior→specific components) paradigms, to capture behavioral relationships. However, these methods still face two critical challenges: (1) severe distribution disparities across behaviors, and (2) negative transfer effects caused by noise in auxiliary behaviors. In this paper, we propose a novel model-agnostic Hierarchical Graph Information Bottleneck (HGIB) framework for multi-behavior recommendation to effectively address these challenges. Following information bottleneck principles, our framework optimizes the learning of compact yet sufficient representations that preserve essential information for target behavior prediction while eliminating task-irrelevant redundancies. To further mitigate interaction noise, we introduce a Graph Refinement Encoder (GRE) that dynamically prunes redundant edges through learnable edge dropout mechanisms. We conduct comprehensive experiments on three real-world public datasets, which demonstrate the superior effectiveness of our framework. Beyond these widely used datasets in the academic community, we further expand our evaluation on several real industrial scenarios and conduct an online A/B testing, showing again a significant improvement in multi-behavior recommendations. The source code of our proposed HGIB is available at https://github.com/zhy99426/HGIB.

CCS Concepts

• **Information systems** → **Recommender systems**.

Keywords

Multi-Behavior Recommendation, Information Bottleneck, Graph Neural Networks

ACM Reference Format:

Hengyu Zhang, Chunxu Shen, Xiangguo Sun, Jie Tan, Yanchao Tan, Yu Rong, Hong Cheng, and Lingling Yi. 2025. Hierarchical Graph Information Bottleneck for Multi-Behavior Recommendation. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems (RecSys '25), September 22–26, 2025, Prague, Czech Republic.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3705328.3748073

1 Introduction

In real-world online platforms, users typically engage with the platform through multiple types of interaction behaviors such as viewing, adding-to-cart, and purchasing. Therefore, the goal of multi-behavior recommendation tasks is to effectively leverage these multiple interaction behaviors to improve the prediction performance of the model for target behaviors (e.g., buy) of the primary focus of the platform.

To achieve this goal, researchers propose many advanced machine learning methods based on deep neural networks (DNNs) [8, 32], graph convolutional networks (GCNs) [13, 26, 31, 44], and attention mechanisms [5, 32, 40, 45] to precisely model user preferences. Subsequent efforts explore incorporating multi-task learning (MTL) paradigms to utilize multi-behavior supervision signals [1, 2, 33, 46] and enhancing representation learning through self-supervised learning techniques like contrastive learning [30, 35].

Recent state-of-the-art (SOTA) approaches predominantly adopt a hierarchical design philosophy to model the relationships between behaviors, primarily divided into two paradigms: cascading and

*Both authors contributed equally to this research.
†Corresponding author.

RecSys '25, Prague, Czech Republic

ACM ISBN 979-8-4007-1364-4/25/09
https://doi.org/10.1145/3705328.3748073

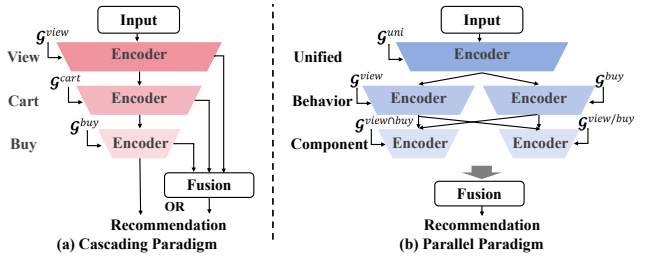


Figure 1: Illustration of the hierarchical model design in cascading and parallel paradigms.

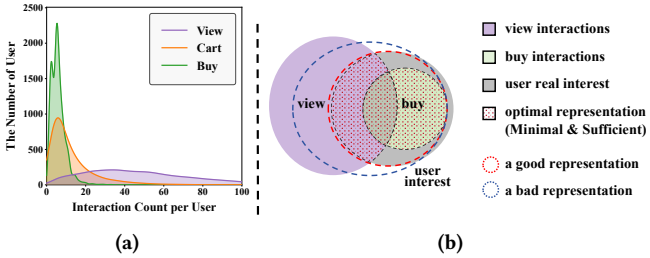


Figure 2: (a) Histogram of user numbers w.r.t interaction counts per user for different behaviors on the Taobao dataset. (b) Information diagram of learned representation, user real interest, and user-item interactions (e.g., view, buy).

parallel paradigms. Methods in the cascading paradigm [3, 18, 26] are inspired by the natural cascading sequence of behaviors, such as view → cart → buy, modeling the relationships hierarchically through cascading encoders, as shown in Figure 1 (a). Since cascading relationships between behaviors are not strict, many other works adopt the parallel paradigm. Methods in the parallel paradigm [15, 22] use parallel encoders for different behaviors and design modules to uncover relationships between them, such as decoupling "view" into "view and buy" and "view but not buy," as depicted in Figure 1 (b).

Despite their success, the existing SOTA approaches still encounter critical challenges:

- **Severe Distribution Disparities Across Behaviors.** As illustrated in Figure 2 (a), there are substantial distributional differences among different behaviors. For instance, in e-commerce platforms, users may browse numerous items in a day but not make a purchase. This imbalance introduces two key issues: On the one hand, auxiliary behaviors (e.g., view) typically exhibit substantially higher interaction frequencies than target behaviors (e.g., buy). So auxiliary behavior signals may introduce popularity bias into target behavior prediction due to their numerical dominance; On the other hand, the long-tail distribution of sparse target behaviors increases model overfitting risks. Figure 2(b) further illustrates this issue: the red circle represents a good, sufficient, and compact user interest representation, while the blue circle is affected by bias from auxiliary behaviors and suffers from overfitting.
- **Negative Transfer from Noisy Auxiliary Behaviors.** Many auxiliary behaviors are implicit feedback and do not directly reflect users' real interests. For example, a user browsing an item does not necessarily indicate a preference for it. Additionally, feedback from accidental clicks is common. These noises in auxiliary behaviors can negatively impact target behavior prediction, leading to a phenomenon known as negative transfer, which degrades model performance.

To overcome these challenges, we offer a simple yet effective framework for multi-behavior recommendation with **H**ierarchical **G**raph **I**nformation **B**ottleneck (**HGIB**). Specifically, our approach integrates information bottleneck principles into hierarchical model design, enforcing the multi-behavior model to retain the relevant information for the target task while minimizing irrelevant information. In particular, we analyze the optimization objective of hierarchical models for multi-behavior recommendation from an information-theoretic perspective, and derive the optimization goal into the loss function by finding a lower bound of the maximized objective and employing HSIC [7] for approximations. Furthermore, to explicitly remove noises from multi-behavior data, we propose the Graph Refinement Encoder (GRE) that prunes noisy interactions, thus further mitigating the negative transfer phenomenon from auxiliary behaviors. The main contributions of our work can be summarized as follows:

- We analyze the optimization objective of hierarchical model design in multi-behavior recommendation from the perspective of the information bottleneck principle and present an innovative Hierarchical Graph Information Bottleneck (HGIB) framework. This framework effectively addresses the challenges of imbalanced multi-behavior data distribution and negative transfer.
- Our proposed Graph Refinement Encoder (GRE) explicitly removes noisy interactions from multi-behavior data, thereby alleviating negative transfer.
- Extensive experiments across three public datasets demonstrate the outstanding superiority and broad generality of our HGIB framework. Moreover, to further demonstrate the industrial application potential of our framework, we conduct comparison experiments on two large-scale industrial datasets with billions of interactions and perform a 15-day online A/B testing.

2 Related Works

2.1 Multi-Behavior Recommendation

The challenge of user interaction sparsity drives significant academic and industrial interest in multi-behavior recommendation systems, which aim to enhance target behavior prediction (e.g., buy) by leveraging auxiliary behavior interactions (e.g., view, cart). Early approaches primarily focus on extending single-behavior recommendation algorithms, exemplified by adopting matrix factorization into CMF [47], the use of Monte Carlo sampling [25], or developing effective negative sampling strategies [4, 14, 19] for multi-behavior data.

The rise of deep learning techniques shifts research focus towards approaches based on neural networks. Researchers devote themselves to using advanced architectures, including deep neural

networks (DNNs), graph convolutional networks (GCNs), and attention mechanisms to extract rich patterns from multi-behavior interactions. For example, DNN-based models like DIPN [8] and MATN [32] employ attention mechanisms to model inter-behavior relationships for final representation aggregation. However, these architectures typically fail to capture high-order user-item interactions in graph structures, leading to GCN-based frameworks as the dominant paradigm. Representative works such as RGCN [26], MBGCN [13], and GNMR [31] construct unified interaction graphs to model global user preferences through graph convolution operations. To better utilize auxiliary behavior signals, the multi-task learning framework is integrated into recommendation systems. Some works like MGNN [46], EHCF [2], GHCF [1], and MBGMN [33] achieve performance improvements through joint prediction of multiple behaviors. Subsequent efforts like CML [30] and MBSSL [35] seek to incorporate self-supervised learning techniques like contrastive learning to further enhance representation learning. More recently, HIRE [5] proposes a lightweight module to directly infer important heterogeneous interactions from data without relying on predefined patterns.

Recent advancements in cascading paradigms, as exemplified by CRGCN [36] and MB-CGCN [3], leverage natural behavioral hierarchies (e.g., view→cart→purchase) to achieve superior prediction accuracy. DA-GCN [50] extends the model of cascading relationships across multiple behaviors by constructing personalized directed acyclic behavior graphs. However, these cascading models face challenges from distributional imbalances across behaviors, where biases and noise in auxiliary behaviors transfer to target behavior prediction, resulting in negative transfer issues. To address this limitation, PKEF [21] introduces a hybrid architecture combining parallel and cascading paradigms for behavior relationship modeling. The superiority of parallel paradigms is also demonstrated in MB-HGCN [37] and HPMR [22]. The state-of-the-art MULE [15] model further demonstrates the superiority of parallel architectures through its unified→behavior→behavior component, achieving superior performance.

2.2 Information Bottleneck for Recommendation

The Information Bottleneck (IB) principle is successfully applied in machine learning to extract optimal intermediate representations that achieve sufficient compression while preserving predictive relevance. This principle has demonstrated significant success across various domains, including image classification, natural language understanding, and graph learning [41]. In recent years, some research studies have tried to adapt the IB framework to recommendation systems, particularly in addressing challenges related to robust representation learning, denoising, debiasing, fairness, and explainability. CGI [29] integrates the IB principle with contrastive learning frameworks through graph augmentation techniques like node and edge dropping to learn robust representations for recommendation tasks. DIB [16, 17] employs the IB framework to alleviate confounding bias in recommendation systems. GBSR [39] and IBMRec [38] leverage IB-based approaches for graph and feature denoising for social and multi-modal recommendation, respectively. FairIB [34] proposes a model-agnostic methodology that effectively balances recommendation accuracy with fairness considerations. These diverse applications underscore the versatility of the IB principle in addressing critical challenges across different recommendation tasks.

3 Preliminaries

3.1 Problem Definition

Given the user set $\mathcal{U} = \{u_1, u_2, \ldots, u_{|\mathcal{U}|}\}$ and item set $\mathcal{V} = \{v_1, v_2, \ldots, v_{|\mathcal{V}|}\}$, we formalize the multi-behavior recommendation task as follows. Let $\mathcal{B}$ denote the behavior set where $b_t \in \mathcal{B}$ represents the target behavior (e.g., buy) which is the primary concern of the platform, and $\mathcal{B}_{aux} = \mathcal{B} \setminus \{b_t\}$ are auxiliary behaviors (e.g., view, add-to-cart). Suppose $|\mathcal{B}|$ is the total number of behavior types, so there are $|\mathcal{B}| - 1$ types of auxiliary behaviors in $\mathcal{B}_{aux}$. For each interaction that user u has interacted with item v under behavior $b \in \mathcal{B}$, a user-item pair (u, v) is included in the edge set $\mathcal{E}^b$, which stores all the user-item interactions under behavior b. So we can treat the user-item interaction data under behavior b as a form of bipartite graph, denoted as $\mathcal{G}^b = (\mathcal{U} \cup \mathcal{V}, \mathcal{E}^b)$. The objective of multi-behavior recommendation is to utilize the data from auxiliary behaviors in $\mathcal{B}_{aux}$ sufficiently to assist in predicting the specific target behavior b_t.

3.2 Information Bottleneck Principle

The Information Bottleneck (IB) principle provides an information-theoretic framework for learning robust representations by discarding information that is not useful from the input. This framework formalizes the trade-off between preserving predictive information about the target task and compressing irrelevant variations in intermediate representation. Formally, given the input data X and the target variable Y, IB aims to find an optimal representation T which is sufficient and compact of input X that maximally preserves information about the label Y while minimizing redundancy, which can be formulated as:

$$\max \underbrace{I(T;Y)}_{\text{Relevance}} - \beta \cdot \underbrace{I(X;T)}_{\text{Compression}}, \tag{1}$$

where $I(T;Y)$ denotes the mutual information between the representation T and the label Y, and $I(X;T)$ means the mutual information between the representation T and the input X. $\beta > 0$ controls the trade-off of the parts of preservation and compression. Recently, the IB principle has become a powerful framework for robust representation learning in pattern recognition, natural language processing, and model explainability.

4 Hierarchical Graph Information Bottleneck (HGIB)

4.1 Overview

In this section, we propose **H**ierarchical **G**raph **I**nformation **B**ottleneck (**HGIB**) framework, to integrate the information bottleneck principle into the hierarchical model design for multi-behavior recommendation. The overall objective of our proposed HGIB is to strategically preserve discriminative information relevant to target behavior prediction while eliminating irrelevant noise.

4.1.1 Abstraction of Hierarchical Model. To capture the complex interaction patterns between users and items, we map the user/item IDs in the given user set $\mathcal{U}$ and item set $\mathcal{V}$ to a learnable dense embedding, which can be formulated as:

$$\boldsymbol{e}_u = E^T \cdot \boldsymbol{o}_u,\ \boldsymbol{e}_v = E^T \cdot \boldsymbol{o}_v, \tag{2}$$

where $\boldsymbol{E} \in \mathbb{R}^{(|\mathcal{U}|+|\mathcal{V}|)\times d}$ denotes the embedding lookup table of user/item IDs, and $\boldsymbol{o}_u, \boldsymbol{o}_v \in \mathbb{R}^{|\mathcal{U}|+|\mathcal{V}|}$ mean the corresponding one-hot vector of user/item IDs.

Denote the input embedding matrix $\boldsymbol{E} \in \mathbb{R}^{(|\mathcal{U}|+|\mathcal{V}|)\times d}$ as $\boldsymbol{E}^0$. We can formulate the abstraction of the hierarchical model for recommendation as:

$$\boldsymbol{E}^{l+1} = \text{Encoder}_{l+1}(\boldsymbol{E}^l), \tag{3}$$

where $\boldsymbol{E}^l$ means the output embedding matrix of the l-th encoder, and assuming there are L encoders in total.

4.1.2 Objective. The objective of our proposed HGIB can be given by:

$$\max \sum_{l=1}^{L} \left[I(\boldsymbol{E}^l; \boldsymbol{Y}) - \beta \cdot I(\boldsymbol{E}^l; \boldsymbol{E}^{l-1}) \right], \tag{4}$$

where $\boldsymbol{Y}$ denotes the label matrix of the target behavior interactions. In Equation 4, the first term constrains that the embedding outputs from each encoder layer preserve task-relevant information crucial for target behavior recommendation, while the second term forces the encoder to progressively compress task-irrelevant components within the learned representations.

4.2 Maximization of $I(\boldsymbol{E}^l; \boldsymbol{Y})$

Due to the intractability of directly computing mutual information, we strategically reformulate the optimization objective by deriving the lower bound of $I(\boldsymbol{E}^l; \boldsymbol{Y})$ as follows:

$$\begin{aligned}
\max\ I(\boldsymbol{E}^l; \boldsymbol{Y}) &= I(\boldsymbol{E}^L; \boldsymbol{Y}) + I(\boldsymbol{E}^l; \boldsymbol{Y}|\boldsymbol{E}^L) - I(\boldsymbol{E}^L; \boldsymbol{Y}|\boldsymbol{E}^l) \\
&\geq I(\boldsymbol{E}^L; \boldsymbol{Y}) - I(\boldsymbol{E}^L; \boldsymbol{Y}|\boldsymbol{E}^l) \\
&= [H(\boldsymbol{Y}) - H(\boldsymbol{Y}|\boldsymbol{E}^L)] - [H(\boldsymbol{E}^L|\boldsymbol{E}^l) - H(\boldsymbol{E}^L|\boldsymbol{E}^l, \boldsymbol{Y})] \\
&\geq -H(\boldsymbol{Y}|\boldsymbol{E}^L) - H(\boldsymbol{E}^L|\boldsymbol{E}^l) \\
&= -H(\boldsymbol{Y}|\boldsymbol{E}^L) - [H(\boldsymbol{E}^L) - I(\boldsymbol{E}^L; \boldsymbol{E}^l)] \\
&= \underbrace{-H(\boldsymbol{Y}|\boldsymbol{E}^L)}_{\text{Recommendation}}\ \underbrace{+I(\boldsymbol{E}^L; \boldsymbol{E}^l)}_{\text{Preservation}}\ \underbrace{-H(\boldsymbol{E}^L)}_{\text{Regularity}}.
\end{aligned} \tag{5}$$

According to the derivation in [27], the second term $I(\boldsymbol{E}^L; \boldsymbol{E}^l)$ is lower bounded by minus InfoNCE [27] loss as follows:

$$I(\boldsymbol{E}^L; \boldsymbol{E}^l) \geq \log(N) - \mathcal{L}_{CL}(\boldsymbol{E}^L, \boldsymbol{E}^l), \tag{6}$$

where N denotes the number of samples including both positive and negative ones, and $\mathcal{L}_{CL}$ means the InfoNCE loss for contrastive learning.

Moreover, the first term $-H(\boldsymbol{Y}|\boldsymbol{E}^L)$ is lower bounded by minus cross entropy loss, i.e., $-H(\boldsymbol{Y}|\boldsymbol{E}^L) \geq -\mathcal{L}_{CE}(\boldsymbol{E}^L, \boldsymbol{Y})$ where $\mathcal{L}_{CE}$ denotes the cross entropy loss, and the third term can be treated as the L_2-norm regularization of the trainable parameters of the model. Thus, the maximum optimization of $I(\boldsymbol{E}^l; \boldsymbol{Y})$ is equivalent to the following training objective:

$$\min \underbrace{\mathcal{L}_{CE}(\boldsymbol{E}^L, \boldsymbol{Y})}_{\mathcal{L}_{rec}} + \alpha \cdot \underbrace{\sum_{l=1}^{L-1} \mathcal{L}_{CL}(\boldsymbol{E}^L, \boldsymbol{E}^l)}_{\mathcal{L}_{pres}} + \lambda \cdot \underbrace{||\Theta||_2^2}_{\mathcal{L}_{reg}}, \tag{7}$$

where $\mathcal{L}_{rec}, \mathcal{L}_{pres}, \mathcal{L}_{reg}$ are corresponding to recommendation, preservation, and regularization in Equation 5, and Θ denotes the entire ensemble of trainable parameters in the model.

4.3 Minimization of $I(\boldsymbol{E}^l; \boldsymbol{E}^{l-1})$

In the discussion of maximization of $I(\boldsymbol{E}^l; \boldsymbol{Y})$, we find the lower bound of the mutual information $I(\boldsymbol{E}^L; \boldsymbol{E}^l)$ according to Equation 6. However, calculating the upper bound of mutual information is an intractable question. To address this challenge, some prior studies [20, 28, 39] demonstrate that the Hilbert-Schmidt Independence Criterion (HSIC) [7] can serve as a great approximation in the minimization of mutual information. Given two random variables $\boldsymbol{A}$ and $\boldsymbol{B}$, HSIC measures dependence between them using kernel matrices. For n observations $\{(a_i, b_i)\}_{i=1}^n$, define kernel matrices $\boldsymbol{K}^A$ (for $\boldsymbol{A}$, $\boldsymbol{K}^A_{ij} = k^A(a_i, a_j)$) and $\boldsymbol{K}^B$ (for $\boldsymbol{B}$, $\boldsymbol{K}^B_{ij} = k^B(b_i, b_j)$), centered by $\boldsymbol{H} = \boldsymbol{I} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^\top$. The empirical HSIC is calculated as:

$$\text{HSIC}(A, B) = \frac{1}{(n-1)^2} \text{tr}\left(\boldsymbol{K}^A \boldsymbol{H} \boldsymbol{K}^B \boldsymbol{H}\right), \tag{8}$$

where $\text{tr}(\cdot)$ denotes the matrix trace. HSIC is non-negative, and equals zero if and only if A and B are independent.

So the minimization optimization of $I(\boldsymbol{E}^l; \boldsymbol{E}^{l-1})$ can be reformulated as follows:

$$\min\ \mathcal{L}_{comp} = \sum_{l=1}^{L} \text{HSIC}(\boldsymbol{E}^l, \boldsymbol{E}^{l-1}), \tag{9}$$

where $\mathcal{L}_{comp}$ correspond to the compression part in Equation 1.

4.4 Denoising via Graph Refinement Encoder (GRE)

In addition to the previously discussed robust representation learning guided by the information bottleneck principle, we further propose a Graph Refinement Encoder (GRE) that adopts a learnable mechanism to further explicitly eliminate noisy interactions in user-item graphs, as shown in Figure 3 (b). To achieve this purpose, we formulate the edge refinement strategy as follows:

$$w_{uv} = \begin{cases} \sigma(\boldsymbol{e}_u^T \boldsymbol{e}_v), & \text{if } \sigma(\boldsymbol{e}_u^T \boldsymbol{e}_v) > \tau \\ 0, & \text{if } \sigma(\boldsymbol{e}_u^T \boldsymbol{e}_v) \leq \tau \end{cases} \tag{10}$$

where w_{uv} means the refined edge weights, and $\sigma(\cdot)$ denotes the sigmoid function and τ is the threshold to prune the noisy interactions. Due to the non-differentiability of the strategy function, we apply the Gumbel-softmax [12] reparameterization trick to enable end-to-end learning for the refinement strategy. After refinement, GRE employs the LightGCN [10] for the graph aggregation process.

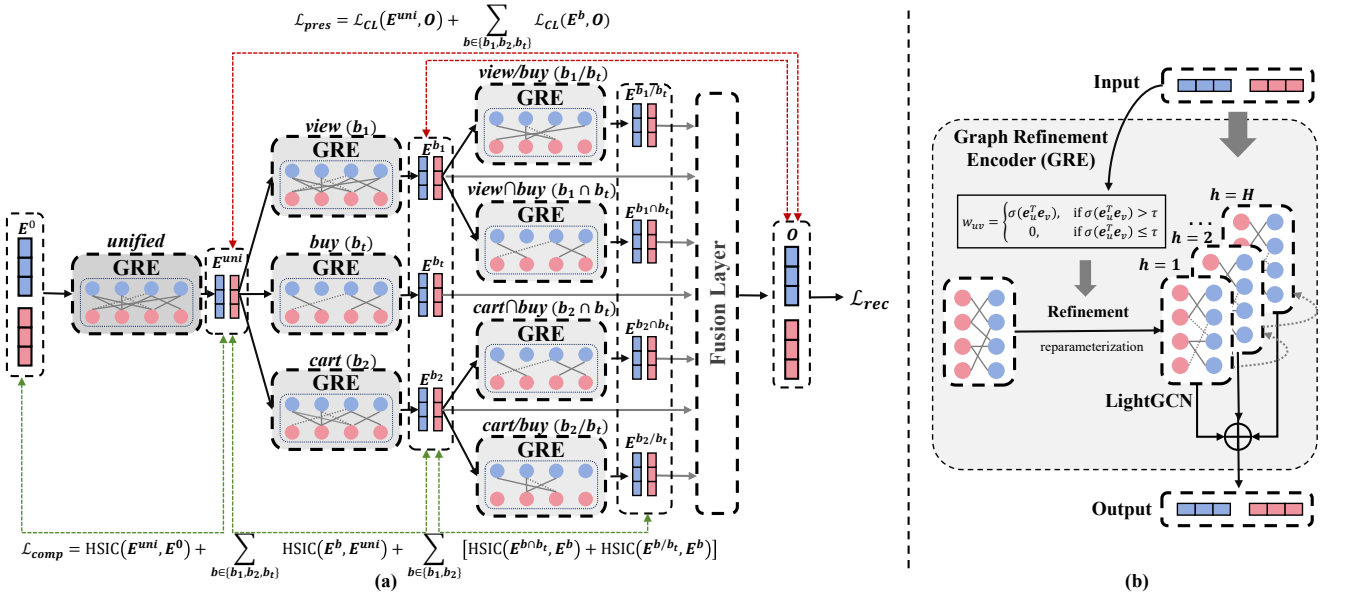


Figure 3: Illustration of the proposed HGIB, with three types of behavior as examples (view as b_1, cart as b_2, buy as b_t). (a) Instantiating of HGIB in the parallel paradigm, where the red dotted line represents $\mathcal{L}_{pres}$ for preservation and the green dotted line represents $\mathcal{L}_{comp}$ for compression. (b) Illustration of Graph Refinement Encoder (GRE), which denoises the noisy interaction by a learnable strategy.

4.5 Instantiation of HGIB

In this section, we instantiate our proposed HGIB with a specific parallel paradigm backbone simplified from the strongest SOTA method MULE [15], as shown in Figure 3 (a).

4.5.1 Model Structure. Given the initial embedding $\boldsymbol{E}^0$ and user-item interaction graphs $\{\mathcal{G}^b\}_{b\in\{b_1,b_2,b_t\}}$, where b_1, b_2, b_t represent view, cart, buy, the model structure of the HGIB instantiation can be summarized as follows:

Unified Encoder:

$$\boldsymbol{E}^{uni} = \text{GRE}(\mathcal{G}^{uni}, \boldsymbol{E}^0), \tag{11}$$

where $\mathcal{G}^{uni} = \bigcup_{b\in\{b_1,b_2,b_t\}}\{\mathcal{G}^b\}$, i.e., merging all the interactions from multiple behaviors, and GRE denotes the graph refinement encoder proposed in Section 4.4.

Behavior-Specific Encoder:

$$\boldsymbol{E}^b = \text{GRE}(\mathcal{G}^b, \boldsymbol{E}^{uni}), \text{ for } b \in \{b_1, b_2, b_t\}, \tag{12}$$

Behavior-Component Encoder:
For $b \in \{b_1, b_2\}$,

$$\boldsymbol{E}^{b\cap b_t} = \text{GRE}(\mathcal{G}^{b\cap b_t}, \boldsymbol{E}^b), \boldsymbol{E}^{b/b_t} = \text{GRE}(\mathcal{G}^{b/b_t}, \boldsymbol{E}^b), \tag{13}$$

where $\mathcal{G}^{b\cap b_t} = \mathcal{G}^b \bigcap \mathcal{G}^{b_t}$,i.e., the intersection of $\mathcal{G}^b$ and $\mathcal{G}^{b_t}$, indicates the user-item interaction graph of view->buy ($b_1 \cap b_t$) and cart->buy ($b_2 \cap b_t$), and $\mathcal{G}^{b/b_t} = \mathcal{G}^b - \mathcal{G}^{b_t}$,i.e., the difference set of $\mathcal{G}^b$ and $\mathcal{G}^{b_t}$, indicates the user-item interaction graph of view-but-not-buy (b_1/b_t) and cart-but-not-buy (b_2/b_t).

Fusion Layer: Finally, we apply the widely-used Target Attention (TA) [23, 43, 48, 49] to aggregate the final representation $\boldsymbol{O}$:

$$\begin{aligned}\hat{\boldsymbol{E}} &= \text{TA}\left(\boldsymbol{E}^t, \{\boldsymbol{E}^b\}_{b\in\{b_1,b_2,b_t\}}\right),\\ \boldsymbol{O} &= \text{TA}\left(\hat{\boldsymbol{E}}, \{\boldsymbol{E}^{b'}\}_{b'\in\{b\cap b_t, b/b_t | b\in\{b_1,b_2\}\}}\right),\end{aligned} \tag{14}$$

where $\text{TA}(\boldsymbol{q}, \boldsymbol{k})$ denotes apply the target attention to aggregate keys $\boldsymbol{k}$ with query $\boldsymbol{q}$.

4.5.2 Optimization Objective. Summarizing the discussion in Section 4.1-4.3, the objective function of HGIB instantiation can be formulated as follows:

$$\min \mathcal{L}_{rec} + \alpha \cdot \mathcal{L}_{pres} + \beta \cdot \mathcal{L}_{comp} + \gamma \cdot \mathcal{L}_{reg}, \tag{15}$$

where $\mathcal{L}_{rec}$ is the cross-entropy loss for target behavior prediction and $\mathcal{L}_{reg}$ for L_2-norm regularization. The preservation loss $\mathcal{L}_{pres}$ and the compression loss $\mathcal{L}_{comp}$ can be given by:

$$\begin{aligned}\mathcal{L}_{pres} &= \mathcal{L}_{CL}\left(\boldsymbol{E}^{uni}, \boldsymbol{O}\right) + \sum_{b\in\{b_1,b_2,b_t\}} \mathcal{L}_{CL}\left(\boldsymbol{E}^b, \boldsymbol{O}\right),\\ \mathcal{L}_{comp} &= \text{HSIC}\left(\boldsymbol{E}^{uni}, \boldsymbol{E}^0\right) + \sum_{b\in\{b_1,b_2,b_t\}} \text{HSIC}\left(\boldsymbol{E}^b, \boldsymbol{E}^{uni}\right)\\ &\quad + \sum_{b\in\{b_1,b_2\}} \left[\text{HSIC}\left(\boldsymbol{E}^{b\cap b_t}, \boldsymbol{E}^b\right) + \text{HSIC}\left(\boldsymbol{E}^{b/b_t}, \boldsymbol{E}^b\right)\right].\end{aligned} \tag{16}$$

4.6 Further Analysis

4.6.1 Complexity Analysis. Compared to the backbone model, the additional time complexity introduced by HGIB arises primarily from the calculation of the auxiliary loss $\mathcal{L}_{pres}, \mathcal{L}_{comp}$ and Graph Refinement Encoder (GRE). Because the time complexity of InfoNCE is $O(B^2d)$ where B denotes the batch size, the additional

complexity from $\mathcal{L}_{pres}$ is $O(B^2Ld)$ where L is the number of encoders. The time complexity of $\mathcal{L}_{comp}$ can be given by $O(L)$ times the complexity of calculating HSIC, which results in $O(B^3L+B^2Ld)$. And the calculation for GRE takes $O(|\mathcal{E}|Ld)$ time. In summary, the additional time complexity introduced by HGIB can be given by $O\left((B^2+|\mathcal{E}|)Ld+B^3L\right)$. Since the time complexity of commonly used backbone models (e.g., LightGCN) is already at least $O(|\mathcal{E}|Ld)$ ($|\mathcal{E}| \gg B$), the proposed HGIB framework does not impose a significant computational overhead on the backbone network.

4.6.2 Compatibility Analysis. Our proposed HGIB is a model-agnostic framework for multi-behavior recommendation, which can be seamlessly compatible with other hierarchical multi-behavior methods. For applications to methods without hierarchical design, HGIB can be degenerated into the standard IB. Our proposed HGIB exhibits strong generalization across different hierarchical multi-behavior methods, as validated by the comprehensive compatibility analysis experiments in Section 5.4.

5 Experiments

To comprehensively evaluate the performance of our proposed HGIB, we conducted extensive experiments on three public real-world datasets to explore the following research questions:

- **(RQ1)** How does HGIB perform compared to other state-of-the-art multi-behavior methods?
- **(RQ2)** How do the different components affect the prediction performance of HGIB, respectively?
- **(RQ3)** Can HGIB be compatible with other multi-behavior methods?
- **(RQ4)** How do different hyperparameters affect the prediction performance of HGIB, respectively?
- **(RQ5)** How does HGIB perform when applied to real industrial scenarios?
- **(RQ6)** Why does HGIB achieve better performance?

5.1 Experiment Setting

5.1.1 Dataset Description. To investigate the effectiveness of our proposed HGIB, we conduct comprehensive experiments on three popular public datasets for multi-behavior recommendation, including Taobao, Tmall, and Jdata datasets. All these datasets are gathered from the top e-commerce platforms, such as taobao.com and jd.com in China. The statistical information of the datasets is summarized in Table 1. Taobao dataset includes three types of behavior: *view*, *add-to-cart*, and *buy*, while Tmall and Jdata datasets also have *add-to-collect* behavior apart from these. For all datasets, we treat *buy* as the target behavior. For preprocessing, we follow the settings of the previous works for these datasets.

Table 1: Statistics of public datasets for the experiment.

Dataset	#Users	#Items	#Views	#Collects	#Carts	#Buys
Taobao	15,449	11,953	873,954	-	195,476	92,180
Tmall	41,738	11,953	1,813,498	221,514	1,996	255,586
Jdata	93,334	24,624	1,681,430	45,613	49,891	321,883

5.1.2 Evaluation Metrics. For all our experiments, we assess the Top-K recommendation performance of methods for predicting the target behavior (i.e., *buy*) by two widely used metrics: HR@K (Hit Ratio) and NDCG@K (Normalized Discounted Cumulative Gain), where $K = 10$ for our evaluation specifically. HR@K is a recall-based metric that measures the average fraction of correct items among the Top-K recommendations. NDCG@K, on the other hand, evaluates the quality of the ranking in the Top-K recommendations by considering the position of each item, assigning higher importance to items ranked higher in the list. And we apply the full-ranking setting to sort the entire item set to get the Top-K recommendation results.

5.1.3 Baseline Methods. To comprehensively validate the effectiveness of our proposed HGIB, we conduct comparisons with diverse baseline models, which can be categorized into three groups: (1) **Single behavior methods**: MF-BPR [24], NCF [11], LightGCN [10]; (2) **Multi-behavior methods without hierarchical design**: RGCN [26], GNMR [31], NMTR [6], MBGCN [13]; (3) **Multi-behavior methods with hierarchical design**: CRGCN [36], MBCGCN [3], HPMR [22], PKEF [21], AutoDCS [18], COPF [42], MULE [15].

5.1.4 Implement Details. For a fair comparison, we strictly follow the data partitioning protocols and settings in prior works [3, 15, 36] for training-testing splitting. For baseline methods, we adopted their official open-source implementations. We configure the embedding dimension as $d = 64$, and set the batch size to 1024 for all implemented methods.

For our proposed HGIB, the source code is available [1]. The hyperparameters of HGIB are set as follows: the coefficient α is set to 1.0 for the Taobao and Tmall dataset and 0.5 for the Jdata dataset; the coefficient $\beta = 50$ and the threshold $\tau = 0.05$ in GRE are set as the default for all datasets. And the regularization coefficient λ is fixed to 0.1. For the model's optimization, we use the Adam optimizer with a learning rate of 5×10^{-4}, training the model for a maximum of 100 epochs.

5.2 Overall Performance (RQ1)

The overall evaluation of the proposed HGIB against thirteen baseline models is summarized in Table 2.

Drawing from these results, we summarize the following observations:

- Leveraging multi-behavior data is essential, as single-behavior methods (e.g., LightGCN) consistently underperform against most of the multi-behavior methods.
- Multi-behavior methods with hierarchical designs consistently achieve superior performance compared to other approaches, regardless of whether they follow a cascading or parallel paradigm. By hierarchically modeling the relationships between multiple behaviors, these methods significantly enhance recommendation accuracy.
- Due to the high conversion rates in the Jdata dataset (e.g., 43% of "collect" behaviors and 57% of "cart" behaviors convert to "buy"), the natural cascading sequence of behaviors is more pronounced. Thus, the cascading paradigm tends to outperform the parallel paradigm in this dataset.

[1] https://github.com/zhy99426/HGIB

Table 2: Overall evaluation results of our proposed HGIB and state-of-the-art baselines on three real-world multi-behavior datasets. Results show the best-performing method in bold and the runner-up underlined. "Rel Impr." indicates the relative improvements compared to the strongest baseline. An asterisk (*) denotes statistical significance (p < 0.05) when comparing HGIB to the strongest baseline results.

Method	Taobao		Tmall		Jdata	
	HR	NDCG	HR	NDCG	HR	NDCG
MF-BPR	0.0076	0.0036	0.0230	0.0207	0.1850	0.1238
NCF	0.0236	0.0128	0.0124	0.0062	0.2090	0.1410
LightGCN	0.0411	0.0240	0.0393	0.0209	0.2252	0.1436
RGCN	0.0215	0.0104	0.0316	0.0157	0.2406	0.1444
GNMR	0.0368	0.0216	0.0393	0.0193	0.3068	0.1581
NMTR	0.0282	0.0137	0.0536	0.0286	0.3142	0.1717
MBGCN	0.0509	0.0294	0.0549	0.0285	0.2803	0.1572
CRGCN	0.0855	0.0439	0.0840	0.0442	0.5001	0.2914
MB-CGCN	0.1233	0.0677	0.0984	0.0558	0.4349	0.2758
HPMR	0.1104	0.0599	0.0956	0.0515	0.3260	0.2029
PKEF	0.1385	0.0785	0.1277	0.0721	0.4334	0.2615
AutoDCS	0.1522	0.0813	0.1432	0.0743	0.6174	0.4559
COPF	0.1552	0.0838	0.1755	0.0967	0.5338	0.3692
MULE	0.1939	0.1109	0.2109	0.1165	0.5820	0.4147
HGIB	**0.2203***	**0.1214***	**0.2427***	**0.1287***	**0.6552***	**0.4747***
Rel Impr.	13.62%	9.47%	15.08%	10.47%	6.12%	4.12%

- Our proposed HGIB framework achieves consistently superior performance across all datasets, demonstrating its exceptional effectiveness. This is attributed to HGIB's use of the information bottleneck principle, which compels the model to retain information relevant to target behavior prediction while discarding irrelevant information. Furthermore, the GRE module filters out noisy interactions in the multi-behavior user-item interaction graphs, thereby mitigating negative transfer.

5.3 Ablation Study (RQ2)

As discussed in Section 4, the main contribution of HGIB includes the introduction of the auxiliary loss $\mathcal{L}_{pres}$, $\mathcal{L}_{comp}$, and the GRE module. To evaluate the effectiveness of these three components, we conduct an ablation study experiment by removing each component from the HGIB framework, resulting in three variant models: "w/o $\mathcal{L}_{pres}$", "w/o $\mathcal{L}_{comp}$" and "w/o GRE", where "w/o $\mathcal{L}_{pres}$" and "w/o $\mathcal{L}_{comp}$" are implemented by removing the corresponding loss, and w/o GRE is implemented by replacing Graph Refinement Encoder (GRE) with LightGCN [10].

The results of the ablation study on three multi-behavior datasets are presented in Table 3. Comparing the results, we can draw the following observations:

- The experiment results demonstrate that eliminating any key component results in a decline in model performance, thereby demonstrating the essential role and effectiveness of each component within the proposed HGIB framework. Specifically, both the preservation of task-relevant information and the compression of redundant information are necessary in the information bottleneck principle for robust representation learning. Furthermore, the explicit denoising implemented by GRE effectively mitigates the negative transfer problem caused by noise interference.
- The variant model "w/o $\mathcal{L}_{pres}$" exhibits the most significant performance degradation compared to HGIB, highlighting the critical importance of retaining task-relevant information in the information bottleneck framework. In multi-behavior recommendation scenarios, the $\mathcal{L}_{pres}$ objective requires the model to extract valuable knowledge specifically beneficial for the target behavior through representation learning. Building upon this foundation, the $\mathcal{L}_{comp}$ and GRE further compel the model to eliminate noise information irrelevant to the target behavior recommendation task.

Table 3: Ablation study results of different components in HGIB. Cyan indicates performance decrease compared to HGIB, with deeper shades representing greater relative changes.

Method	Taobao		Tmall		Jdata	
	HR	NDCG	HR	NDCG	HR	NDCG
HGIB	**0.2203**	**0.1214**	**0.2427**	**0.1287**	**0.6552**	**0.4747**
w/o $\mathcal{L}_{pres}$	0.1870	0.1034	0.1799	0.0980	0.6172	0.4551
w/o $\mathcal{L}_{comp}$	0.2188	0.1204	0.2322	0.1222	0.6397	0.4534
w/o GRE	0.2110	0.1167	0.2200	0.1164	0.6319	0.4491

5.4 Compatibility Analysis with Different Backbones (RQ3)

Our proposed HGIB is a model-agnostic framework capable of being integrated with other multi-behavior models with hierarchical design. To evaluate its adaptability, we implement the SOTA model of the cascading paradigm, AutoDCS [18], and of the parallel paradigm, MULE [15], as backbones to integrate into the HGIB framework. We conduct compatibility analysis experiments through comparison of performance between original backbone models and their enhanced variants integrated into the HGIB framework, and the results are presented in Table 4. The "Base" in Table 4 means the backbone model of HGIB introduced in Section 4.5.1.

As the experimental results in Table 4 show, HGIB consistently achieves significant performance gains across different backbone models. The performance improvement mainly stems from the fact that HGIB follows the information bottleneck principle enforces preservation of knowledge relevant to the prediction of the target behavior during hierarchical representation learning, while compressing noise and bias introduced by auxiliary behaviors. It demonstrates the broad effectiveness and general superiority of our proposed HGIB framework for multi-behavior recommendations.

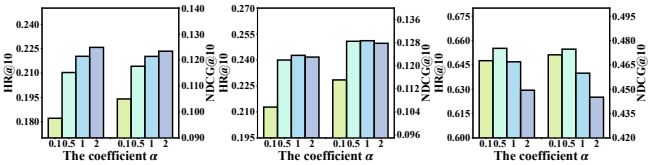
5.5 Hyperparameter Analysis (RQ4)

In this section, we will discuss the key hyperparameters α and β in our proposed HGIB controlling the extent of preservation and compression in representation learning, respectively. To evaluate how these hyperparameters influence HGIB's performance, we conduct a comparative study by varying their performance across

Table 4: Compatibility analysis with different backbone models on three multi-behavior datasets. "Rel Impr." indicates the relative improvements compared to the corresponding backbones, and deeper orange shades represent greater gains.

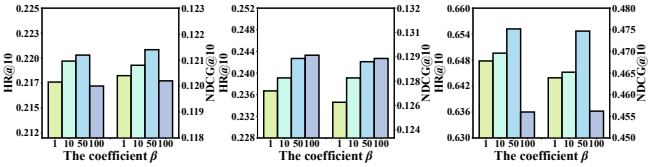
Method	Taobao		Tmall		Jdata	
	HR	NDCG	HR	NDCG	HR	NDCG
Base	0.1679	0.0970	0.1655	0.0890	0.6063	0.4442
HGIB	0.2203	0.1214	0.2427	0.1287	0.6552	0.4747
Rel Impr.	31.21%	25.15%	46.65%	44.61%	8.07%	6.87%
AutoDCS	0.1522	0.0813	0.1432	0.0743	0.6174	0.4559
HGIB(AutoDCS)	0.1674	0.0926	0.1578	0.0841	0.6198	0.4571
Rel Impr.	9.99%	13.90%	10.20%	13.19%	0.39%	0.26%
MULE	0.1939	0.1109	0.2109	0.1165	0.5820	0.4147
HGIB(MULE)	0.2183	0.1192	0.2395	0.1268	0.6326	0.4475
Rel Impr.	12.58%	7.48%	13.56%	8.84%	8.69%	7.91%

public multi-behavior recommendation datasets. During experiments, while testing different configurations of α and β, all other model parameters are maintained at their default settings to isolate their individual effects.



(a) Taobao (b) Tmall (c) Jdata

Figure 4: Performance with the coefficient α for $\mathcal{L}_{pres}$.



(a) Taobao (b) Tmall (c) Jdata

Figure 5: Performance with the coefficient β for $\mathcal{L}_{comp}$.

The comparative analysis of hyperparameters is illustrated in Figure 4 and 5, leading to the following insights:

- **The coefficient α:** For Taobao and Tmall datasets, as α increases, the model is compelled to preserve task-relevant information in intermediate representations during the learning process, leading to gradual performance improvement. And performance gains plateau or even slightly decline when α exceeds to some extent. For the Jdata dataset, due to its inherent characteristics with a high conversion rate, the prediction task is relatively simple. Excessively large α values can disrupt the model's focus on the main recommendation task, resulting in peak performance occurring at $\alpha = 0.5$.
- **The coefficient β:** Generally, HGIB achieves peak performance with a moderate β value, so we set the default $\beta = 50$ for all datasets. This occurs because overly small β values fail to leverage the information bottleneck principle to compress task-irrelevant information; while excessively large β values lead to over-compression, impairing the model's prediction accuracy.

5.6 Application on Industrial Scenarios (RQ5)

Beyond the e-commerce scenarios covered by the three public datasets, our proposed HGIB framework demonstrates extensive applicability across diverse real-world industrial scenarios for multi-behavior recommendations. To validate this, we conduct comparative experiments on two industrial datasets from **WeChat**, one of China's largest social platforms, regarding news recommendation and live-streaming recommendation scenarios.

Table 5: Statistics of industry datasets.

News		Live Stream	
Metric	**Count**	**Metric**	**Count**
#Users	26.81 million	#Users	32.25 million
#News	2.21 million	#Streamers	0.59 million
#Clicks	3.14 billion	#Clicks	2.64 billion
#Finished	1.46 billion	#Likes	0.15 billion
#Likes	29.64 million	#Comments	32.57 million
#Shares	42.79 million	#Gifts	16.60 million

5.6.1 Experiment Description. The industrial datasets contain billions of interaction records from tens of millions of users across two scenarios: news recommendation (in a 15-day period) and live streaming recommendation (in a 90-day period). 'Shares' is the target behavior for the news recommendation to promote social interaction and user stickiness, and 'Gifts' is the target behavior for the live stream recommendation to increase the revenue of live streamers and the platform and promote the commercialization of the platform.

All data are subjected to rigorous privacy-preserving processing, and we follow the 'leave-one-out' evaluation strategy for testing. The detailed statistics of industrial datasets are shown in Table 5.

Table 6: Performance comparison on industrial datasets.

Method	News		Live Stream	
	HR	NDCG	HR	NDCG
MULE	0.1448	0.0701	0.1471	0.0712
HGIB	**0.1663**	**0.0825**	**0.1552**	**0.0764**
Rel Impr.	14.85%	17.69%	5.51%	7.30%

5.6.2 Performance Comparison. We compare our proposed HGIB with the strongest SOTA multi-behavior baseline, MULE [15], and evaluate performance using HR@10 and NDCG@10. The results, presented in Table 6, show that our method achieves relative improvements of 14.85%/17.69% in News Recommendation and 5.51%/7.30% in Live Stream Recommendation over MULE in terms of HR and NDCG. These results demonstrate that our proposed HGIB can effectively leverage the rich and diverse multi-behavior data of users for large-scale industrial recommendations.

5.7 Online A/B Testing (RQ5)

To evaluate our model's online performance in real industrial scenarios, we conducted a 15-day online A/B testing in the news recommendation ranking task of the WeChat platform from April 20, 2025, to May 5, 2025. WeChat Subscription Accounts provide daily news recommendation services for billions of users. We randomly served 2% online traffic with our model, which contains over 20 million users, and another 2% with the SOTA baseline model, MULE.

We used widely used metrics including **Click-Through Rate** (CTR) and **Social Sharing Rate** (SSR), to evaluate the online effectiveness. The experimental results demonstrated statistically significant improvements of **0.32%** in CTR ($p<0.05$) and **1.65%** in SSR ($p<0.01$) over the SOTA baseline. The longitudinal validation demonstrates HGIB's effectiveness in enhancing user engagement on large-scale social platforms.

5.8 Why It Works? (RQ6)

In this section, we will discuss why our proposed HGIB achieves significant performance improvements compared to other baselines. To investigate this, we introduce the concept of information abundance from [9], which measures the richness of information learned by embeddings. A low information abundance indicates that the embedding set tends to collapse into a low-rank matrix.

Definition 5.1 (Information Abundance). Given an embedding matrix $\boldsymbol{E} \in \mathbb{R}^{M\times d}$ where M denotes the number of embeddings, and its singular value decomposition $\boldsymbol{E} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V} = \sum_{k=1}^{d} \sigma_k \boldsymbol{u}_k \boldsymbol{v}_k^\top$. The information abundance of $\boldsymbol{E}$ is defined as:

$$\mathrm{IA}(\boldsymbol{E}) = \frac{\|\boldsymbol{\sigma}\|_1}{\|\boldsymbol{\sigma}\|_\infty},$$

i.e., the sum of all singular values normalized by the maximum singular value.

We visualize the information abundance of embeddings output by each encoder on the Taobao dataset in our proposed HGIB, the backbone Base, and the baseline MULE, as shown in Figure 6.

In Figure 6, we can clearly observe that the embeddings learned by HGIB exhibit the highest information abundance. Additionally, the information abundance of embeddings in HGIB decreases with increasing encoder levels. This aligns with the essence of the hierarchical funnel design: the shallow unified graph contains the richest interaction data, including views, carts, and buys, while subsequent encoders model subgraphs that capture different behaviors and their relationships. As the encoder level increases, HGIB extracts task-relevant information while discarding other redundant information. In contrast, the shallow-encoder embeddings learned by the Base and MULE show insufficiently learned knowledge with low information abundance. This deviates from the principle of hierarchical design, limiting the learning of subsequent encoders and the overall performance of the model.

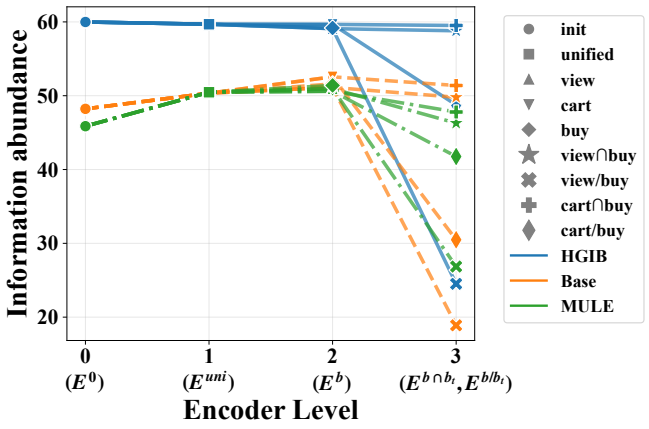


Figure 6: Comparison of information abundance of the embeddings in different models.

6 Conclusions

In this paper, we highlight critical challenges faced by the multi-behavior recommendation methods, including imbalanced behavior distributions and negative transfer caused by noisy interactions. To address these challenges, we proposed the Hierarchical Graph Information Bottleneck (HGIB) framework, which integrates information bottleneck principles into hierarchical multi-behavior modeling to retain task-relevant knowledge while compressing noise and irrelevant information. By optimizing an information-theoretic objective and employing the Graph Refinement Encoder (GRE) to prune noisy interactions explicitly, HGIB performs robust representation learning to mitigate overfitting and negative transfer effects. Extensive experiments on public and industrial datasets validate the superiority and generality of our approach, demonstrating its effectiveness in multi-behavior representation learning and its potential to enhance real-world recommendation systems.

Acknowledgments

This work is sponsored by Tencent WeChat Rhino-Bird Focused Research Program. Xiangguo Sun and Hong Cheng are supported by project #MMT-p2-23 of the Shun Hing Institute of Advanced Engineering, The Chinese University of Hong Kong and by grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (No. CUHK 14217622).

References

[1] Chong Chen, Weizhi Ma, Min Zhang, Zhaowei Wang, Xiuqiang He, Chenyang Wang, Yiqun Liu, and Shaoping Ma. 2021. Graph Heterogeneous Multi-Relational Recommendation. In *AAAI*. AAAI Press, 3958–3966.

[2] Chong Chen, Min Zhang, Yongfeng Zhang, Weizhi Ma, Yiqun Liu, and Shaoping Ma. 2020. Efficient Heterogeneous Collaborative Filtering without Negative Sampling for Recommendation. In *AAAI*. AAAI Press, 19–26.

[3] Zhiyong Cheng, Sai Han, Fan Liu, Lei Zhu, Zan Gao, and Yuxin Peng. 2023. Multi-Behavior Recommendation with Cascading Graph Convolution Networks. In *WWW*. ACM, 1181–1189.

[4] Jingtao Ding, Guanghui Yu, Xiangnan He, Yuhan Quan, Yong Li, Tat-Seng Chua, Depeng Jin, and Jiajie Yu. 2018. Improving Implicit Recommender Systems with View Data. In *IJCAI*. ijcai.org, 3343–3349.

[5] Shuheng Fang, Kangfei Zhao, Yu Rong, Jeffrey Xu Yu, and Zhixun Li. 2025. All-in-One: Heterogeneous Interaction Modeling for Cold-Start Rating Prediction . In *ICDE*. IEEE Computer Society, Los Alamitos, CA, USA, 1537–1550. doi:10.1109/ICDE65448.2025.00119

[6] Chen Gao, Xiangnan He, Dahua Gan, Xiangning Chen, Fuli Feng, Yong Li, Tat-Seng Chua, and Depeng Jin. 2019. Neural Multi-task Recommendation from Multi-behavior Data. In *ICDE*. IEEE, 1554–1557.

[7] Arthur Gretton, Olivier Bousquet, Alexander J. Smola, and Bernhard Schölkopf. 2005. Measuring Statistical Dependence with Hilbert-Schmidt Norms. In *ALT (Lecture Notes in Computer Science, Vol. 3734)*. Springer, 63–77.

[8] Long Guo, Lifeng Hua, Rongfei Jia, Binqiang Zhao, Xiaobo Wang, and Bin Cui. 2019. Buying or Browsing?: Predicting Real-time Purchasing Intent using Attention-based Deep Network with Multiple Behavior. In *KDD*. ACM, 1984–1992.

[9] Xingzhuo Guo, Junwei Pan, Ximei Wang, Baixu Chen, Jie Jiang, and Mingsheng Long. 2024. On the Embedding Collapse when Scaling up Recommendation Models. In *ICML*. OpenReview.net.

[10] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yong-Dong Zhang, and Meng Wang. 2020. LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation. In *SIGIR*. ACM, 639–648.

[11] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural Collaborative Filtering. In *WWW*. ACM, 173–182.

[12] Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical Reparameterization with Gumbel-Softmax. In *ICLR (Poster)*. OpenReview.net.

[13] Bowen Jin, Chen Gao, Xiangnan He, Depeng Jin, and Yong Li. 2020. Multi-behavior Recommendation with Graph Convolutional Networks. In *SIGIR*. ACM, 659–668.

[14] Artus Krohn-Grimberghe, Lucas Drumond, Christoph Freudenthaler, and Lars Schmidt-Thieme. 2012. Multi-relational matrix factorization using bayesian personalized ranking for social network data. In *WSDM*. ACM, 173–182.

[15] Seunghan Lee, Geonwoo Ko, Hyun-Je Song, and Jinhong Jung. 2024. MuLe: Multi-Grained Graph Learning for Multi-Behavior Recommendation. In *CIKM*. ACM, 1163–1173.

[16] Dugang Liu, Pengxiang Cheng, Hong Zhu, Zhenhua Dong, Xiuqiang He, Weike Pan, and Zhong Ming. 2021. Mitigating Confounding Bias in Recommendation via Information Bottleneck. In *RecSys*. ACM, 351–360.

[17] Dugang Liu, Pengxiang Cheng, Hong Zhu, Zhenhua Dong, Xiuqiang He, Weike Pan, and Zhong Ming. 2023. Debiased Representation Learning in Recommendation via Information Bottleneck. *Trans. Recomm. Syst.* 1, 1 (2023), 1–27.

[18] Dugang Liu, Shenxian Xian, Yuhao Wu, Chaohua Yang, Xing Tang, Xiuqiang He, and Zhong Ming. 2024. AutoDCS: Automated Decision Chain Selection in Deep Recommender Systems. In *SIGIR*. ACM, 956–965.

[19] Babak Loni, Roberto Pagano, Martha A. Larson, and Alan Hanjalic. 2016. Bayesian Personalized Ranking with Multi-Channel User Feedback. In *RecSys*. ACM, 361–364.

[20] Kurt Wan-Duo Ma, J. P. Lewis, and W. Bastiaan Kleijn. 2020. The HSIC Bottleneck: Deep Learning without Back-Propagation. In *AAAI*. AAAI Press, 5085–5092.

[21] Chang Meng, Chenhao Zhai, Yu Yang, Hengyu Zhang, and Xiu Li. 2023. Parallel Knowledge Enhancement based Framework for Multi-behavior Recommendation. In *CIKM*. ACM, 1797–1806.

[22] Chang Meng, Hengyu Zhang, Wei Guo, Huifeng Guo, Haotian Liu, Yingxue Zhang, Hongkun Zheng, Ruiming Tang, Xiu Li, and Rui Zhang. 2023. Hierarchical Projection Enhanced Multi-behavior Recommendation. In *KDD*. ACM, 4649–4660.

[23] Qi Pi, Guorui Zhou, Yujing Zhang, Zhe Wang, Lejian Ren, Ying Fan, Xiaoqiang Zhu, and Kun Gai. 2020. Search-based User Interest Modeling with Lifelong Sequential Behavior Data for Click-Through Rate Prediction. In *CIKM*. ACM, 2685–2692.

[24] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian Personalized Ranking from Implicit Feedback. In *UAI*. AUAI Press, 452–461.

[25] Yu Rong, Xiao Wen, and Hong Cheng. 2014. A Monte Carlo algorithm for cold start recommendation. In *WWW*. ACM, 327–336.

[26] Michael Sejr Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2018. Modeling Relational Data with Graph Convolutional Networks. In *ESWC (Lecture Notes in Computer Science, Vol. 10843)*. Springer, 593–607.

[27] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation Learning with Contrastive Predictive Coding. *CoRR* abs/1807.03748 (2018).

[28] Zifeng Wang, Tong Jian, Aria Masoomi, Stratis Ioannidis, and Jennifer G. Dy. 2021. Revisiting Hilbert-Schmidt Information Bottleneck for Adversarial Robustness. In *NeurIPS*. 586–597.

[29] Chunyu Wei, Jian Liang, Di Liu, and Fei Wang. 2022. Contrastive Graph Structure Learning via Information Bottleneck for Recommendation. In *NeurIPS*.

[30] Wei Wei, Chao Huang, Lianghao Xia, Yong Xu, Jiashu Zhao, and Dawei Yin. 2022. Contrastive Meta Learning with Behavior Multiplicity for Recommendation. In *WSDM*. ACM, 1120–1128.

[31] Lianghao Xia, Chao Huang, Yong Xu, Peng Dai, Mengyin Lu, and Liefeng Bo. 2021. Multi-Behavior Enhanced Recommendation with Cross-Interaction Collaborative Relation Modeling. In *ICDE*. IEEE, 1931–1936.

[32] Lianghao Xia, Chao Huang, Yong Xu, Peng Dai, Bo Zhang, and Liefeng Bo. 2020. Multiplex Behavioral Relation Learning for Recommendation via Memory Augmented Transformer Network. In *SIGIR*. ACM, 2397–2406.

[33] Lianghao Xia, Yong Xu, Chao Huang, Peng Dai, and Liefeng Bo. 2021. Graph Meta Network for Multi-Behavior Recommendation. In *SIGIR*. ACM, 757–766.

[34] Junsong Xie, Yonghui Yang, Zihan Wang, and Le Wu. 2024. Learning Fair Representations for Recommendation via Information Bottleneck Principle. In *IJCAI*. ijcai.org, 2469–2477.

[35] Jingcao Xu, Chaokun Wang, Cheng Wu, Yang Song, Kai Zheng, Xiaowei Wang, Changping Wang, Guorui Zhou, and Kun Gai. 2023. Multi-behavior Self-supervised Learning for Recommendation. In *SIGIR*. ACM, 496–505.

[36] Mingshi Yan, Zhiyong Cheng, Chen Gao, Jing Sun, Fan Liu, Fuming Sun, and Haojie Li. 2024. Cascading Residual Graph Convolutional Network for Multi-Behavior Recommendation. *ACM Trans. Inf. Syst.* 42, 1 (2024), 10:1–10:26.

[37] Mingshi Yan, Zhiyong Cheng, Jing Sun, Fuming Sun, and Yuxin Peng. 2023. MB-HGCN: A Hierarchical Graph Convolutional Network for Multi-behavior Recommendation. *CoRR* abs/2306.10679 (2023).

[38] Yonghui Yang, Le Wu, Zhuangzhuang He, Zhengwei Wu, Richang Hong, and Meng Wang. 2025. Less is More: Information Bottleneck Denoised Multimedia Recommendation. *CoRR* abs/2501.12175 (2025).

[39] Yonghui Yang, Le Wu, Zihan Wang, Zhuangzhuang He, Richang Hong, and Meng Wang. 2024. Graph Bottlenecked Social Recommendation. In *KDD*. ACM, 3853–3862.

[40] Junchi Yu, Tingyang Xu, Yu Rong, Yatao Bian, Junzhou Huang, and Ran He. 2021. Graph Information Bottleneck for Subgraph Recognition. In *ICLR*. OpenReview.net.

[41] Junchi Yu, Tingyang Xu, Yu Rong, Yatao Bian, Junzhou Huang, and Ran He. 2021. Graph Information Bottleneck for Subgraph Recognition. In *ICLR*.

[42] Chenhao Zhai, Chang Meng, Yu Yang, Kexin Zhang, Xuhao Zhao, and Xiu Li. 2025. Combinatorial Optimization Perspective based Framework for Multi-behavior Recommendation. In *KDD (1)*. ACM, 1891–1902.

[43] Hengyu Zhang, Junwei Pan, Dapeng Liu, Jie Jiang, and Xiu Li. 2024. Deep Pattern Network for Click-Through Rate Prediction. In *SIGIR*. ACM, 1189–1199.

[44] Hengyu Zhang, Chunxu Shen, Xiangguo Sun, Jie Tan, Yu Rong, Chengzhi Piao, Hong Cheng, and Lingling Yi. 2025. Adaptive Graph Integration for Cross-Domain Recommendation via Heterogeneous Graph Coordinators. In *SIGIR*. ACM, 1860–1869.

[45] Hengyu Zhang, Enming Yuan, Wei Guo, Zhicheng He, Jiarui Qin, Huifeng Guo, Bo Chen, Xiu Li, and Ruiming Tang. 2022. Disentangling Past-Future Modeling in Sequential Recommendation via Dual Networks. In *CIKM*. ACM, 2549–2558.

[46] Weifeng Zhang, Jingwen Mao, Yi Cao, and Congfu Xu. 2020. Multiplex Graph Neural Networks for Multi-behavior Recommendation. In *CIKM*. ACM, 2313–2316.

[47] Zhe Zhao, Zhiyuan Cheng, Lichan Hong, and Ed Huai-hsin Chi. 2015. Improving User Topic Interest Profiles by Behavior Factorization. In *WWW*. ACM, 1406–1416.

[48] Guorui Zhou, Na Mou, Ying Fan, Qi Pi, Weijie Bian, Chang Zhou, Xiaoqiang Zhu, and Kun Gai. 2019. Deep Interest Evolution Network for Click-Through Rate Prediction. In *AAAI*. AAAI Press, 5941–5948.

[49] Guorui Zhou, Xiaoqiang Zhu, Chengru Song, Ying Fan, Han Zhu, Xiao Ma, Yanghui Yan, Junqi Jin, Han Li, and Kun Gai. 2018. Deep Interest Network for Click-Through Rate Prediction. In *KDD*. ACM, 1059–1068.

[50] Xi Zhu, Fake Lin, Ziwei Zhao, Tong Xu, Xiangyu Zhao, Zikai Yin, Xueying Li, and Enhong Chen. 2025. Multi-Behavior Recommendation with Personalized Directed Acyclic Behavior Graphs. *ACM Trans. Inf. Syst.* 43, 1 (2025), 20:1–20:30.