MotiR: Motivation-aware Retrieval for Long-Tail Recommendation

Kaichen Zhao[*,1,2†] **Mingming Li**[1†] **Haiquan Zhao**[2]
Kuien Liu[4,5] **Zhixu Li**[3‡] **Xueying Li**[1‡]
[1] Taobao&Tmall Group, China [2] School of Computer Science, Fudan University
[3] Renmin University of China [4] Academy of Cyber, Beijing, 100846, China
[5] Institute of Software Chinese Academy of Sciences, Beijing, 100190, China
`{22210240394,22210240393}@m.fudan.edu.cn`
`{mingcong.lmm,xiaoming.lxy}@taobao.com`
`zhixuli@ruc.edu.cn` `kuien@iscas.ac.cn`

Abstract

In the retrieval stage of recommendation systems, two-tower models are widely adopted for their efficiency as a predominant paradigm. However, this method, which relies on collaborative filtering signals, exhibits limitations in modeling similarity for long-tail items. To address this issue, we propose a **Moti**vation-aware **R**etrieval for Long-Tail Recommendation, named **MotiR**. The purchase motivations generated by LLMs represent a condensed abstraction of items' intrinsic attributes. By effectively integrating them with traditional item features, this approach enables the two-tower model to capture semantic-level similarities among long-tail items. Furthermore, a gated network-based adaptive weighting mechanism dynamically adjusts representation weights: emphasizing semantic modeling for long-tail items while preserving collaborative signal advantages for popular items. Experimental results demonstrate **60.5%** Hit@10 improvements over existing methods on Amazon Books. Industrial deployment in Taobao&Tmall Group 88VIP scenarios achieves over **4%** CTR and CVR improvement, validating the effectiveness of our method.

1 Introduction

The primary goal of product recommendation systems is to build personalized interest prediction models by analyzing user attributes and historical behavior data. This enables accurate and relevant recommendations. In rapidly growing commercial ecosystems with expanding user bases and product catalogs, adopting an efficient two-stage recommendation framework (retrieval and ranking) has become a key strategy to enhance user retention and boost transaction conversion rates.

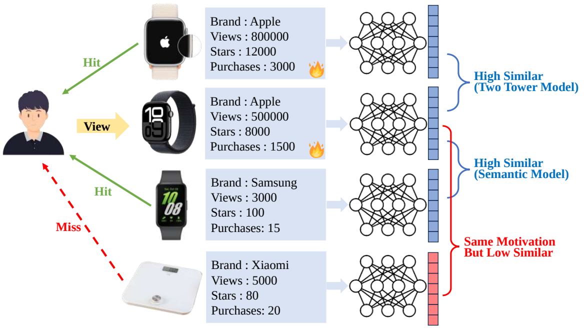


Figure 1: Problems with Existing Retrieval Models in Similarity Modeling of Long-tail Items.

In typical retrieval-stage architectures, the two-tower model (Huang et al., 2013; Covington et al., 2016; Li et al., 2019; Lv et al., 2019) encodes user and item features independently into embedding vectors, using inner product operations to measure user-item interaction probabilities. These models are trained primarily using collaborative filtering signals, which rely on constructing positive and negative sample pairs from user-item interaction records. An ideal two-tower model should satisfy two key properties: (1) maximizing the cosine similarity between user embeddings and the embeddings of their historically interacted items, and (2) maximizing the cosine similarity between embeddings of similar items. While these models are effective in strengthening Property 1, since there is no explicit supervisory signal, they face significant challenges in modeling Property 2, especially for long-tail items with sparse interactions. To address the long-tail issue, some studies (Yi et al., 2019; Huang et al., 2020; Pan et al., 2019; Yao et al., 2021; Zhao et al., 2020) have explored sampling strategies and data augmentation. However, these efforts remain fundamentally reliant on collaborative signals, limiting their effectiveness in fully resolving the problem.

With the development of semantic models, some works (Liu et al., 2022; Li et al., 2023b; Zhang et al., 2024; Ren et al., 2024; Xi et al., 2024) have

[*] Work done during internship at Taobao&Tmall Group
[†] Equal Contribution
[‡] Corresponding authors

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track), pages 934–945
July 28-30, 2025 ©2025 Association for Computational Linguistics

also tried to introduce semantic information into the recommendation system. However, these approaches predominantly leverage item descriptions as training corpora. These texts typically serve as explanations of an item's functionalities or qualities, but they often fail to capture the latent associations between different items adequately. For example, when a user interacts with an item, there typically exists an underlying purchase motivation driving this behavior, and may interact with other items sharing the same motivational attributes. However, purchase motivations represent intrinsic properties embedded within items, yet they often do not exist in item description texts. As a result, these methods still do not solve the problem of modeling similarities of long-tail items fundamentally. As shown in Figure 1, a fitness enthusiast male seeks to purchase an Apple Watch to track his physical activities. While semantic information enables the retrieval of a Samsung Watch (a long-tail item), it fails the retrieval of a Xiaomi Smart Scale which would also align with his fitness goals.

To address the issue of long-tail items, this paper proposes an LLM-driven purchase motivation extract framework. In detail, we utilize LLM to extract the purchase motivations that are embedded behind the item descriptions and convert them into embeddings by using a pre-trained semantic model. Thus, provides similarity associations for long-tail items from the perspective of purchase motivation, enhancing the two-tower model's capacity to learn Property 2.

Besides, collaborative-signal item representations and semantic motivation representations exhibit significant complementary characteristics across different data density scenarios: For popular items with frequent interactions, collaborative filtering-based signal sufficiently for recommendation; while for long-tail items with sparse interactions, purchase motivation provides supplementation through semantic associations. Accordingly, we design a gated network-based adaptive fusion mechanism that dynamically adjusts the weighting coefficients between these two representation types, achieving an optimal combination of item features.

To verify the effectiveness of our proposed method, we conduct experiments on several popular datasets including Amazon Books and Amazon Beauty and Personal Care. Results demonstrate significant improvements in Hit Ratio metrics through motivation feature integration (over **60.5%** Hit@10 improvements on Amazon Books). In real-world deployment for Taobao&Tmall Group 88VIP scenarios, the MotiR achieves over 4% improvement in click-through rate (CTR) and conversion rate (CVR), verifying the practical value of our approach.

Our contributions are listed as follows:

1. We propose a **Moti**vation-aware **R**etrieval method (**MotiR**), which introduces purchase motivation information to achieve effective recommendations of long-tail items with similar motivation.

2. We have innovatively introduced a gated network that dynamically assigns weights based on the popularity of different items, which can effectively recommend both popular and long-tail items.

3. We achieved **60.5%** Hit@10 increase on public datasets and an additional **4%** CTR and CVR gains in Taobao&Tmall Group 88VIP scenarios, demonstrating the effectiveness of our approach.

2 Related Work

2.1 Two-Tower Model

Deep learning techniques have significantly enhanced recommendation systems through end-to-end feature learning, in which two-tower models have emerged as a mainstream architecture for industrial retrieval stages due to their efficient inference. The Wide Deep (Cheng et al., 2016) pioneered the integration of wide linear models with deep neural networks to balance memorization and generalization capabilities, while YouTube DNN (Covington et al., 2016) achieved large-scale video recommendation by modeling deep user behavior sequences. Li et al. (Li et al., 2019) introduced a multi-interest retrieval network to capture diverse interests from user interaction histories. However, existing methods exhibit an over-reliance on collaborative filtering signals, neglecting semantic-level similarity relationships between items, which limits their performance in long-tail scenarios (He et al., 2020).

2.2 LLM For Recommendation System

With recent breakthroughs in large language models (LLMs), researchers have explored leveraging LLMs' common sense knowledge for recommendation tasks: Gao et al. (Hou et al., 2024) demonstrated the potential of LLMs as zero-shot rankers, while P5 (Geng et al., 2022) established a unified generative recommendation framework via prompt engineering. Nonetheless, directly fine-tuning LLMs faces challenges such as high computational costs and latency (Li et al., 2023a). Moreover, current approaches fail to sufficiently exploit fine-grained semantic expressions of user purchase motivations from the perspective of user interest modeling.

3 Method

Addressing the persistent challenge of inadequate similarity modeling for long-tail items in conventional two-tower models (see appendix A.1 for detailed analyses), We employ LLMs to extract item purchase motivations (3.1) and systematically integrate them into the two-tower architecture (3.2). Building upon this foundation, we propose a gated network mechanism that dynamically modulates the weighting between the item tower and semantic tower (3.3), implemented through a three-phase progressive training framework (3.4). These methodological innovations and their detailed implementations will be systematically elaborated in subsequent sections.

3.1 Motivation-Aware Item Representation

We aim to address the insufficient capability of existing retrieval models in modeling similarity relationships for long-tail items. The essence of user consumption behavior can be attributed to the matching between item intrinsic attributes and user demand motivations. Based on this, we propose the **Purchasing Motivation Consistency Hypothesis**: when a user interacts with a particular item, they are more likely to engage with other items that share the same purchase motivation. This hypothesis provides a novel theoretical perspective for item similarity modeling — achieving semantic alignment of similar items through mining their implicit purchasing motivations.

To realize this hypothesis, we innovatively introduce large language models as prior knowledge distillers, which can parse potential user purchasing motivation sets from item description texts. The motivation set of each item is transformed into a motivation vector $\mathbf{m}_i \in \mathbb{R}^d$ via a semantic encoder (Xiao et al., 2024), constituting the item semantic representation. This approach offers two critical advantages:

- **Prior Knowledge Guidance**: The motivation vector encoding process operates independently of user interaction data, directly leveraging LLM-internalized knowledge about item attributes and human consumption custom, which breaks away from the dependency of collaborative filtering data.
- **Semantic Transferability**: Mapping motivation texts to a continuous vector space through pre-trained semantic embedding models, ensures the similarity of similar purchase motivations in the vector space.

3.2 Multimodal Feature Fusion Architecture

The final item representation $\mathbf{e}_i \in \mathbb{R}^{2d}$ is constructed through dual-channel feature concatenation:

$$\mathbf{e}_i = \left(\underbrace{f_{\text{ID}}(i)}_{\text{Collaborative Signal}}, \underbrace{f_{\text{Motivation}}(\mathbf{m}_i)}_{\text{Semantic Prior}} \right)$$

Where $f_{\text{ID}}(\cdot)$ denotes the traditional ID-based feature encoder (item tower), and $f_{\text{Motivation}}(\cdot)$ represents the motivation feature encoder (semantic tower). This architecture achieves dual complementary effects:

- **Data Sufficiency Compensation**: The ID representation captures explicit collaborative patterns through massive interaction data, dominating precise recommendations for high-frequency items.
- **Semantic Robustness Enhancement**: The motivation representation provides cross-instance similarity association for low-frequency items via LLM-extracted semantic priors.

This fusion mechanism essentially constructs a joint optimization space for collaborative signals and semantic priors. In interaction-sparse regions, the semantic similarity of motivation vectors guides the model to establish a more reasonable item association, significantly improving the traditional two-tower model's deficiency in optimizing Property II (vector alignment of similar items).

3.3 Dynamic Feature Fusion Mechanism

The item tower and the semantic tower respectively model explicit interaction patterns and implicit semantic attributes of commodities, forming complementary representation spaces. However, naive

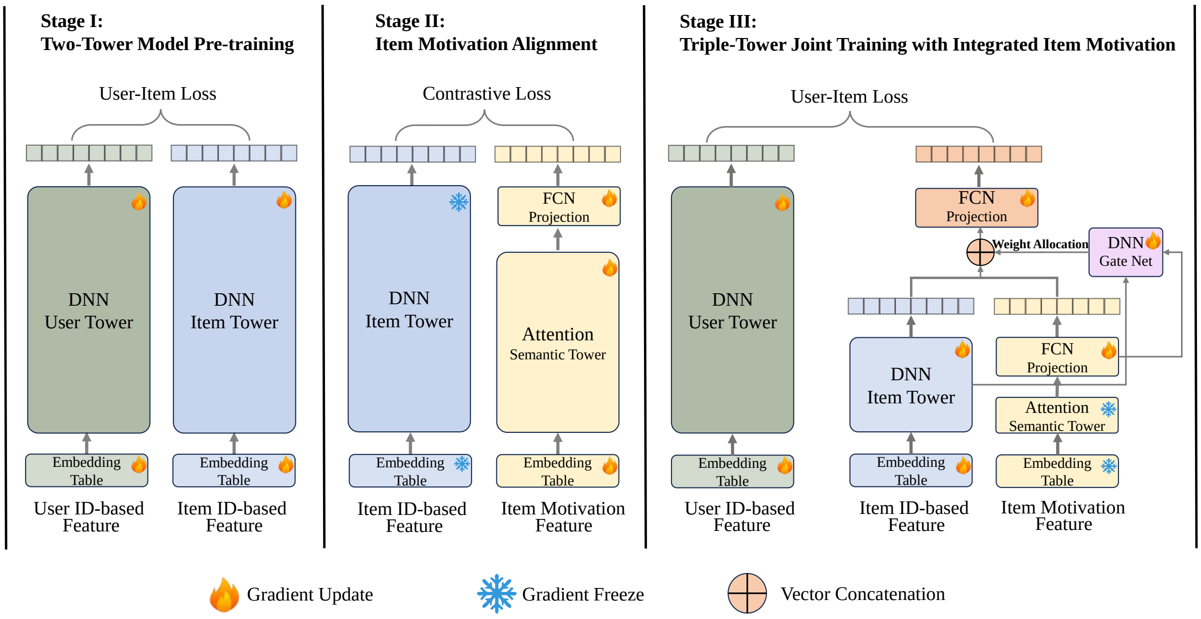


Figure 2: Three-Stage Retrieval Model Training Framework with Integrated Motivational Signals.

feature concatenation fails to achieve adaptive coordination of modal advantages.

Inspired by PEPNet (Chang et al., 2023) personalized bias mechanism, we propose an **Adaptive Gated Fusion Network** that dynamically modulates modal weights based on commodity characteristics. The mathematical formulation is defined as:

$$\mathbf{e}_i = \alpha_i \cdot \mathbf{h}_{\text{CF}}^{(i)} + (1 - \alpha_i) \cdot \mathbf{h}_{\text{Sem}}^{(i)}$$

where the gating coefficient $\alpha_i \in [0, 1]$ is generated through:

$$\alpha_i = \sigma \left(\mathbf{W}_g \cdot \text{Concat}(\mathbf{h}_{\text{CF}}^{(i)}, \mathbf{h}_{\text{Sem}}^{(i)}) + b_g \right)$$

Here, $\mathbf{W}_g \in \mathbb{R}^{(d_m+1)\times 1}$ and $b_g \in \mathbb{R}$ are learnable parameters, with $\sigma(\cdot)$ being the Sigmoid activation function.

The gated network adaptively learns weighting strategies for the item tower and semantic tower based on intrinsic item attributes. Qualitative analysis reveals that for frequently interacted items, it augments weights on the collaborative signal-driven item tower while prioritizing the semantic tower for long-tail items.

3.4 Three-Stage Training Approach

Since the semantic features extracted by the LLM and the item embeddings from the item tower reside in distinct feature spaces, directly incorporating motivational semantics may cause feature distribution shifts, making it difficult for the item tower to effectively interpret semantic information. To address this challenge, we employ a contrastive learning approach to align the feature spaces between the item tower and semantic tower before feature fusion, thereby enhancing convergence speed and training stability (Wang et al., 2024). As shown in Figure 2, the training pipeline consists of three stages:

- **stage 1**: Independently train the two-tower retrieval model without semantic features.
- **stage 2**: Align the item tower embeddings and semantic model vectors into a unified feature space through contrastive learning.
- **stage 3**: Jointly fine-tune the model by integrating semantic features into item representations.

4 Experiment

4.1 Datasets and Evaluation Metrics

The experiments are conducted on two publicly available e-commerce datasets: Amazon Beauty and Amazon Books. Specifically, the Books dataset contains approximately 4.4 million items and 10 million user-item interactions. The Beauty dataset covers 11 million user-item interactions and 1 million items.

To simulate real-world sequential recommendation scenarios, user behavior sequences are chronologically split to construct "next-item prediction"

Methods	Books				Beauty			
	hit@10	hit@50	hit@100	hit@500	hit@10	hit@50	hit@100	hit@500
WALS	1.42%	3.97%	5.28%	10.93%	1.92%	5.49%	6.95%	11.93%
YoutubeDNN	2.53%	7.76%	12.90%	19.54%	3.30%	8.27%	15.19%	23.44%
MaxMF	2.85%	8.62%	13.04%	21.37%	3.59%	9.06%	16.24%	25.10%
Mind	3.09%	11.01%	16.31%	24.59%	4.86%	13.22%	20.87%	31.77%
SASRec (2023)	2.92%	7.29%	-	-	4.25%	11.58%	19.67%	29.79%
HSTU	4.69%	10.66%	-	-	-	-	-	-
MotiR (ours)	**4.96%**	**15.29%**	**20.01%**	**31.07%**	**6.46%**	**16.28%**	**24.22%**	**36.40%**

Table 1: Main Results of MotiR and Other Mainstream Methods.

tasks for evaluation. Model performance is quantified using Hit Ratio@k (hit@k), defined as: for a given user if the ground-truth interacted item appears in the top-k recommended list after full-item ranking, the retrieval is considered successful.

$$\text{hit@}k = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \mathbf{1}\left(i_u^{\text{gt}} \in \mathcal{L}_u^k \right)$$

Where $\mathcal{U}$ denotes the set of users, i_u^{gt} is the ground-truth interacted item for user u, $\mathcal{L}_u^k$ represents the top-k recommended items after full-item ranking and $\mathbf{1}(\cdot)$ is an indicator function (1 if true, 0 otherwise)

4.2 Experiment Settings

We compare the proposed MotiR with the following retrieval models: **WALS (Aberger, 2014)** and **MaxMF (Weston et al., 2013)** are recommendation algorithms based on traditional collaborative filtering mechanisms. **YouTube DNN (Covington et al., 2016)** and **MIND (Li et al., 2019)** introduce deep neural networks into recommendation systems, representing mainstream baseline models for retrieval in current research. **SASRec (Kang and McAuley, 2018)** and **HSTU (Zhai et al., 2024)** are the research of introducing large language models with transformer architecture into recommendation systems.

The experimental configurations were established as follows: In academic research scenarios, we adopted PyTorch 1.13 deep learning framework for prototype development, while employing TensorFlow 1.12 framework for distributed training in industrial application scenarios. The entire training process was accelerated by 8 NVIDIA V100 GPUs, with the batch size set to 512. Our LLM-based (Achiam et al., 2023) motive parsing framework automatically extracts semantic features through structured prompts, and the appendix A.2 shows the prompt template. The training procedure was systematically divided into three distinct stages with differentiated optimization objectives. Detailed training parameters and configurations for each stage are provided in the appendix A.3.

interaction	item tower	semantic tower	item nums
$[5-10)$	32.76%	68.24%	2.48M
$[10-20)$	43.28%	56.72%	1.04M
$[20-50)$	53.24%	46.76%	0.76M
$[50, \infty)$	60.59%	39.41%	0.12M

Table 2: Weight Allocation Results of the Gated Network Between Item Tower and Semantic Tower.

4.3 Main Results

The experimental results on Amazon Books and Amazon Beauty datasets demonstrate the superior performance of MotiR compared to mainstream baseline methods. As shown in table 1, in the Books domain, our method achieves a hit@10 of 4.96%, representing a 60.5% relative improvement over the collaborative signal-based model (Mind) and beyond the LLM-based model over 5% (HSTU). Similarly, in the Beauty domain, the hit@10 metric improves from 4.86% to 6.46%, with a 32.9% relative gain.

4.4 Ablation Study

4.4.1 Gated Network Weight Allocation

The ablation study on gated network weight allocation reveals a clear correlation between item interaction frequency and different item representation models (item tower and semantic tower). As shown in table 2, for items with sparse interactions (5-10 interactions), the semantic tower dominates with 68.24% weight allocation, while the item tower only accounts for 32.76%. This weighting pattern gradually reverses as interaction frequency increases. The inverse proportionality between interaction frequency and semantic tower weight quantitatively verifies our core hypothesis: the gated network automatically establishes

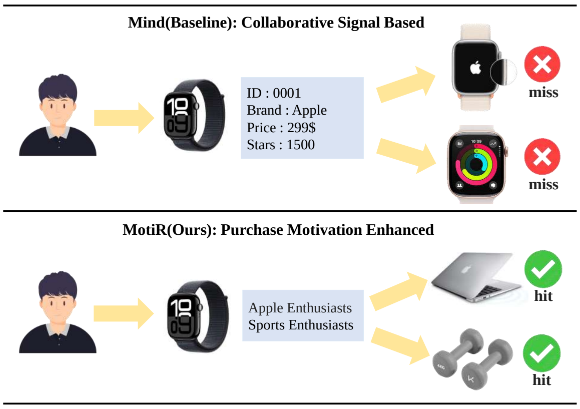


Figure 3: Recommendations for Apple Fitness Trackers with Between Mind and MotiR.

a "collaborative-to-semantic" continuum based on item popularity.

4.4.2 Different Semantic Information

To systematically validate the effectiveness of semantic information in item similarity modeling, this study conducts a comparative analysis of different semantic features on retrieval performance. As shown in table 3, experimental results on the public Amazon dataset demonstrate that LLM-generated purchase motivation features achieve significant improvements in model Hit Ratio. This finding substantiates our core hypothesis - that the motivation semantics distilled through LLMs can effectively capture deep-level associations between items. Compared to the surface-level information provided by item titles, motivation descriptions enable feature extension through contextual reasoning. Meanwhile, relative to the redundant textual content in item descriptions, the motivation extraction process achieves effective noise reduction and clustering for item features.

4.4.3 Different Large Language Models

We also analyze the extraction performance of different LLMs on purchase motivations. We conducted experiments using several API-based LLMs (GPT-4 (Achiam et al., 2023), ChatGPT (Ouyang et al., 2022), Qwen2.5-Max (Yang et al., 2024)) and some open-source LLMs (Qwen2.5-7B-Instruct (Yang et al., 2024), Baichuan2-7B-Chat (Yang et al., 2023), Meta-Llama-3-8B-Instruct (Grattafiori et al., 2024)). Table 5 demonstrates the impact of different LLM-generated purchase motivations on MotiR's hit ratio.

Among all LLMs, GPT-4 achieved the best performance in motivation extraction. However, it is noteworthy that there exists no significant performance gap between different LLMs. Even open-source models with relatively smaller parameters like Qwen2.5-7B-Instruct can attain nearly comparable effectiveness to GPT-4. This observation suggests that our method does not heavily depend on the semantic comprehension capabilities of LLMs, most LLMs can extract reasonable purchase motivations from product descriptions, thereby enabling the two-tower model to better capture similarities among long-tail items. Consequently, practical production scenarios may consider adopting relatively compact open-source LLMs to reduce cost and time overhead for purchase motivation generation.

	hit@10	hit@100	hit@500
Non-semantic	3.09%	16.31%	24.59%
Item Title	4.02%	17.90%	26.97%
Item Description	4.56%	18.71%	29.87%
Purchase Motivation	**4.96%**	**20.01%**	**31.07%**

Table 3: Impact of Diverse Semantic Information on Model Hit Ratio.

	hit@100	hit@500	hit@3000
Mind	7.45%	14.60%	28.20%
MotiR (ours)	**10.04%**	**18.98%**	**37.02%**

Table 4: Real Scenery Results of MotiR and Baseline Method.

5 Industrial Application

5.1 Revenue Analysis

As the most valuable consumer cohort with the highest purchasing power and loyalty on the Taobao&Tmall Group, the 88VIP membership has reached a scale of tens of millions, sustaining daily active users (DAU) at the ten-million level, while its annual contribution to Gross Merchandise Volume (GMV) has surpassed RMB 2 trillion.

The online baseline retrieval model constitutes an enhanced version based on the Mind (Li et al., 2019) model architecture, with multi-level Squeeze-and-Excitation (Hu et al., 2018) (SENet) layers incorporated into the feature interaction module, and systematic optimization of data sampling strategies being implemented during the training stage.

In online A/B testing for Taobao&Tmall Group 88VIP homepage recommendations, As shown in table 4, our proposed MotiR model achieved a relative improvement exceeding 20% in Hit Ratio compared to the baseline system. Additionally, the model delivered relative **gains of 4.76% in Click-**

Methods	Books				Beauty			
	hit@10	hit@50	hit@100	hit@500	hit@10	hit@50	hit@100	hit@500
GPT-4	**4.96%**	**15.29%**	**20.01%**	**31.07%**	**6.46%**	**16.28%**	**24.22%**	**36.40%**
Qwen2.5-Max	4.85%	15.20%	19.71%	30.59%	6.40%	16.11%	23.90%	36.01%
ChatGPT	4.79%	15.10%	19.44%	30.25%	6.27%	15.96%	23.64%	35.81%
Qwen2.5-7B-Instruct	4.77%	15.04%	19.46%	30.30%	6.25%	15.82%	23.59%	35.72%
Baichuan2-7B-Chat	4.62%	14.77%	18.97%	29.64%	6.06%	15.29%	22.70%	33.95%
Meta-Llama-3-8B-Instruct	4.03%	12.54%	15.45%	25.58%	5.30%	14.07%	19.59%	30.56%

Table 5: Impact of Different LLM on Model Hit Ratio.

Through Rate (CTR) and 4.35% in Conversion Rate (CVR) (p<0.01), generating substantial business value for the platform. Figure 3 demonstrates a case analysis of the retrieval model enhanced with purchase motivations in Taobao&Tmall internal datasets. Our method gets item correlations from the purchase motivation perspective, thereby enabling enhanced capture of user interest.

The long-tail effect proves particularly prominent in real-world recommendation scenarios: weekly interacted items by Taobao&Tmall Group 88VIP users constitute less than 30% of the entire item catalog. This phenomenon proves the necessity of modeling item similarity through purchasing motivation. Evaluation reveals that the semantic representation module attains an average weight allocation of 62.3±1.5% during model inference, strongly validating the critical role of semantic features in long-tail item recommendation.

5.2 Computational Overhead Analysis

We conduct a detailed time complexity analysis for each training phase:

- **Stage 1**: Conventional two-tower model training requires 20 epochs, accounting for approximately 60% of the total training time.
- **Stage 2**: The contrastive alignment process completes within 0.2 epoch, consuming merely 5% of the computational budget.
- **Stage 3**: Since the weights of the semantic model are no longer trained except for the projection layer, we implement an optimized training method where item embeddings from the semantic model's base layer are precomputed offline, which consumes 20% of the total time. The subsequent joint fine-tuning of semantic projections and two-tower parameters completes in 3 epochs, requiring 15% additional computation.

In these stages, the training of the two-tower model still takes up most of the time, while the additional time overhead caused by the introduction of semantic information is acceptable. For industrial deployment, the embeddings of all items are precalculated offline. During real-time serving, the enhanced retrieval system maintains identical computational complexity to conventional two-tower architectures, as it only requires standard vector similarity calculations between user and item embeddings. This design ensures our method incurs no additional computational overhead during online inference while achieving significant performance improvements.

6 Conclusion

Cause of the traditional two-tower model has poor modeling capabilities for long-tail items similarity, this paper proposes a **Moti**vation **R**etrieval method (**MotiR**). We leverage LLM to extract purchase motivations, constructing semantic embedding spaces to capture implicit associations. A gated network enables data density-aware adaptive fusion: emphasizing semantic representations for long-tail items while preserving collaborative advantages for popular items. Our Method effectively alleviates the problem of insufficient similarity modeling capabilities of traditional retrieval models in long-tail items. Real-world deployment in Taobao&Tmall Group 88VIP scenarios achieves over 4% CTR and CVR gains.

7 Acknowledgement

This work was supported by Alibaba Group through Alibaba Research Intern Program, Suzhou Key Laboratory of Artificial Intelligence and Social Governance Technologies (SZS2023007), Smart Social Governance Technology and Innovative Application Platform (YZCXPT2023101), and the Leadership Talent Program (Science and Education) of SIP.

References

Christopher R Aberger. 2014. Recommender: An analysis of collaborative filtering techniques. *Personal and Ubiquitous Computing Journal*, 5.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jianxin Chang, Chenbin Zhang, Yiqun Hui, Dewei Leng, Yanan Niu, Yang Song, and Kun Gai. 2023. Pepnet: Parameter and embedding personalized network for infusing with personalized prior information. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3795–3804.

Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, and 1 others. 2016. Wide & deep learning for recommender systems. In *Proceedings of the 1st workshop on deep learning for recommender systems*, pages 7–10.

Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*, pages 191–198.

Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2022. Recommendation as language processing (rlp): A unified pretrain, personalized prompt & predict paradigm (p5). In *Proceedings of the 16th ACM conference on recommender systems*, pages 299–315.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. 2020. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 639–648.

Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. 2024. Large language models are zero-shot rankers for recommender systems. In *European Conference on Information Retrieval*, pages 364–381. Springer.

Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141.

Jui-Ting Huang, Ashish Sharma, Shuying Sun, Li Xia, David Zhang, Philip Pronin, Janani Padmanabhan, Giuseppe Ottaviano, and Linjun Yang. 2020. Embedding-based retrieval in facebook search. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2553–2561.

Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 2333–2338.

Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *2018 IEEE international conference on data mining (ICDM)*, pages 197–206. IEEE.

Chao Li, Zhiyuan Liu, Mengmeng Wu, Yuchi Xu, Huan Zhao, Pipei Huang, Guoliang Kang, Qiwei Chen, Wei Li, and Dik Lun Lee. 2019. Multi-interest network with dynamic routing for recommendation at tmall. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 2615–2623.

Jiacheng Li, Ming Wang, Jin Li, Jinmiao Fu, Xin Shen, Jingbo Shang, and Julian McAuley. 2023a. Text is all you need: Learning language representations for sequential recommendation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1258–1267.

Xiangyang Li, Bo Chen, Lu Hou, and Ruiming Tang. 2023b. Ctrl: Connect collaborative and language model for ctr prediction. *ACM Transactions on Recommender Systems*.

Guang Liu, Jie Yang, and Ledell Wu. 2022. Ptab: Using the pre-trained language model for modeling tabular data. *arXiv preprint arXiv:2209.08060*.

Fuyu Lv, Taiwei Jin, Changlong Yu, Fei Sun, Quan Lin, Keping Yang, and Wilfred Ng. 2019. Sdm: Sequential deep matching model for online large-scale recommender system. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 2635–2643.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Feiyang Pan, Shuokai Li, Xiang Ao, Pingzhong Tang, and Qing He. 2019. Warm up cold-start advertisements: Improving ctr predictions via learning to learn id embeddings. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 695–704.

Xubin Ren, Wei Wei, Lianghao Xia, Lixin Su, Suqi Cheng, Junfeng Wang, Dawei Yin, and Chao Huang. 2024. Representation learning with large language models for recommendation. In *Proceedings of the ACM Web Conference 2024*, pages 3464–3475.

Xingmei Wang, Weiwen Liu, Xiaolong Chen, Qi Liu, Xu Huang, Yichao Wang, Xiangyang Li, Yasheng Wang, Zhenhua Dong, Defu Lian, and 1 others. 2024. Cela: Cost-efficient language model alignment for ctr prediction. *arXiv preprint arXiv:2405.10596*.

Jason Weston, Ron J Weiss, and Hector Yee. 2013. Nonlinear latent factorization by embedding multiple user interests. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 65–68.

Yunjia Xi, Weiwen Liu, Jianghao Lin, Xiaoling Cai, Hong Zhu, Jieming Zhu, Bo Chen, Ruiming Tang, Weinan Zhang, and Yong Yu. 2024. Towards open-world recommendation with knowledge augmentation from large language models. In *Proceedings of the 18th ACM Conference on Recommender Systems*, pages 12–22.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 641–649.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, and 1 others. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Tiansheng Yao, Xinyang Yi, Derek Zhiyuan Cheng, Felix Yu, Ting Chen, Aditya Menon, Lichan Hong, Ed H Chi, Steve Tjoa, Jieqi Kang, and 1 others. 2021. Self-supervised learning for large-scale item recommendations. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pages 4321–4330.

Xinyang Yi, Ji Yang, Lichan Hong, Derek Zhiyuan Cheng, Lukasz Heldt, Aditee Kumthekar, Zhe Zhao, Li Wei, and Ed Chi. 2019. Sampling-bias-corrected neural modeling for large corpus item recommendations. In *Proceedings of the 13th ACM conference on recommender systems*, pages 269–277.

Jiaqi Zhai, Lucy Liao, Xing Liu, Yueming Wang, Rui Li, Xuan Cao, Leon Gao, Zhaojie Gong, Fangda Gu, Michael He, and 1 others. 2024. Actions speak louder than words: Trillion-parameter sequential transducers for generative recommendations. *arXiv preprint arXiv:2402.17152*.

Chiyu Zhang, Yifei Sun, Jun Chen, Jie Lei, Muhammad Abdul-Mageed, Sinong Wang, Rong Jin, Sem Park, Ning Yao, and Bo Long. 2024. Spar: Personalized content-based recommendation via long engagement attention. *arXiv preprint arXiv:2402.10555*.

Cheng Zhao, Chenliang Li, Rong Xiao, Hongbo Deng, and Aixin Sun. 2020. Catn: Cross-domain recommendation for cold-start users via aspect transfer network. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pages 229–238.

A Appendix

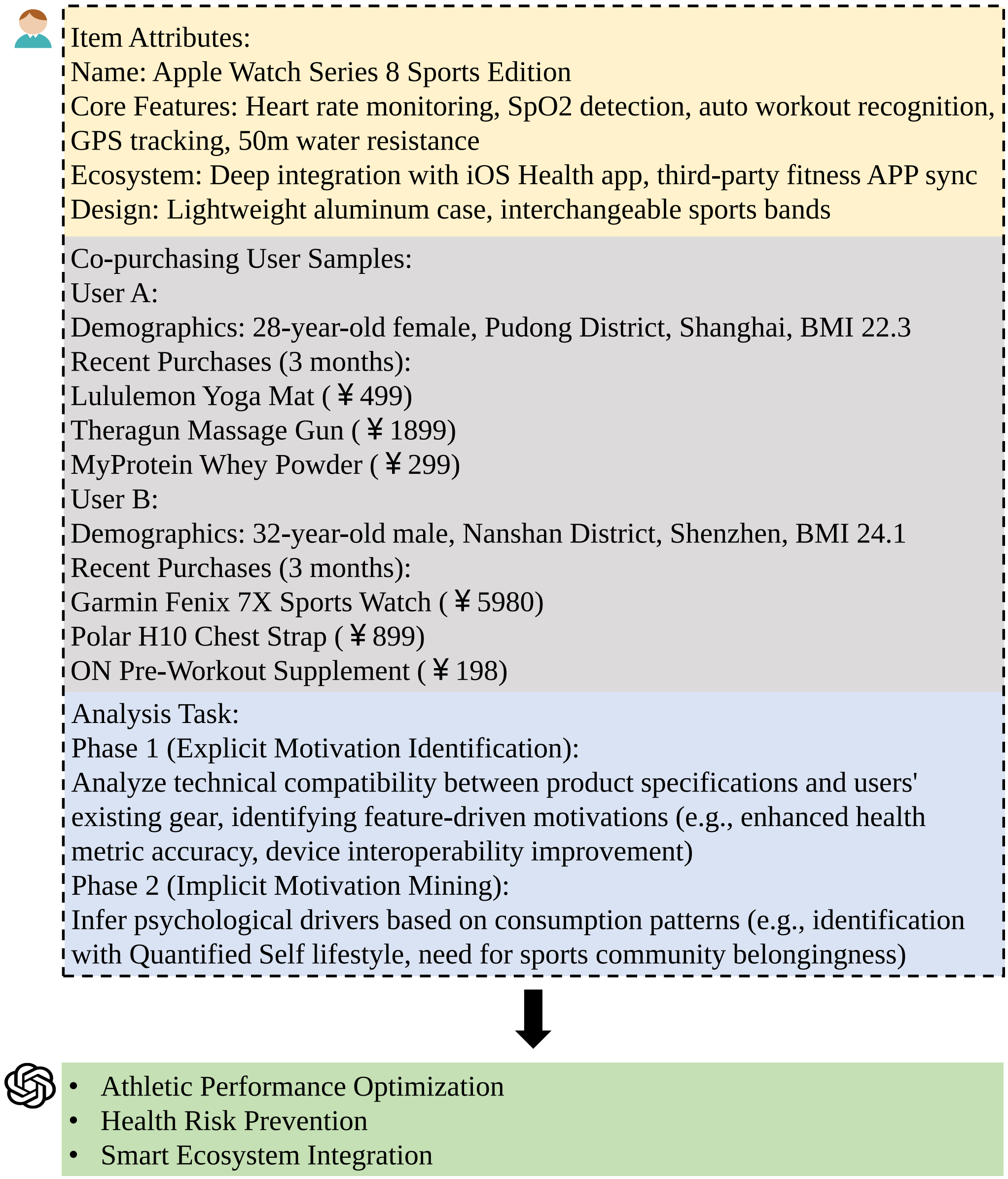
A.1 Properties of Two-Tower Models

The modeling logic of traditional two-tower recommendation systems is established on the **collaborative filtering hypothesis**: if a user has interacted with item A, they are more likely to interact with items similar to A. To satisfy this hypothesis, an ideal two-tower model must simultaneously guarantee two critical properties: (1) **User-Item Interaction Explicit Alignment**: The cosine similarity between a user's representation vector and the vectors of their historically interacted items should be maximized. (2) **Item-Item Semantic Implicit Alignment**: The cosine similarity between item pairs with semantic similarity should be maximized in the vector space.

Existing two-tower models primarily train through user-item collaborative signals. Within this framework, the optimization objective of the first property is achieved via explicit supervisory signals, while the learning of the second property suffers from an **inherent deficiency** — the model can only implicitly capture item similarity through statistical patterns in user behavior, rather than receiving explicit supervision. Specifically, when two items are frequently interacted with by the same users, the model passively adjusts their vector similarity. Notably, there exists no explicit supervisory signal requiring the cosine similarity between similar items to be maximized.

Through theoretical analysis, this paper reveals two fundamental limitations of traditional approaches in modeling the second property:

1. **Representation Distortion in Long-Tail Items.** Interaction data in recommendation systems generally follows a long-tail distribution. Under the collaborative signal-based learning mechanism, insufficient training samples for cold items lead to inadequate updates



Item Attributes:
Name: Apple Watch Series 8 Sports Edition
Core Features: Heart rate monitoring, SpO2 detection, auto workout recognition, GPS tracking, 50m water resistance
Ecosystem: Deep integration with iOS Health app, third-party fitness APP sync
Design: Lightweight aluminum case, interchangeable sports bands

Co-purchasing User Samples:
User A:
Demographics: 28-year-old female, Pudong District, Shanghai, BMI 22.3
Recent Purchases (3 months):
Lululemon Yoga Mat (￥499)
Theragun Massage Gun (￥1899)
MyProtein Whey Powder (￥299)
User B:
Demographics: 32-year-old male, Nanshan District, Shenzhen, BMI 24.1
Recent Purchases (3 months):
Garmin Fenix 7X Sports Watch (￥5980)
Polar H10 Chest Strap (￥899)
ON Pre-Workout Supplement (￥198)

Analysis Task:
Phase 1 (Explicit Motivation Identification):
Analyze technical compatibility between product specifications and users' existing gear, identifying feature-driven motivations (e.g., enhanced health metric accuracy, device interoperability improvement)
Phase 2 (Implicit Motivation Mining):
Infer psychological drivers based on consumption patterns (e.g., identification with Quantified Self lifestyle, need for sports community belongingness)

- Athletic Performance Optimization
- Health Risk Prevention
- Smart Ecosystem Integration

Figure 4: Purchace Motivation Extract Prompt for LLM.

of their representation vectors, making it difficult to accurately reflect their semantic attributes.

2. **Semantic Disconnection in Feature Encoding.** Traditional ID-based feature encoding schemes exhibit an inherent flaw. When independently mapping semantically related features A and B through ID embeddings, the geometric relationships in the vector space become decoupled from the original feature semantic similarity. Resulting in similar items losing their exact similarity after encoding.

A.2 Motivation Extract

- **Core Semantic Feature Extraction**: Employing GPT-4 (Achiam et al., 2023) as the central semantic parsing engine, we generate purchase motivation descriptions through deep semantic reasoning on item descriptions. For public benchmarks (e.g., Amazon datasets), input features are strictly limited to item titles and official descriptions.

- **Behavioral Feature Augmentation**: For real-world e-commerce scenarios, we design a multimodal feature fusion strategy: beyond basic item descriptions, co-purchasing behavior features are incorporated. Specifi-

epoch	similarity	top10 (bs=512)	hit@10	hit@50	hit@100	hit@500
0	0.01	7.92%	4.79%	14.90%	19.56%	30.39%
0.2	0.54	60.19%	4.96%	15.29%	20.01%	31.07%
0.5	0.81	71.72%	4.82%	15.08%	19.49%	30.21%
1	0.92	78.96%	4.29%	13.16%	17.15%	27.10%
2	0.98	85.23%	3.57%	10.96%	15.60%	24.19%

Table 6: Impact of Contrastive Learning on Amazon Books dataset.

cally, two randomly selected users with co-purchase relationships are sampled for each target item. Their demographic attributes (age, gender, location) and recent purchase history (anonymized) form supplementary contextual signals, and provide more extensive background knowledge for the extraction of purchase motivations. Figure 4 shows a prompt for LLM to extract purchase motivations in a real scenario.

A.3 Training Details

The model training process adopts a progressive three-stage optimization strategy, with hyperparameter configurations and training objectives detailed as follows:

1. **Base Two-Tower Model Pretraining**: In the initial stage, we independently train the user-item two-tower model for 10 epochs with a dynamically decaying learning rate (from 1e-3 to 1e-4). This stage establishes the fundamental collaborative representation space between users and items. Training employs the Adam optimizer with a batch size of 512 and a dropout ratio of 0.2. To enhance positive-negative sample discrimination, we set the temperature parameter to 0.05 and adopt a balanced global negative sampling and in-batch negative sampling strategy (64 global negatives and 64 in-batch negatives per sample), optimized through a sampled softmax loss function.

2. **Semantic Representation Alignment**: The second stage introduces contrastive learning with a BGE-pretrained semantic encoder. Conducted over 0.5 epochs, this stage projects the 256-dimensional semantic features into 128-dimensional space through a learnable projection layer while keeping the item tower parameters frozen. The learning rate linearly decays from 3e-4 to 5e-5. This alignment process geometrically maps semantic and collaborative representations into a unified feature space, laying the foundation for subsequent fusion. Notably, the potential representation homogenization caused by contrastive learning will be thoroughly analyzed in the following Section.

3. **Multimodal Fusion Fine-tuning**: The final stage involves 3 epochs of joint optimization focusing on the gated network and semantic projection layer. We freeze the base parameters of the semantic model while updating its terminal projection layer, and resume parameter updates for the two-tower model. The learning rate decays from 1e-4 to 1e-5. The gated network utilizes a two-layer fully connected architecture, taking concatenated vectors from the item tower (128-dimensional) and semantic tower (128-dimensional) as input, and outputs a 2D weight vector for dynamic feature fusion. In addition, after the two vectors are concatenated, a layer of projection is performed to restore the dimension from 256 to 128.

A.4 Impact of Contrastive Learning

Experimental results reveal a non-linear relationship between contrastive learning duration and model performance. As shown in table 6, while the cosine similarity between semantic and item tower monotonically increases with training epochs, the retrieval performance metrics (hit@k) exhibit a significant inverted U-shaped curve. When contrastive learning proceeds for 0.2 epochs, the similarity reaches 0.54 with peak Hit Ratio metrics. However, extending training to 2 epochs results in similarity rising to 0.98 but the Hit Ratio declines significantly, approaching the baseline performance without semantic modeling.

This phenomenon demonstrates the dual effects of contrastive learning: Proper feature alignment helps reduce the huge feature differences between the two item representation models, providing a foundation for subsequent fusion; whereas excessive alignment causes over-homogenization between semantic and item tower, diminishing the complementary benefits of semantic modeling.

The impact of contrastive learning on model performance is similar in Amazon Books datasets and real-world applications in the industry. On Taobao&Tmall Group 88VIP, we terminate contrastive learning when the cross-modal similarity threshold arrives at 0.5 to 0.6 and subsequently initiating the multimodal fusion tuning stage leads to optimal retrieval performance.