VIETNAM NATIONAL UNIVERSITY

UNIVERSITY OF ENGINEERING AND TECHNOLOGY

NGUYỄN TRẦN NGỌC LINH

ROBUST AND ADAPTIVE RECOMMENDATION BY DEEP MODELING OF CANONICAL AND AUXILIARY DATA

MAJOR: INFORMATION SYSTEMS

MAJOR CODE: 9480104

SUMMARY OF PhD DISSERTATION INFORMATION SYSTEM

HA NOI - 2026

The dissertation was completed at: University of Engineering and Technology, Vietnam National University, Hanoi

Supervisors: Assoc.Prof. Phan Xuan Hieu

Co-supervisors: Dr. Tran Mai Vu

Reviewer 1: Dr. Do Thanh Ha

Reviewer 2: Assoc.Prof. Bui Thu Lam

Reviewer 3: Assoc.Prof. Do Van Thanh

The dissertation will be defended before the Vietnam National University-level PhD Thesis Evaluation Committee at Hanoi University Of Engineering and Technology, VNU

at 8h30 A.M. Thursday, 18th June 2026

PhD **SUPERVISORS**

CONFIRMATION FROM THE TRAINING INSTITUTION

The dissertation can be found at:

- National Library of Vietnam
- Information Center - Library, Vietnam National University, Hanoi

Abstract

Modern recommendation systems are increasingly required to operate in large-scale digital ecosystems where user bases grow continuously, item spaces expand rapidly, and interaction patterns evolve in unpredictable ways.

This dissertation develops deep learning methods for robust and adaptive recommendation through deep modeling of both canonical data (user-item interactions) and auxiliary data (side information, domain context, and conversational signals). The research addresses three fundamental challenges across four interconnected directions.

First, the dissertation investigates scalable recommendation through ID-free user representations, neural soft clustering, and contrastive learning that helps maintaining recommendation quality at the scale of real-world recommendation.

Second, the research explores robust fusion of canonical and auxiliary data through attention-based weight generation mechanisms and masked graph contrastive learning. These techniques dynamically balance the contribution of behavioral embeddings and side information.

Third, the dissertation develops continual learning mechanisms for adaptive multi-domain recommendation.

Fourth, the research proposes hybrid conversational recommendation that bridges canonical and auxiliary data through graph neural network-based preference modeling integrated with large language models and retrieval-augmented generation.

Extensive experiments on benchmark datasets and industrial deployment with real products from Viettel (one of the largest corporations in Vietnam), validating the effectiveness of the proposed models, demonstrating consistent improvements over state-of-the-art baselines.

Keywords: *Deep learning, recommendation systems, canonical data, auxiliary data, data sparsity, cold-start problem, scalability, soft clustering, contrastive learning, graph neural networks, side information fusion, continual learning, multi-domain recommendation, conversational recommendation, large language models, retrieval-augmented generation.*

Chapter 1

Literature Review of Background and Methods

1.1 Problem Definition and Formulation

1.1.1 Overview of Recommendation Problem

Problem Statement

Formally, a recommender system operates over three fundamental entities: a set of users $\mathcal{U} = \{u_1, u_2, \ldots, u_M\}$, a set of items $\mathcal{V} = \{v_1, v_2, \ldots, v_N\}$, and an interaction matrix $R \in \mathbb{R}^{M \times N}$ between them . The core of any recommendation method is a utility function that quantifies the relevance of an item to a user:

$$f : \mathcal{U} \times \mathcal{V} \to \mathcal{D} \tag{1.1}$$

where $\mathcal{D}$ represents the domain of utility values.

The primary objective of a recommender system is to identify, for each user $u \in \mathcal{U}$, the items that maximize the utility function:

$$v_u^* = \arg\max_{v \in \mathcal{V}} f(u, v) \tag{1.2}$$

In practice, systems typically generate a ranked list of top-K items rather than a single recommendation, producing an ordered set $\mathcal{D}_u = \{v_1, v_2, \ldots, v_K\}$ where items are sorted in descending order by their predicted utility scores. The recommendation task thus becomes:

$$\mathcal{D}_u = \text{Top-}K_{v \in \mathcal{V}} \left(f(u, v) \right) \tag{1.3}$$

Interaction History Representation

User item interactions history are conventionally represented as a rating matrix $\mathbf{R} \in \mathbb{R}^{M \times N}$, where $M = |\mathcal{U}|$ denotes the number of users and $N = |\mathcal{V}|$ denotes the number of items. Each entry $r_{u,v}$ in this matrix represents the observed interaction between user u and item v.

For explicit feedback systems, $r_{u,v}$ captures a user's rating that directly reflects their preference, such as a 5 star rating on a movie or a numerical score for a product. For implicit feedback systems, $r_{u,v}$ encodes behavioral signals such as clicks, views, or purchases which indirectly indicate user interest without explicit preference statements.

A critical characteristic of the rating matrix is that only a small fraction of entries are observed. Let $\Omega \subseteq \mathcal{U} \times \mathcal{V}$ denote the set of observed user item pairs. The recommendation task can then be formulated as matrix completion: given the partially observed matrix $\mathbf{R}_\Omega$, predict the missing entries $\mathbf{R}_{\bar{\Omega}}$ where $\bar{\Omega} = (\mathcal{U} \times \mathcal{V}) \setminus \Omega$ represents unobserved interactions.

$$\hat{\mathbf{R}} = \mathcal{F}(\mathbf{R}_\Omega; \Theta) \tag{1.4}$$

where $\mathcal{F}$ represents the recommendation model parameterized by Θ, and $\hat{\mathbf{R}}$ denotes the predicted complete rating matrix.

The density of the rating matrix, defined as $|\Omega|/(M \times N)$, is typically extremely low in real-world systems often less than 1% for large scale platforms. This extreme sparsity fundamentally shapes the design of recommendation algorithms and motivates many of the advanced techniques discussed in this dissertation.

1.1.2 Research Scope and Objectives

This dissertation develops deep learning based solutions that address the interconnected challenges of sparsity, cold-start, and scalability. The proposed approaches share a common objective: learning robust, transferable representations that capture meaningful user preferences without requiring exhaustive interaction histories or expensive computational resources.

Chapter 2

Robust Recommendation via Interaction Embedding and Soft Clustering

2.1 Introduction

Modern recommender systems have become essential components of web scale applications, supporting millions of users across ecommerce platforms, streaming services, and social networks. Despite significant advances in deep learning based recommendation approaches, several fundamental challenges persist that limit the practical deployment and scalability of these systems. This section identifies the key scalability challenges in modern recommender systems and uses them as the primary motivation for the subsequent analysis in this chapter.

2.2 EfficientRec: Scalable ID-Free Recommendation via Soft Clustering and Contrastive Learning

To address the identified gaps, this chapter proposes EfficientRec, a novel framework for robust large scale recommendation via interaction embedding and soft clustering. The research objectives guiding the development of EfficientRec are as follows.

This section presents the proposed EfficientRec architecture, a scalable recommendation framework designed to address the fundamental limitations of conventional user-ID based recommendation models. The architecture eliminates the dependency on explicit user-IDentifiers by constructing user representations dynamically from behavioral signals, thereby achieving computational complexity that is independent of the user population size. This design enables the system to scale to large user bases while main-

taining consistent recommendation quality and supporting seamless integration of new users without requiring model retraining.

The proposed model consists of three principal components that work together to provide personalized recommendations.

The overall architecture is illustrated in Figure 2.1, which provides a high level view of how the three components interact to produce personalized recommendations.

The first component is the "Interaction Embedding model", which is responsible for constructing compact and informative user representations by aggregating information from the user's historical interactions with items in the system.

The second component is the "Clustering Model", which organizes users into preference aware groups using contrastive learning and soft clustering techniques. The clustering model learns to map user representations into a latent preference space where each dimension corresponds to a distinct preference cluster. The contrastive learning objective ensures that users with similar preferences are mapped to similar regions in the preference space, while users with different preferences are separated well.

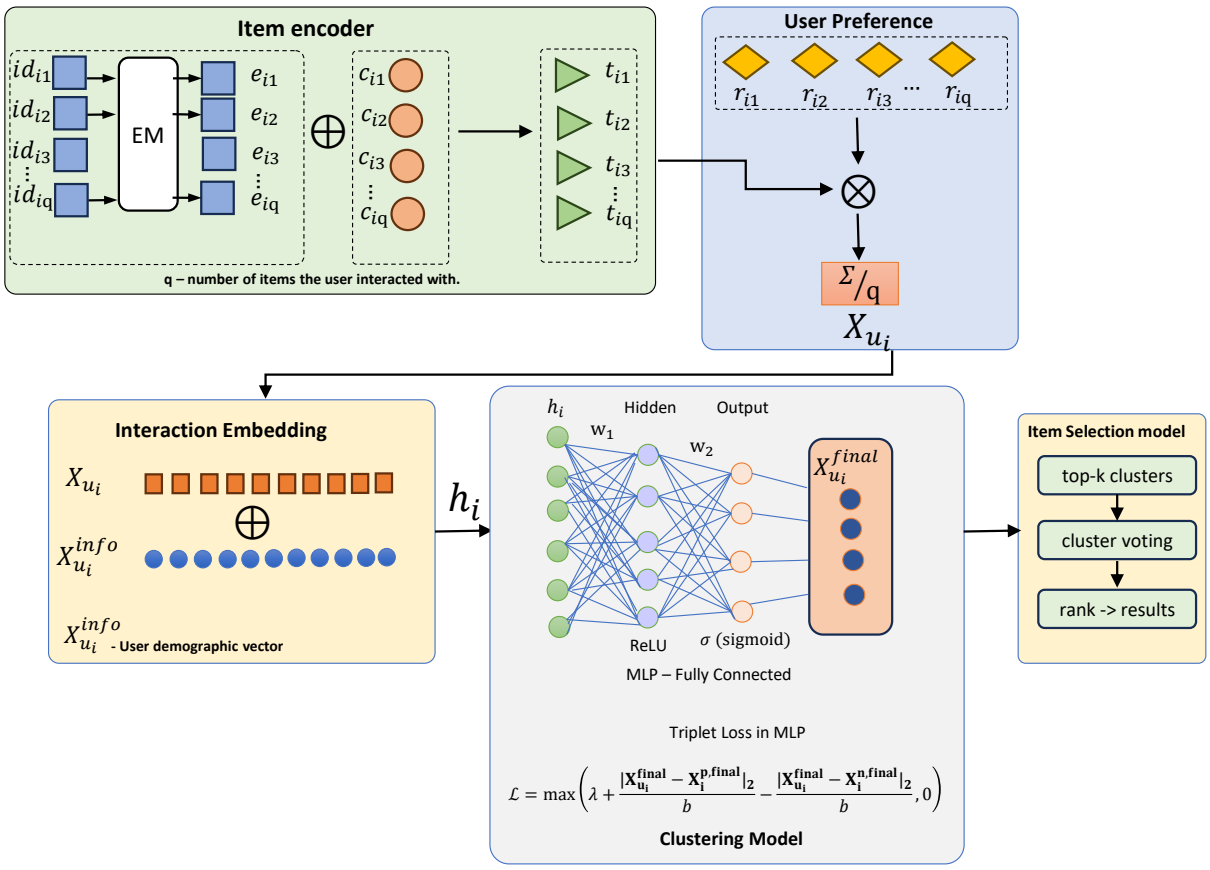

Figure 2.1: EfficientRec Overall Architecture

The third component is the "Item Selection model", which efficiently generates

personalized recommendations through a two phases cluster based voting mechanism. In the offline phase, the model precomputed preference scores for each cluster and item pair based on the aggregated ratings from all users belonging to that cluster. In the online phase, the model computes the target user's cluster membership and generates recommendations by aggregating the precomputed scores from the user's most relevant clusters. This two phases design significantly reduces the computational cost of recommendation generation compared to methods that must score all items for each user.

2.3 Experimental Results

Table 2.1: Overall Performance Comparison (All Users) @30, reported as mean ± standard deviation over 5 random seeds.

Model	Recall@30	NDCG@30	Category
Proposed Method			
EfficientRec	$\mathbf{0.1994 \pm 0.0028}^{\dagger}$	$\mathbf{0.1178 \pm 0.0027}^{\text{ns}}$	Proposed
Graph-based Methods			
NGCF	$\underline{0.1958 \pm 0.0013}$	$\underline{0.1174 \pm 0.0007}$	Graph
LINKX	0.1928 ± 0.0022	0.1160 ± 0.0012	Graph
GraphSAGE	0.1658 ± 0.0006	0.0970 ± 0.0004	Graph
GAT	0.1639 ± 0.0001	0.0946 ± 0.0001	Graph
LightGCN	0.1636 ± 0.0003	0.0945 ± 0.0001	Graph
Contrastive Learning Methods			
SSL4Rec	0.1644 ± 0.0001	0.0950 ± 0.0012	SSL
MixGCF	0.1591 ± 0.0001	0.0910 ± 0.0001	CL
SGL	0.0192 ± 0.0015	0.0565 ± 0.0011	CL
Clustering-based and MF Methods			
SCoC	0.1584 ± 0.0001	0.1107 ± 0.0004	Clustering
FCCF	0.1090 ± 0.0001	0.1072 ± 0.0012	Clustering
BPR-MF	0.1074 ± 0.0003	0.0912 ± 0.0001	MF
FCM-Rec	0.1067 ± 0.0001	0.1068 ± 0.0001	Clustering

Bold indicates the best result and underline the second best.

Table 2.2: Results of the Online Experiments on TV360

Methods	**TV360 Films**		**TV360 Videos**	
	ACPU	ADPU	ACPU	ADPU
2DNNs	0.0152	13.896	0.0322	10.526
ALS	0.0121	12.190	0.0160	4.277
ER Interaction Split	**0.0186**	**15.018**	**0.0421**	<u>12.290</u>
ER User Group Split	<u>0.0176</u>	<u>14.272</u>	<u>0.0381</u>	**13.155**

Bold = Best, <u>Underline</u> = Second Best.

2.4 Chapter Summary

This chapter introduced EfficientRec, a scalable, ID-free recommendation framework that replaces per-user embeddings with behavior-driven representations learned from interaction subsets through deep interaction modeling, soft clustering, and contrastive learning. By computing each user vector on demand rather than storing it, the design lowers the user-side memory footprint from $O(M \cdot d)$ to $O(K \cdot d)$, scales sublinearly with the user population, and extends naturally to new and low-interaction users without retraining. Offline experiments on public benchmarks together with online A/B testing on a production streaming platform confirm that EfficientRec matches or exceeds strong graph-based baselines while substantially reducing resource cost.

These observations point to several natural extensions. Learning the number of clusters, or updating clusters incrementally as new interactions stream in, would let the model track a shifting population more closely, while incorporating richer multimodal side information would give a stronger prior precisely where behavioral evidence is thinnest. Finally, compressing the model through quantization, pruning, and distillation would bring inference within the latency and memory budgets needed for on-device deployment.

Chapter 3

Boosting Recommendation via Graph-based Fusion of Canonical Interactions and Auxiliary Side Information

3.1 Introduction

This chapter presents how side information, interaction patterns, and graph learning can be cohesively combined to improve modern recommendation performance. Through detailed architectural modelling and empirical analysis, this chapter demonstrates that augmenting GNN-based recommendation with structured side information fusion produces more accurate, robust, and semantically aligned recommendation outcomes. These works have been published in peer-reviewed conferences, including:[P2] "GIFT4Rec: An Effective Side Information Fusion Technique Apply to Graph Neural Network for Cold-Start Recommendation" (ACIIDS 2023), and [P3] "The Masked Simple Graph Contrastive Learning for Recommendation" (KSE 2024).

3.2 GIFT4Rec: Auxiliary Information Fusion with Attention-based and Meta-Learning Techniques for Cold-Start Recommendation

3.2.1 Gift4Rec: Model Architecture and Components

This section presents the proposed GIFT4Rec architecture, a unified framework for integrating side information into graph-based recommendation systems. The architecture addresses the fundamental challenge of balancing behavioral signals from user item interactions with semantic information from user attributes, enabling robust recommendation across both warm-start and cold-start scenarios.

The overall architecture is illustrated in Figure 3.1, which provides a high-level view of how the three components interact together for producing personalized recommendations.

The first component is the “GNN Interaction Module”, which learns user and item representations by propagating information through the user item interaction graph. Unlike content-based approaches that rely solely on feature matching, this component captures collaborative signals from the global interaction structure, enabling discovery of preference patterns that emerge from collective user behavior.

The second component is the "Local Side Information Fusion Module" (LSIF), which adaptively combines behavioral embeddings with side information embeddings for each individual user. The key insight is that the optimal fusion strategy varies across users some users have rich interaction histories that provide strong preference signals, while others have limited interactions where side information becomes more valuable. This component learns personalized fusion weights through an attention-based mechanism called Attention DropoutNet (ADN).

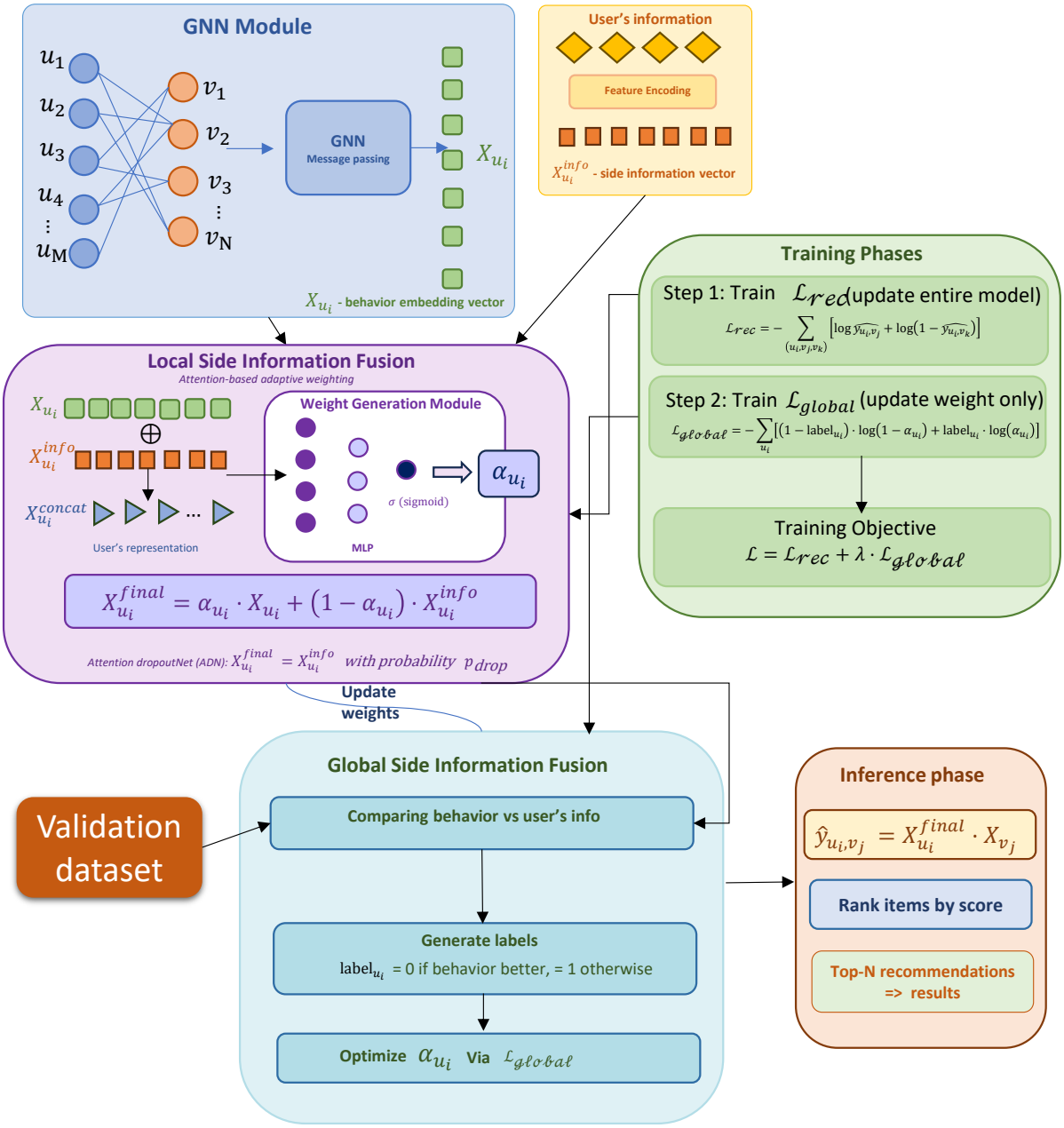

Figure 3.1: GIFT4Rec overall architecture

The third component is the "Global Side Information Fusion" (GSIF) module, which provides meta-level supervision for the weight generation process by evaluating which information source (behavioral or side information) supports better recommendation performance on validation data. This global perspective ensures that the learned fusion weights generalize well beyond the training interactions.

3.2.2 Experimental Results

a) Overall Performance

Table 3.1: GIFT4Rec – Overall Performance Comparison (All Users) @30, reported as mean ± standard deviation over 5 random seeds.

Model	Recall@30	NDCG@30	Category
GIFT4Rec	$\mathbf{0.2162 \pm 0.0001}^{\dagger}$	$\mathbf{0.1263 \pm 0.0001}^{\dagger}$	Proposed
KGAT	0.1793 ± 0.0008	0.1067 ± 0.0007	Aux+KG
LINKX	0.1928 ± 0.0022	0.1160 ± 0.0012	Aux+Graph
GAT	0.1639 ± 0.0001	0.0946 ± 0.0001	Aux+Attn
KGAT DropoutNet	0.0908 ± 0.0003	0.0359 ± 0.0008	Aux+KG
NGCF	0.1958 ± 0.0013	$\underline{0.1178 \pm 0.0007}$	GNN
LightGCN	0.1636 ± 0.0003	0.0945 ± 0.0001	GNN
SSL4Rec	0.1644 ± 0.0001	0.0950 ± 0.0012	SSL
EfficientRec (Ch.2)	$\underline{0.1994 \pm 0.0028}$	0.1174 ± 0.0027	Proposed

Bold = best, underline = second best.

3.3 The Masked Simple Graph Contrastive Learning for Recommendation

3.3.1 MaskSimGCL: Model Architecture and Components

This section presents the proposed MaskSimGCL (Masked Simple Graph Contrastive Learning) architecture, a novel framework designed to address the limitations of existing graph-based contrastive learning methods for recommendation. The architecture extends the SimGCL framework by integrating learnable masking mechanisms that adaptively identify and weight the importance of different embedding dimensions, thereby achieving more robust representation learning under sparse data conditions.

The proposed model consists of four principal components that operate together to deliver personalized recommendations. The first component is the graph neural network backbone, which is responsible for learning user and item representations through message passing operations on the user item bipartite graph. Following the LightGCN design, this component employs simplified graph convolutions that propagate collaborative signals without feature transformation, capturing neighborhood patterns through layer-wise aggregation.

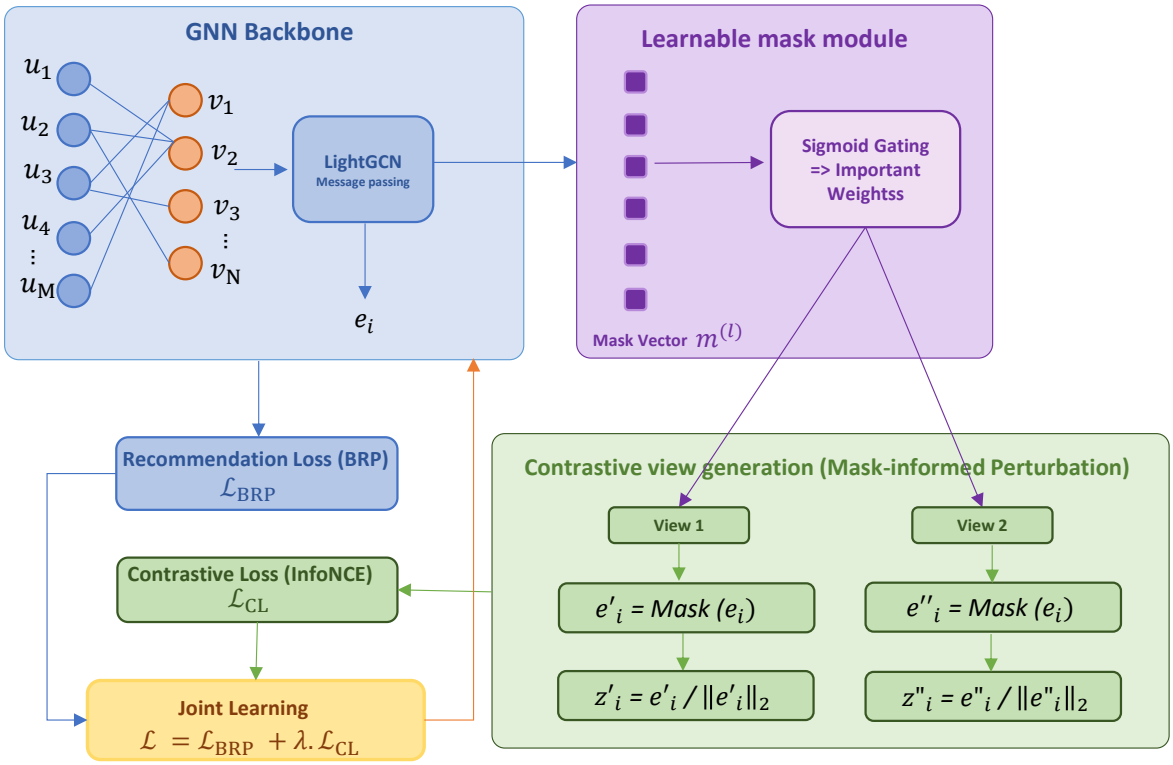

Figure 3.2: MaskSimGCL overall architecture

The second component is the learnable mask module, which introduces trainable mask vectors at each graph neural network layer. These masks serve as importance filters that adaptively weight each dimension of the node embeddings-based on their relevance to the recommendation task. By focusing model capacity on informative parameters while suppressing noisy or redundant dimensions, this component provides implicit regularization that effectively mitigates overfitting in sparse data environments.

The third component is the contrastive view Generation module, which constructs augmented representations for contrastive learning. Unlike SimGCL's uniform noise injection, MaskSimGCL employs mask informed perturbations that apply differential noise magnitudes-based on the learned importance scores. Dimensions identified as less important receive larger perturbations, while critical dimensions are preserved with smaller noise, resulting in consistent contrastive views that enhance representation learning.

The fourth component is the joint optimization framework, which combines the supervised recommendation objective with the self-supervised contrastive learning objective. This multitask learning formulation enables the model to simultaneously optimize for accurating preference prediction and robust representation learning, with the contrastive loss providing uniformity regularization that promotes more evenly distributed

embeddings in the representation space.

3.3.2 Experimental Results

Overall Performance

Table 3.2: MaskSimGCL - Overall Performance Comparison (All Users) @30, reported as mean ± standard deviation over 5 random seeds.

Model	Recall@30	NDCG@30	Category
MaskSimGCL	$\mathbf{0.2404 \pm 0.0008}^{\dagger}$	$\mathbf{0.1322 \pm 0.0004}^{\dagger}$	Proposed
XSimGCL	$\underline{0.2301 \pm 0.0004}$	0.1250 ± 0.0002	GCL
SimGCL	0.2292 ± 0.0012	0.1249 ± 0.0008	GCL
LightGCL	0.2128 ± 0.0003	0.1103 ± 0.0001	GCL
DirectAU	0.2110 ± 0.0009	0.1226 ± 0.0003	GCL
SGL	0.0192 ± 0.0015	0.0565 ± 0.0011	GCL
SSL4Rec	0.1644 ± 0.0001	0.0950 ± 0.0012	GNN
LightGCN	0.1636 ± 0.0003	0.0945 ± 0.0001	GNN
GIFT4Rec (Section 3.2)	0.2162 ± 0.0001	$\underline{0.1263 \pm 0.0001}$	Proposed
EfficientRec (Chapter 2)	0.1994 ± 0.0028	0.1174 ± 0.0027	Proposed

Bold = best, underline = second best.

3.4 Chapter Summary

This chapter addressed sparsity and cold-start through two complementary contributions. GIFT4Rec performs adaptive side-information fusion, using an attention-based weight-generation module to compute user-specific fusion weights and to combine behavioural embeddings with auxiliary signals through local and global modules optimized under meta-learning principles. MaskSimGCL complements it by strengthening representation robustness with masked graph contrastive learning. Together they yield consistent gains over graph-based baselines in both warm-start and cold-start settings, showing that principled fusion and contrastive regularization improve recommendation under data scarcity.

Chapter 4

Enhancing Multi-Domain Recommendation with Continual Learning

4.1 Introduction

Modern recommendation systems increasingly operate across multiple service domains within unified platforms, where users interact with heterogeneous content categories such as e-commerce products, video streaming, music services, and news feeds. These multi-domain environments present unique challenges that extend beyond traditional single domain recommendation paradigms. While cross domain recommendation (CDR) has emerged as a promising direction to leverage rich information across domains, existing approaches predominantly focus on improving target domain performance while often neglecting the preservation of source domain knowledge and the fairness of performance across all participating domains.

4.1.1 Model Architecture and Components

The CNL4Rec framework introduces a task masking based continual learning mechanism that operates at the embedding level. The central design principle is to treat each domain as a sequential learning task and apply domain specific masks that identify and protect important latent dimensions for each domain.

The architecture introduces learnable mask vectors for each domain that have the same dimensionality as the user and item embeddings. These masks identify which latent dimensions are essential for representing domain specific behavioral patterns, enabling the system to selectively update only relevant parameters during training. Parameters

deemed unimportant for the current domain remain protected to maintain performance on previously learned domains.

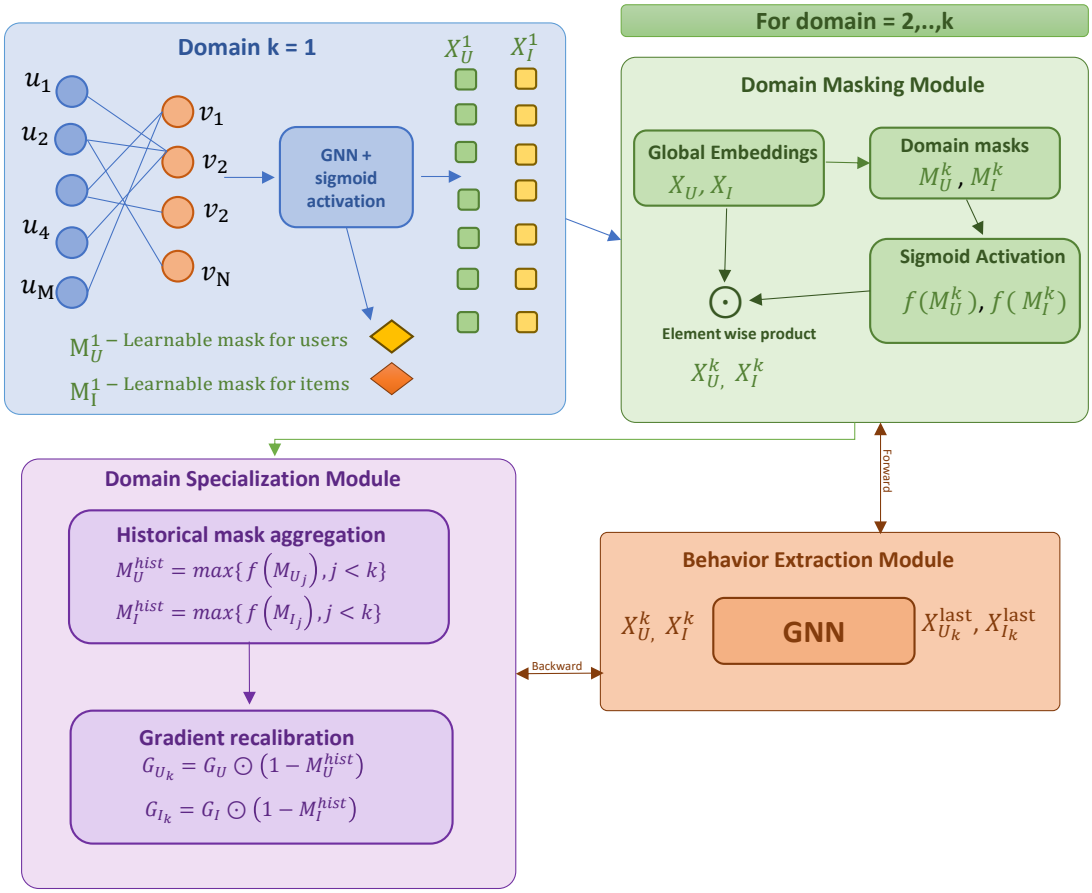

Figure 4.1: CNL4Rec Overall Architecture

As illustrated in Figure 4.1, CNL4Rec addresses the continual multi-domain recommendation problem through three tightly integrated modules. The "Domain Masking Module" learns domain-specific binary masks over the shared user and item embedding matrices, identifying which latent dimensions are relevant for each domain and producing filtered domain-specific representations $\mathbf{X}_U^k$ and $\mathbf{X}_I^k$. Its role is to partition the shared embedding space into domain-specific subspaces without requiring separate parameters per domain. The "Domain Specialization Module" protects previously acquired knowledge by aggregating the masks of all past domains into a historical mask $\mathrm{M}^{\mathrm{hist}}$, which is then used to zero out gradients flowing into dimensions already claimed by prior domains during backpropagation. This selective gradient suppression is the core anti-forgetting mechanism, ensuring that learning a new domain does not overwrite representations essential to earlier ones. The "Behavior Extraction Module" receives the masked, domain-specific embeddings and passes them through a GNN backbone that aggregates high-order collaborative signals within the current domain, producing final embeddings

used for preference scoring and recommendation ranking. Together, the three modules implement domain knowledge isolation at the embedding level, knowledge protection at the gradient level, and preference learning at the inference level collectively resolving the stability plasticity trade off inherent in continual multi-domain recommendation.

4.1.2 Experimental Results

A. Overall results

a. Results on MovieLens-1M

Table 4.1: Performance Comparison on MovieLens-1M (Recall@30). CV(%)=coefficient of variation across genres (lower is more balanced); Min=worst-domain Recall@30 (higher is better). **Bold**=Best, Underline=Second Best.

Method	Action	Comedy	Sci-Fi	Thriller	Drama	Mean	Std	CV(%)	Min
CNL4Rec	**0.2481**	**0.2583**	**0.1865**	**0.2349**	0.2163	**0.2288**	0.0254	**11.1**	**0.1865**
CTNet	0.1600	0.0838	0.1342	0.2034	**0.2268**	0.1616	0.0506	31.3	0.0838
DIIT	0.1246	0.0632	0.1106	0.1181	0.1061	0.1045	0.0216	20.7	0.0632
KEEP	0.0598	0.0060	0.0841	0.1005	0.1796	0.0860	0.0567	65.9	0.0060
MF	0.0744	0.0060	0.0286	0.0485	0.2061	0.0727	0.0704	96.8	0.0060
ECAT	0.0063	0.0273	0.0173	0.0744	0.1665	0.0583	0.0588	100.8	0.0063

b. Results on Yelp

Table 4.2: Performance Comparison on Yelp (Recall@30). CV(%)=coefficient of variation across categories (lower is more balanced); Min=worst-domain Recall@30 (higher is better). **Bold**=Best, Underline=Second Best.

Method	Restaurant	Shopping	Food	Beauty	Health	Mean	Std	CV(%)	Min
CNL4Rec	**0.0194**	0.0150	0.0184	0.0166	**0.0182**	**0.0174**	0.0015	8.8	0.0150
ECAT	0.0164	0.0172	0.0162	**0.0172**	0.0154	0.0165	0.0007	**4.1**	0.0154
CTNet	0.0146	0.0152	**0.0192**	0.0148	0.0179	0.0164	0.0019	11.4	0.0146
KEEP	0.0158	**0.0182**	0.0156	0.0156	0.0164	0.0165	0.0010	6.0	**0.0156**
MF	0.0150	0.0158	0.0179	0.0162	0.0156	0.0163	0.0010	6.1	0.0150
DIIT	0.0138	0.0154	0.0150	0.0156	0.0178	0.0155	0.0013	8.4	0.0138

c. Results on Amazon

Table 4.3: Performance Comparison on Amazon (Recall@30). CV(%)=coefficient of variation across categories (lower is more balanced); Min=worst-domain Recall@30 (higher is better). **Bold**=Best, Underline=Second Best.

Method	Electronics	Books	Movies	Home	Sports	Mean	Std	CV(%)	Min
CNL4Rec	0.0139	0.0125	0.0135	**0.0133**	0.0126	**0.0132**	0.0005	**4.1**	**0.0125**
ECAT	**0.0147**	0.0130	0.0128	0.0132	0.0116	0.0131	0.0010	7.6	0.0116
MF	0.0129	0.0109	**0.0170**	0.0125	0.0124	0.0131	0.0020	15.6	0.0109
KEEP	0.0114	**0.0145**	0.0142	0.0118	0.0127	0.0129	0.0012	9.6	0.0114
CTNet	0.0115	0.0120	0.0137	0.0113	**0.0136**	0.0124	0.0010	8.3	0.0113
DIIT	0.0134	0.0135	0.0105	0.0103	0.0116	0.0119	0.0014	11.6	0.0103

4.2 Chapter Summary

This chapter presented CNL4Rec, a continual-learning framework for adaptive multi-domain recommendation. Domain masking identifies the parameters important to each domain, while domain specialization modulates gradient updates so that knowledge from previously seen domains is preserved as new domains are learned, mitigating catastrophic forgetting. A fairness objective further encourages balanced performance across domains rather than optimizing a single target. Across multiple multi-domain benchmarks, CNL4Rec attains the best mean and worst-domain performance with low cross-domain variance, outperforming representative cross-domain baselines.

The approach rests on assumptions that also delimit its scope. It works best when domain boundaries are reasonably well defined, and is therefore less suited to highly overlapping or rapidly evolving domains; its accuracy is also sensitive to the order in which domains arrive, degrading when they are presented from smallest to largest. Because the accumulated union of importance masks grows with each new domain, memory and complexity rise as domains continue to accumulate.

Chapter 5

Conversational Recommendation with a GNN and RAG-Based Hybrid System

5.1 Introduction

This chapter proposes a film recommendation chatbot that combines the understanding of past behaviors from historical interaction data studied through a graph based deep learning model and real time preferences acquired from large language models (LLMs) via ensemble learning. This framework accentuates the complementary strengths of both advanced techniques to provide meaningful recommendations to users. In particular, graph based deep learning excels in deploying interactions into graph structured data, thus uncovering latent patterns and complex relationships between node entities. The graph based model also uses scalable algorithms and distributed computing techniques to handle vast volumes of interaction data, ensuring the system remains responsive and accurate as user and content data grow. On the other hand, LLMs collect individual preferences through conversational sessions, process natural language input, and flexibly adapt to immediate requirements. By combining these components, the system achieves a comprehensive, robust, and adaptive recommendation approach that balances past behavior insights with current user context, optimized for real time performance and large scale data processing.

This work was published in “Improving Retrieval-Augmented Generation for Scalable Movie Chatbots via Graph Based Recommendation Models” 2025 [P5].

5.2 CG-RAG: Conversational Recommendation via Graph-Enhanced Retrieval-Augmented Generation

5.2.1 Overall Architecture of CG-RAG

The recommendation generator integrates dialogue data with both contextual understanding and user behavior analysis to provide relevant movie suggestions. Our method comprises three main components: the conversational engine, the recommendation engine, and the feature matching and retrieval layer.

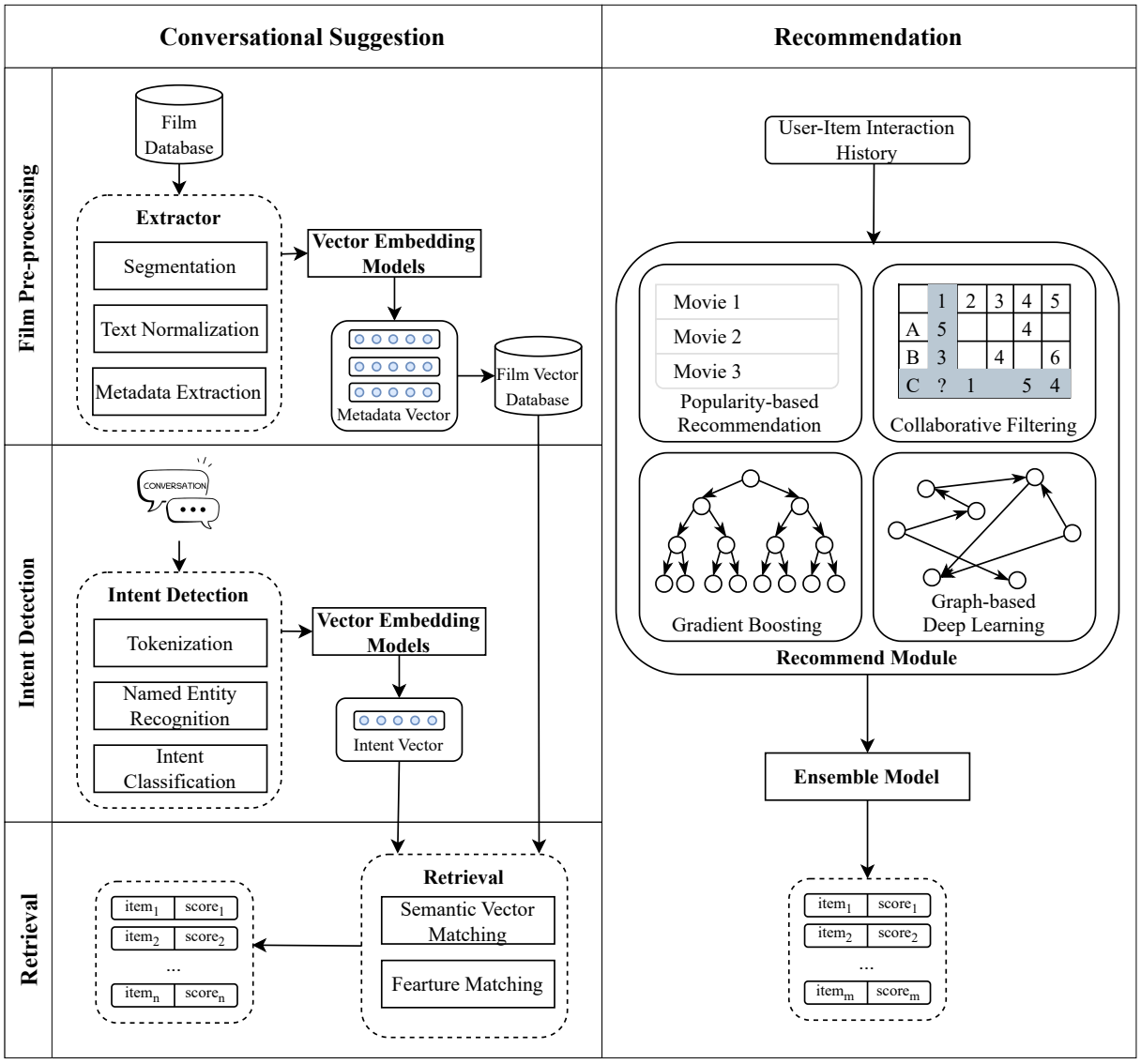

Figure 5.1: Architecture of the conversational suggestion and recommendation generator, comprising a conversational engine (intent detection producing an intent vector), a graph-based recommendation engine, and a feature-matching and retrieval layer, combined through a shared fusion layer to generate the final conversational recommendations.

Figure 5.1 illustrates the overall architecture of the proposed hybrid conversational recommendation system, which is organized into two parallel branches converging at a shared fusion layer. The *conversational engine* processes the user's natural language query through three sequential stages: a pre-processing stage that normalizes the raw query and simultaneously extracts a Film Vector and a Metadata Vector from the film database, encoding item content and structural attributes respectively; an Intent Detection stage that applies Named Entity Recognition and vector embedding to identify domain-specific entities (e.g., genres, actors) and classify the user's intent into a compact Intent Vector; and a Retrieval stage that computes semantic similarity between the Intent Vector and item vectors to produce a ranked list of semantically relevant candidates. The *recommendation engine* operates independently on historical user-item interaction data and is built upon a systematic empirical evaluation of representative deep recommendation methods. Specifically, we benchmark twelve models spanning graph propagation, graph contrastive learning, and embedding-alignment/self-supervised approaches. The behavioral ranking score derived from historical user-item interaction patterns is independently fused with the retrieval score produced by the conversational retrieval engine, which grounds candidate items in the current dialogue context via RAG. Finally, the *feature matching and retrieval layer* serves as the fusion point of the entire system, aligning the candidate lists from both branches by item identity and jointly re-scoring each item based on its semantic relevance to the expressed user intent and its predicted preference strength from behavioral history, thereby producing a final recommendation list that is simultaneously context-aware and personalized.

5.3 Experimental Results

Table 5.1: Performance comparison between baseline methods on Movielens-1M and TV360 datasets. R@K refers to Recall top-K, and P@K represents Precision top-K. The best result for each dataset is highlighted in bold, while the second best is determined by underline.

Dataset	Movielens-1M					TV360				
Metric	R@30	P@30	R@50	P@50	Time	R@30	P@30	R@50	P@50	Time
LightGCN	0.1628	0.1040	0.2373	0.0964	1e-6	0.0423	0.1018	0.0701	0.1184	1e-6
GAT	0.1650	0.0564	0.2395	0.0515	1e-5	0.0323	0.1067	0.0529	**0.1406**	1e-5
PMLP	0.1292	0.0467	0.1920	0.0430	1e-6	0.0127	0.0433	0.0223	0.0449	1e-5
GraphSAGE	0.1482	0.0526	0.2152	0.0482	1e-6	0.0400	0.1246	0.0666	0.1251	1e-6
LINKX	0.1184	0.0406	0.1672	0.0371	1e-6	0.0376	0.1212	0.0609	0.1176	1e-6
MixGCF	0.0192	0.0208	0.0432	0.0256	1e-5	0.0369	0.0269	0.0550	0.0248	1e-5
SGL	0.0192	0.0208	0.0295	0.0196	1e-6	0.0681	0.0449	0.0933	0.0385	1e-6
SimGCL	0.2125	0.1467	0.2872	0.1224	1e-6	0.1038	0.0672	0.1467	0.0593	1e-5
XSimGCL	0.2209	0.1494	0.2905	0.1267	1e-6	0.1725	0.1026	0.2362	0.0885	1e-6
NCL	**0.2497**	**0.1990**	**0.3351**	**0.1685**	1e-6	0.0915	0.0584	0.1296	0.0514	1e-6
SSL4Rec	0.1933	0.1208	0.2713	0.1068	1e-5	0.1082	0.0730	0.1644	0.0678	1e-5
DirectAU	0.1315	0.0835	0.1820	0.0723	1e-5	**0.1932**	**0.1250**	**0.2714**	0.1095	1e-5

Bold = Best, Underline = Second Best

Conversational Retrieval Engine Result

Table 5.2: Recall@K scores in Movielens-1M dataset across 3 LLMs. The results are measured in various K values from 10 to 200, accompanied by answer generation time in seconds. The generated response outputs the top film provided by the chatbot modules.

	Recall@10	**Recall@20**	**Recall@30**	**Recall@50**	**Recall@100**	**Recall@200**	**Time (s)**
Gemma-2B	0.186	0.201	0.214	0.259	0.401	0.829	2.05
LLaMA-3B	0.204	0.227	0.239	0.291	0.428	0.869	2.28
Qwen-3B	**0.207**	**0.232**	**0.246**	**0.299**	**0.435**	**0.888**	**2.42**

Bold = Best, Underline = Second Best

Table 5.3: Recall@K scores on TV360 dataset across 3 LLMs. The results are measured in various K values from 10 to 200, accompanied by answer generation time in seconds. The generated response outputs the top film provided by both modules.

	Recall@10	**Recall@20**	**Recall@30**	**Recall@50**	**Recall@100**	**Recall@200**	**Time (s)**
Gemma-2B	0.060	0.071	0.084	0.106	0.221	0.703	1.88
LLaMA-3B	0.072	**0.086**	0.093	0.112	0.226	0.731	2.13
Qwen-3B	**0.073**	0.083	**0.095**	**0.118**	**0.233**	**0.740**	**2.18**

Bold = Best, Underline = Second Best

Ensemble Engine Result

Table 5.4: Performance comparison between baseline recommendation methods combined with conversational engine on Movielens-1M and TV360 datasets. R@K refers to Recall top-K and P@K represents Precision top-K

Dataset	Movielens-1M				TV360			
Metric	R@10	R@20	R@30	R@50	R@10	R@20	R@30	R@50
XSimGCL	<u>0.2046</u>	<u>0.2353</u>	<u>0.2503</u>	<u>0.3012</u>	**0.1258**	**0.1532**	**0.1812**	**0.2681**
NCL	**0.2104**	**0.2387**	**0.2524**	**0.3402**	0.0657	0.0884	0.1021	0.1458
DirectAU	0.1365	0.1581	0.1872	0.2447	<u>0.1106</u>	<u>0.1582</u>	<u>0.1808</u>	<u>0.2606</u>

Bold = Best, <u>Underline</u> = Second Best

5.4 Chapter Summary

This chapter introduced CG-RAG, a hybrid conversational recommendation architecture that bridges long-term user behavior with real-time conversational intent. A graph neural network models structured long-term preferences, a large language model handles intent extraction and natural-language dialogue, and a hybrid retrieval stage combining sparse (BM25) and dense retrieval grounds recommendations in the actual item catalog. Because candidate items are constrained to retrieved catalog entries and the generator is conditioned on retrieved metadata, the system produces context-aware, grounded recommendations through interactive dialogue rather than static ranked lists.

A few limitations should be acknowledged. The evaluation rests largely on conversations that were drafted automatically and then revised by human annotators, rather than collected from real users at scale, and the language-model component introduces inference latency that may constrain real-time deployment. The dissertation also stops short of reporting dedicated quantitative metrics for hallucination and cross-turn consistency.

Each of these gaps suggests a way forward. Evaluating the system on genuine user conversations, and benchmarking it against LLM-based recommendation while reporting computational cost alongside accuracy, would place its performance on firmer ground. Latency could be reduced through LLM compression, approximate nearest neighbor retrieval, and caching, and the reliability of the dialogue could be quantified with explicit grounding metrics, such as the fraction of recommended items matching the retrieved catalog, together with measures of multi-turn consistency.

Conclusions

Summary of Contributions

This dissertation develops deep models for modern recommendation systems, organized around three major contributions that correspond to the three fundamental challenges identified in the introduction: robust and scalable modeling of canonical and auxiliary data, adaptive multi-domain recommendation, and conversational recommendation.

The first contribution is a collection of models for robust and scalable deep modeling of canonical and auxiliary data, addressing scalability, data sparsity, and cold-start.

The second contribution is an adaptive multi-domain recommendation model based on continual learning.

The third contribution is a hybrid conversational recommendation architecture that bridges long-term user behavior with real-time, temporally evolving user intent.

Overall, the proposed models achieve well performance across multiple public benchmarks and a real-world industrial dataset, establishing both theoretical insights and practical solutions that bridge the gap between academic research and large-scale industrial deployment.

Limitations

Despite its contributions, this dissertation has several limitations that point to future research.

Computational Resources: Training graph neural networks, contrastive modules, and continual learning updates remains resource intensive at large scale, even though EfficientRec already lowers the user's side memory footprint to $O(K \cdot d)$.

Auxiliary Information Quality, Robustness, and Fairness: Side information is often noisy, incomplete, or biased, and biased attributes can degrade fairness. The dissertation partially mitigates this and further solutions include denoising and uncertainty aware weighting for noise, imputation and self-supervised pretraining for missing data, and inverse propensity or counterfactual debiasing with exposure-parity constraints for bias and fairness.

Domain Boundary Assumptions: The continual-learning framework assumes reasonably clear domain boundaries and is less effective for highly overlapping or rapidly evolving domains, or for scenarios involving task reordering and domain recurrence.

Conversational System Reliability: Integrating GNNs, large language models, and retrieval-augmented generation enhances reasoning and interaction but introduces inference latency, hallucination risk, and controllability, privacy, and safety concerns that are not yet fully resolved.

Evaluation Scope: Although experiments span multiple public benchmarks and a real-world industrial dataset, they cannot cover the full diversity of practical scenarios; generalization to domains such as finance, healthcare, and education requires further empirical validation and domain-specific adaptation.

Future Work

Future research can extend the findings of this dissertation in several promising directions.

Production Oriented Optimization: Developing fine tuning strategies that continuously adapt models based on real user interactions, system logs, and behavioral drift.

Multimodal Personalization: Expanding the recommendation framework beyond structured interaction and textual signals to incorporate multimodal information including images, audio, video, and device level behavioral cues. **Efficient Deployment:** Exploring lightweight architectures for on device inference, cost efficient LLM compression, and privacy preserving multimodal learning for deploying recommendation systems at scale. .

Fairness and Sustainability: Incorporating explicit fairness constraints and bias mitigation techniques into the optimization objectives.

List of Publications

[P 1] Vu Hong Quan, Le Hoang Ngan, Le Minh Duc, **Nguyen Tran Ngoc Linh**, Le Hoang Quynh. "EfficientRec: An Unlimited User Scale Recommendation System Based on Clustering and User's Interaction Embedding Profile." *In Asian Conference on Intelligent Information and Database Systems (ACIIDS)*, pp. 681-696. Springer Nature Singapore, 2022. ***(Scopus Conference)***

[P 2] **Nguyen Tran Ngoc Linh**, Vu Chi Dung, Le Hoang Ngan, Hoang Anh Dung, Phan Xuan Hieu, Ha Quang Thuy, Le Hoang Quynh, Tran Mai Vu. "GIFT4Rec: An Effective Side Information Fusion Technique Apply to Graph Neural Network for Cold-Start Recommendation." *In Asian Conference on Intelligent Information and Database Systems (ACIIDS)*, pp. 334-345. Springer Nature Singapore, 2023. ***(Scopus Conference)***

[P 3] **Nguyen Tran Ngoc Linh**, Le Hoang Ngan, Hoang Anh Dung, Phan Xuan Hieu, Ha Quang Thuy, Le Hoang Quynh, Tran Mai Vu. "The Masked Simple Graph Contrastive Learning for Recommendation." *In 2024 16th International Conference on Knowledge and System Engineering (KSE)*, pp. 156–160. IEEE, 2024.***(Scopus Conference)***

[P 4] **Nguyen Tran Ngoc Linh**, Vu Chi Dung, Le Hoang Ngan, Hoang Anh Dung, Phan Xuan Hieu, Ha Quang Thuy, Le Hoang Quynh, Tran Mai Vu. "Continual Learning Based on Task Masking for Multi-domain Recommendation." *In Asian Conference on Intelligent Information and Database Systems (ACIIDS)*, pp. 257-266. Springer Nature Singapore, 2024.***(Scopus Conference)***

[P 5] **Nguyen Tran Ngoc Linh**, Hoang Anh Dung, Tran Manh Cuong, Vu Minh Thanh, Vu The Anh, Nguyen Xuan Bach, Bui Tuan Nghia, Le Hoang Quynh, Vuong Thi Hai Yen, Tran Mai Vu. "Improving Retrieval-Augmented Generation for Scalable Movie Chatbots via Graph Based Recommendation Models" Submitted to IEEE Access- under minor revision (Round 3), 2026. ***(Q1 Journal)***