Harnessing Multimodal Large Language Models for Personalized Product Search with Query-aware Refinement

Beibei Zhang
State Key Laboratory for Novel Software Technology, Nanjing University
Nanjing, China
zhangbb@smail.nju.edu.cn

Yanan Lu
Tencent
Beijing, China
yananlu@tencent.com

Ruobing Xie
Tencent
Beijing, China
xrbsnowing@163.com

Zongyi Li
Huazhong University of Science and Technology
Wuhan, China
zongyili@hust.edu.cn

Siyuan Xing
Tencent
Beijing, China
siyuanxing@tencent.com

Tongwei Ren*
State Key Laboratory for Novel Software Technology, Nanjing University
Nanjing, China
rentw@nju.edu.cn

Fen Lin
Tencent
Beijing, China
felicialin@tencent.com

Abstract

Personalized product search (PPS) aims to retrieve products relevant to the given query considering user preferences within their purchase histories. Since large language models (LLM) exhibit impressive potential in content understanding and reasoning, current methods explore to leverage LLM to comprehend the complicated relationships among user, query and product to improve the search performance of PPS. Despite the progress, LLM-based PPS solutions merely take textual contents into consideration, neglecting multimodal contents which play a critical role for product search. Motivated by this, we propose a novel framework, HMPPS, for **H**arnessing **M**ultimodal large language models (MLLM) to deal with **P**ersonalized **P**roduct **S**earch based on multimodal contents. Nevertheless, the redundancy and noise in PPS input stand for a great challenge to apply MLLM for PPS, which not only misleads MLLM to generate inaccurate search results but also increases the computation expense of MLLM. To deal with this problem, we additionally design two query-aware refinement modules for HMPPS: 1) a perspective-guided summarization module that generates refined product descriptions around core perspectives relevant to search query, reducing noise and redundancy within textual contents; and 2) a two-stage training paradigm that introduces search query for user history filtering based on multimodal representations, capturing precise user preferences and decreasing the inference cost. Extensive experiments are conducted on four public datasets to demonstrate the effectiveness of HMPPS. Furthermore, HMPPS is deployed on an online search system with billion-level daily active users and achieves an evident gain in A/B testing.

CCS Concepts

• **Information systems** → **Personalization**;

Keywords

Product Search, Personalization, Multimodal Search, Multimodal Large Language Models

ACM Reference Format:
Beibei Zhang, Yanan Lu, Ruobing Xie, Zongyi Li, Siyuan Xing, Tongwei Ren, and Fen Lin. 2025. Harnessing Multimodal Large Language Models for Personalized Product Search with Query-aware Refinement. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3746027.3754898

1 Introduction

Product search aims to return a ranked list of products in response to the given query submitted by users, which plays a vital role in online shopping services. However, since limited queries are too ambiguous to effectively express the underlying preferences of users, the returned search results cannot satisfy user purchase intents precisely. To deal with this problem, more and more researches focus on personalized product search, which additionally takes user purchase history into consideration to model user preferences, contributing to capturing the exact purchase intents of users [1–4, 6, 13, 17, 28, 42].

User, product and query are three core concepts in PPS. How to effectively represent the three and construct their complex relationships comprise the primary challenges of PPS. The primary practice in PPS is to convert user and product IDs into embedding vectors with trainable parameters and then conduct interaction modeling to predict the relevance among them [3, 13, 16, 28, 39]. ID-based solutions are adept at learning the inherent pattern from the abundant log data for PPS and their limited-size embeddings

*Corresponding author.

MM '25, Dublin, Ireland

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3754898

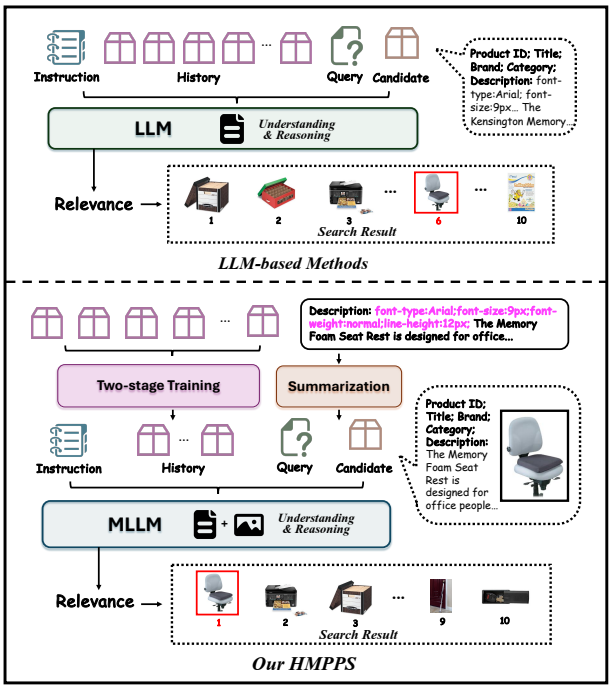

Figure 1: Comparison between existing LLM-based methods *vs.* HMPPS. Here, the target product is labeled with a red box.

contribute to the storage and computation efficiency. Nevertheless, highly relying on ID-based data results in an inadequate understanding of the valuable content-based information of PPS, leading to a sub-optimal search result [25]. Moreover, the negative impact of data bias, *e.g.,* popularity bias [8], is inevitable for limited shopping logs. As a result, the low frequency and new unseen samples are necessarily insufficiently learned, harming PPS accuracy ultimately [26, 41].

To make up for the weakness of ID-based methods, there have been some endeavors to exploit content-based information for PPS [6, 15, 34]. Particularly noteworthy is that the emergence of large language models, which exhibits revolutionized contribution to natural language processing, has inspired researchers to leverage their remarkable content understanding and reasoning capabilities to get rid of data restriction for PPS [42]. As shown in the top half of Figure 1, these approaches convert PPS into a language understanding task to predict the relevance among user history, query and product, which serves as a reranking accordance for a group of limited candidate products. Despite the progress, LLM-based PPS methods achieve sub-optimal search performance for PPS since they merely depend on textual contents, neglecting multimodal information which play a critical role for product search. Visual display figures can be regarded as a powerful complement for PPS especially in cases of inadequate textual descriptions.

Motivated by these observations, we propose a novel framework, HMPPS, for harnessing multimodal large language models to deal with PPS based on multimodal contents. As the bottom part of Figure 1 shows, we convert PPS into a multimodal language understanding task, which utilizes various multimodal contents to predict the relevance among user history, query and candidate product. The remarkable multimodal content understanding ability of pre-trained MLLM contributes to processing and reasoning among diverse multimodal contents, achieving a thorough comprehension for PPS. Leveraging HMPPS to rerank the search results filtered by existing ID-based methods, the entire search performance of PPS can be effectively enhanced.

Nevertheless, the redundancy and noise in PPS input stand for a great challenge to apply MLLM for PPS, which not only misleads MLLM to generate inaccurate search results but also increases the computation expense of MLLM. Useless information in product descriptions, *e.g.,* font settings in Figure 1, are unhelpful for MLLM to understand the relations between product and query, which may even result in mistakes and hallucination. Similarly, irrelevant purchased products in user history can cause mistakes in comprehending user preferences for current search. Except for the accuracy reduction, limited context size and high computation cost of MLLM make it a huge burden to process overlong sequences caused by redundancy.

To address these issues, we additionally design two query-aware refinement modules for product description and user history refinement: 1) a perspective-guided description summarization module that leverages an efficient LLM to obtain core perspectives relevant to search query and then summarize product descriptions around these perspectives, reflecting user search preferences; and 2) a two-stage training paradigm that trains HMPPS for two stages where the first stage is trained on random user history to implicitly learn the correlation among user history, query and candidate product. The second stage is trained on selected user history filtered by the relevance between query and product multimodal representations extracted by the first-stage model, capturing more precise user preferences. Both of these two modules contribute to not only improving the input robustness but also relieving the inference cost of HMPPS since the input size is decreased with the refined description and limited user history.

We conduct extensive experiments on four public datasets to demonstrate the effectiveness of HMPPS. The experimental results confirm the prominent advantage of HMPPS in enhancing the search performance of PPS. We also deploy HMPPS on an online search system with billion-level daily active users and achieve an evident gain in A/B testing, which validates the practicability of HMPPS. It is worth noting that, profiting from the remarkable generalization of MLLM, training HMPPS using small-scale pre-trained MLLMs (*e.g.,* InternVL2-1B[9]) on small-scale training samples (*e.g.,* 10% of the entire training set) can achieve an obvious improvement, proving the training efficiency of HMPPS.

The main contributions of our work can be summarized as follows:

- We propose a novel method, HMPPS, which utilizes pre-trained MLLMs to deal with PPS based on multimodal contents, enhancing the entire search performance of PPS.
- We design a perspective-guided description summarization module for HMPPS, utilizing LLM to generate refined

summaries around core perspectives relevant to search query, reducing the redundancy and noise in data.
- We design a two-stage training paradigm for HMPPS to obtain limited user history relevant with the input query and candidate product, which improves the search accuracy and decreases the inference cost of HMPPS.

2 Related Work

Personalized Product Search. Previous PPS solutions tend to convert user, product and query IDs to embedding vectors and then predict the relevance among them using various interaction modeling methods [1, 3–6, 10, 12, 16–18, 30, 35, 39]. Since ID-based methods heavily rely on dataset quality, inevitable data bias caused by the limited dataset collection results in insufficient learning of low frequency samples. As a result, several methods utilize content-based information to deal with these problems from semantic aspects [6, 15, 34], where LLM-based solutions exhibit significant superiority due to the remarkable content understanding and reasoning capabilities within LLMs [42]. Nevertheless, LLM-based methods merely depend on textual contents, neglecting multimodal information which play a critical role for PPS. The proposed HMPPS is targeted for addressing this issue by introducing MLLM to deal with PPS based on multimodal contents.

Some PPS methods [1, 11, 14, 21, 27, 28, 36] commit to conduct complicated user history modeling to extract precise user preferences, taking lifelong historical behaviors as model input, which is impractical for MLLM application with limited context size. Most relevant to our work is QIN [13], which designs a cascaded strategy to filter irrelevant historical products via product and query representation matching. However, off-the-shelf representation extractors in QIN cannot well adapt to PPS domain, weak in understanding the relations among product and query in search scenario. Our two-stage training paradigm not only extracts product and query representations using a powerful finetuned MLLM but also captures more precise relations between query and product based on multimodal information.

Large Language Models for Description Summarization. Large language models have emerged as powerful tools in the field of natural language processing, for which more and more methods propose to leverage LLMs for product description summarization to reduce noise and redundancy [25, 31, 37]. To prevent hallucinations in generated summaries, some methods [38], motivated by information factorization, leverage manually collected factors to guide LLM summarization, which results in high labor cost and limited generalization ability.

Different from these methods, the perspective-guided description summarization module in HMPPS leverages LLM to automatically extract core perspectives based on product descriptions and search queries, reflecting user search preferences. Additionally, we apply chain of thought and one-shot demonstration to ensure the reliability of the extracted perspectives and generated summaries.

3 Problem Formulation

PPS aims to retrieve products relevant to the given query considering user preferences within their purchase histories. Let U, P and Q denote the sets of users, products and queries, respectively. Each user u has a chronologically ordered purchase history which is composed of products $H_u = \{p_1^u, p_2^u, ..., p_{N_u}^u\}$, where N_u stands for the number of previously purchased products. The target of PPS is to predict the probability of u purchasing the candidate product $p_i \in P$ in response to the given query q based on his/her purchase history H_u:

$$y_{u,q,p_i} = \mathcal{F}_\theta\left(H_u, q, p_i\right), \tag{1}$$

where y_{u,q,p_i} indicates the purchase probability, and $\mathcal{F}$ refers to the PPS solution with learnable parameters θ.

The standard PPS output is a ranking list of all products in P according to their purchased probabilities obtained by Equation 1. However, HMPPS aims to capture exact but costly fine-grained relationships while ID-based methods can efficiently generate massive relevance scores with limited-size user, product and query embeddings. As a result, we decide to obtain top-K_p candidates $P' = \{p'_1, p'_2, ..., p'_{K_p}\}$, where $K_p \ll N_P$, N_P denotes the product number of P, based on the search results of an existing ID-based method M_{ID}. And HMPPS plays the role of reranker that predicts y_{u,q,p_i} only for $p_i \in P'$, which is a common practice in content-based PPS [6, 42]. This formulation combines the complementary advantages of HMPPS and ID-based methods, boosting the entire PPS performance.

4 Method

The overall framework of HMPPS is shown in Figure 2, which consists of three main components: 1) MLLM-based PPS. With instructed prompts, we convert PPS into a multimodal language understanding task and leverage MLLM to predict the relevance among user history, query and product based on multimodal contents, which comprises the search backbone of HMPPS; 2) perspective-guided description summarization. We firstly collect search-relevant perspectives based on product descriptions and search queries and then conduct summarization according to these limited perspectives with the help of LLM, which serves as one input refinement module for HMPPS; and 3) two-stage training paradigm. We firstly train MLLM with random user history to implicitly learn the correlation among user, query and product. The first-stage finetuned MLLM is then leveraged to select historical products relevant to search query and candidate product, serving for the second-stage training to earn a more accurate prediction. This paradigm can be regarded as the other input refinement module for HMPPS.

4.1 MLLM-based PPS

As shown in Figure 2, we design a specific template to transform the point-wise reranker of PPS into MLLM formulation by aggregating instruction text $Inst$, user history H_u, query q and candidate product p_i into an instructed prompt, which enforces the purchase decision output d to be either "yes" or "no". Since HMPPS concentrates on adequately mining valuable contents for PPS enhancement, we take various multimodal information to represent product, including textual product ID, title, brand, category, description and visual display figure. We leverage the general language generation loss

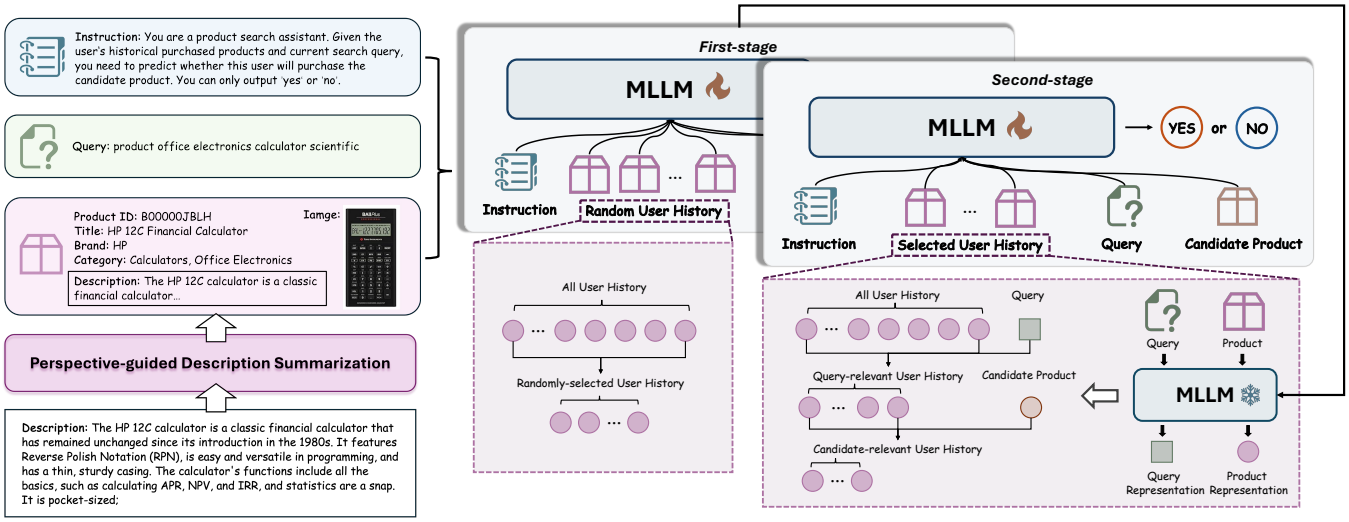

Figure 2: An overview of HMPPS. User history, query and candidate product are fed into MLLM to generate the search decision guided by the designed instruction based on abundant multimodal contents. The product description is additionally refined with a perspective-guided description summarization module. During training, for the first stage, the MLLM is trained to learn the relationship between query and user history implicitly, which is then leveraged to extract product and query representations for user history selection. For the second stage, this MLLM is further trained with selected query-relevant user history for more accurate prediction.

to finetune the pre-trained MLLM model, migrating its powerful understanding and reasoning ability to the PPS domain:

$$\mathcal{L} = -\sum_{l=1}^{L} \log \Pr(d_l | d_{<l}, [Inst; H_u; q; p_i]), \tag{2}$$

where d_l is the l-th word in the decision text, L is the length of the decision text, and $\Pr(\cdot)$ denotes the next word probability predicted by HMPPS.

Contrast learning based on positive and negative samples is proved to be an effective solution to boost the ranking accuracy [19]. As a result, for each search log $\langle H_u, q \rangle$, except for the target product, we randomly sample K_s^n candidate products from the entire product set P as simple negatives. Moreover, we obtain K_h^n hard negatives from the search result P' of existing ID-based method M_{ID} according to their predicted relevance scores. All positive and negative candidates participate in the HMPPS optimization of Equation 2 by assigning "yes" as the positive sample decision and "no" as the negative sample decision.

During the inference stage, since the target of PPS is to predict the relevance score instead of a discrete token, we intercept the probability distribution of the predicted next word and conduct a bidimensional softmax over the estimated scores corresponding to "yes" and "no" to obtain the final purchase probability y_{u,q,p_i}:

$$y_{u,q,p_i} = \frac{\exp(s_{yes})}{\exp(s_{yes}) + \exp(s_{no})}, \tag{3}$$

where s_{yes} and s_{no} refer to the estimated scores related to "yes" and "no" in the predicted probability distribution of the output word, respectively.

4.2 Perspective-guided Description Summarization

Factorization has been proved to be an efficient approach to reduce hallucination for information refinement [38], motivated by which we design a perspective-guided summarization module that utilizes LLM to summarize product descriptions with two relative prompts: 1) perspective extraction prompt that finds out perspectives that users concern for product search by exploring the relationships between the product and search query; and 2) summary generation prompt that instructs LLM to summarize product descriptions concentrating on the search-relevant perspectives collected by the previous step. Except for hallucination reduction, by introducing search-relevant perspectives, the generated summaries can reflect user search preferences more accurately, pandering to the PPS target.

4.2.1 Perspective Extraction Prompt. In-context learning is an effective paradigm to improve LLM generation performance with a few demonstrations composed by example inputs and outputs [7]. Therefore, except for the task instruction, we append one demonstration to the perspective extraction prompt to constraint the LLM output and boost the generation quality.

After collecting perspectives from all samples, we retain the top-K_d frequent perspectives as core perspectives that users concern most during product search. Since user search preferences vary in different scenarios, we provide demonstration and collect perspectives respectively for each dataset.

4.2.2 Summary Generation Prompt. We employ LLM to summarize product descriptions centering on the collected core perspectives to obtain refined descriptions which are more correlative with

user search preferences. Since description summarization is a challenging language processing task, we utilize chain of thought to reduce LLM hallucination [33], enhancing robustness of the generated summary. To be specific, except for summary generation, we explicitly include a reasoning demand in task instruction to enforce LLM to execute the reasoning procedure before outputting summarization result. One-shot demonstration is also available here to help LLM better understand the complicated summarization task.

The summarized descriptions are regarded as the product description of MLLM input to take part in the PPS prediction in section 4.1. Even though the summarization module involves LLM inference for two steps, it can be conducted off-line for only once and the summarized results can be saved for future use, which yeilds immeasurable benefits with negligible expense.

4.3 Two-stage Training Paradigm

User preferences tend to vary for different query and candidate product [1]. To capture exact user preferences, it is necessary to stand out correlative purchased products from the entire user history exclusively with the given query and candidate product. However, due to the limited context size and high computation cost, it is impractical to feed the entire set of historical products, together with query and candidate product, into MLLM to directly capture prominent purchased products. Consequently, we propose a two-stage training paradigm to efficiently extract precise user history.

In the first stage, to cover the lifelong sequential user history with MLLM of limited input size, we randomly select K_{s_1} chronological purchased products from the entire user history H_u, which subsequently take part in the MLLM training with the PPS optimization target in Section 4.1. The transformer architecture of MLLM is adept at capturing dynamic fine-grained relationships among user history, query and candidate product, assigning higher attention to relevant historical products and lower attention to irrelevant historical products. Thus, after the first-stage optimization, the finetuned MLLM has learned the implicit relevance between products and queries for the specific search scenario. We then apply this MLLM as encoder to extract multimodal representations of each product and query by averaging embeddings of all output tokens:

$$f_p = \text{Avg}\left(\mathcal{F}_{MLLM}\left([t_p; v_p]; \theta^{s_1}\right)\right), \tag{4}$$

$$f_q = \text{Avg}\left(\mathcal{F}_{MLLM}\left(t_q; \theta^{s_1}\right)\right), \tag{5}$$

where Avg denotes the average operation on embeddings, $\mathcal{F}_{MLLM}$ denotes the MLLM architecture with the first-stage learned parameters θ^{s_1}, t_p is the combination of product textual information, including product title, brand, category and description, v_p is the product display figure and t_q refers to the query. Cosine similarity is calculated based on the extracted representations to obtain the product-query and product-product relevance:

$$\begin{aligned} r_{p_j^u x} &= cos(f_{p_j^u}, f_x) \\ &= \frac{f_{p_j^u} \cdot f_x}{\|f_{p_j^u}\| \|f_x\|}, x \in \{q, p_i\}, \end{aligned} \tag{6}$$

where $cos(\cdot)$ denotes the cosine similarity calculation between vectors and $\|\|$ denotes the euclidean norm of vectors.

According to the relevance $r_{p_j^u q}$ between historical product and query, we firstly collect $2K_{s_2}$ products from the entire historical product set H_u to construct a query-relevant user history set H'_u. Then we pick out K_{s_2} products from H'_u according to $r_{p_j^u p_i}$ as the final selected user history set H''_u, which takes the relevance between historical product and candidate product into consideration.

We take H''_u as the user history input to further train the MLLM for the second stage of HMPPS, which learns more precise relationships among the relevant user history, query and candidate product. On one hand, a more accurate ranking result can be generated due to the irrelevant historical products have been filtered out. On the other hand, the product number of user history has been decreased for $K_{s_2} < K_{s_1} < N_u$, which reduces the final inference cost of HMPPS.

During inference, since product and query representations have been extracted by the first-stage MLLM and saved in advance, user history selection can be directly conducted via calculating representation similarity. The selected historical products are then regarded as user history to feed the second-stage MLLM to generate the final search decision.

5 Experiments

We conduct extensive experiments on multiple public available datasets to prove the effectiveness of HMPPS. Specifically, our experiments are intended to answer the following research questions (RQs):

- **RQ1** Can HMPPS improve the entire search performance for PPS by reranking the candidate products filtered by existing ID-based methods?
- **RQ2** Does HMPPS outperform other content-based PPS solutions?
- **RQ3** Does each component of HMPPS take effect?
- **RQ4** Does HMPPS have the potential for boosting PPS performance with different training scales?
- **RQ5** How does HMPPS make up for the limitations of existing PPS approaches?
- **RQ6** Can HMPPS benefit the real-world online search system?

5.1 Experimental Setup

5.1.1 Datasets and Evaluation Metrics. We take 5-core Amazon product search dataset [20] as our experimental corpus. Following previous PPS works[1, 3, 4], we select four diverse subsets of the Amazon dataset to conduct experiments, which are *Office Products* (Office), *Cell Phones & Accessories* (Cell), *Beauty* and *Sports & Outdoors* (Sports). To evaluate the PPS performance of HMPPS, we utilize three typical search metrics which are Mean Reciprocal Rank (MRR), Normalized Discounted Cumulative Gain (NDCG) and Recall.

5.1.2 Implementation Details. We take InternVL2-1B [9] as the MLLM backbone. The size of reranking candidate product set K_p is set to 10. We utilize Qwen2.5-14B [40] for description summarization and the number of core perspectives K_d is set as 20. The ID-based method M_{ID} is UniSAR, generating the basic search results for HMPPS reranking. The number of simple negative samples K_s^n is 2 and the number of hard negatives K_h^n is 3. The

Table 1: Comparison results of PPS performance improvement by utilizing HMPPS to rerank the candidate products filtered by different ID-based PPS methods on all datasets. Here, M, N and R denote the metrics of MRR, NDCG and Recall, respectively, Rel.Impr refers to the relative improvement rate of HMPPS against the basic search approaches among all metrics, and the best result is in bold.

Model	Office Products				Cell Phones & Accessories				Beauty				Sports & Outdoors			
	M@8	N@4	R@4	R@1	M@8	N@4	R@4	R@1	M@8	N@4	R@4	R@1	M@8	N@4	R@4	R@1
HEM	0.243	0.239	0.344	0.114	0.081	0.078	0.109	0.040	0.137	0.132	0.180	0.073	0.169	0.167	0.227	0.096
+ HMPPS	**0.288**	**0.292**	**0.392**	**0.169**	**0.101**	**0.104**	**0.136**	**0.062**	**0.155**	**0.157**	**0.208**	**0.091**	**0.185**	**0.192**	**0.257**	**0.111**
Rel.Impr	18.52%	22.18%	13.95%	48.25%	24.69%	33.33%	24.77%	55.00%	13.14%	18.94%	15.56%	24.66%	9.47%	14.97%	13.22%	15.62%
ZAM	0.222	0.216	0.303	0.111	0.069	0.065	0.092	0.033	0.101	0.098	0.135	0.052	0.103	0.101	0.139	0.056
+ HMPPS	**0.274**	**0.280**	**0.372**	**0.163**	**0.097**	**0.100**	**0.128**	**0.063**	**0.134**	**0.137**	**0.175**	**0.086**	**0.137**	**0.144**	**0.188**	**0.089**
Rel.Impr	23.42%	29.63%	22.77%	46.85%	40.58%	53.85%	39.13%	90.91%	32.67%	39.80%	29.63%	65.38%	33.01%	42.57%	35.25%	58.93%
DREM	0.193	0.186	0.265	0.093	0.095	0.091	0.125	0.050	0.135	0.131	0.182	0.070	0.163	0.161	0.218	0.092
+ HMPPS	**0.240**	**0.246**	**0.330**	**0.141**	**0.106**	**0.108**	**0.145**	**0.061**	**0.157**	**0.160**	**0.213**	**0.093**	**0.182**	**0.189**	**0.253**	**0.111**
Rel.Impr	24.35%	32.26%	24.53%	51.61%	11.58%	18.68%	16.00%	22.00%	16.30%	22.14%	17.03%	32.86%	11.66%	17.39%	16.06%	20.65%
DREM-HGN	0.241	0.232	0.314	0.130	0.066	0.062	0.086	0.034	0.141	0.136	0.183	0.077	0.120	0.117	0.164	0.062
+ HMPPS	**0.262**	**0.269**	**0.365**	**0.147**	**0.080**	**0.081**	**0.106**	**0.049**	**0.161**	**0.163**	**0.215**	**0.099**	**0.154**	**0.160**	**0.211**	**0.098**
Rel.Impr	8.71%	15.95%	16.24%	13.08%	21.21%	30.65%	23.26%	44.12%	14.18%	19.85%	17.49%	28.57%	28.33%	36.75%	28.66%	58.06%
CAMI	0.244	0.239	0.330	0.129	0.088	0.086	0.121	0.045	0.130	0.125	0.172	0.068	0.170	0.168	0.227	0.096
+ HMPPS	**0.268**	**0.271**	**0.366**	**0.154**	**0.103**	**0.104**	**0.136**	**0.063**	**0.156**	**0.159**	**0.210**	**0.095**	**0.189**	**0.196**	**0.264**	**0.113**
Rel.Impr	9.84%	13.39%	10.91%	19.38%	17.05%	20.93%	12.40%	40.00%	20.00%	27.20%	22.09%	39.71%	11.18%	16.67%	16.30%	17.71%
UniSAR	0.368	0.370	0.475	0.232	0.155	0.152	0.205	0.086	0.130	0.125	0.173	0.064	0.181	0.173	0.211	0.128
+ HMPPS	**0.398**	**0.414**	**0.532**	**0.256**	**0.190**	**0.194**	**0.249**	**0.124**	**0.184**	**0.188**	**0.238**	**0.121**	**0.198**	**0.207**	**0.269**	**0.130**
Rel.Impr	8.15%	11.89%	12.00%	10.34%	22.58%	27.63%	21.46%	44.19%	41.54%	50.40%	37.57 %	89.06%	9.39%	19.65%	27.49%	1.56%

Table 2: Comparison results of PPS performance of HMPPS *vs.* other content-based PPS methods on all datasets.

Model	Office Products				Cell Phones & Accessories				Beauty				Sports & Outdoors			
	M@8	N@4	R@4	R@1	M@8	N@4	R@4	R@1	M@8	N@4	R@4	R@1	M@8	N@4	R@4	R@1
RTM	0.371	0.382	0.499	0.228	0.130	0.130	0.189	0.058	0.143	0.146	0.204	0.074	0.162	0.169	0.239	0.088
InstructRec	0.382	0.398	0.521	0.239	0.184	0.187	0.241	0.117	0.175	0.180	0.232	0.110	0.171	0.178	0.244	0.098
HMPPS	**0.398**	**0.414**	**0.532**	**0.256**	**0.190**	**0.194**	**0.249**	**0.124**	**0.184**	**0.188**	**0.238**	**0.121**	**0.198**	**0.207**	**0.269**	**0.130**

product number K_{s_1} of randomly selected user history is set as {5, 7, 7, 7} for Office, Cell, Beauty and Sports datasets, respectively. The product number K_{s_2} of user history selected based on the relevance is set as {2, 3, 3, 3} for the second-stage of HMPPS. Low-rank adaption (LoRA) is adopted for parameter-efficient finetuning MLLM. Benefitting from the remarkable generalization of MLLM, we train HMPPS on merely 10% training data.

5.2 Performance Comparison (RQ1 & RQ2)

To prove the effectiveness of HMPPS for PPS, we integrate it with six different representative PPS solutions to rerank their search results. The evaluation results are shown in Table 1, which reveals the following observations: 1) HMPPS can effectively utilize MLLM to rerank the candidate products filtered by existing PPS models based on multimodal contents, resulting in obvious improvement and enhancing the entire search performance for PPS; and 2) HMPPS can adapt to various types of existing PPS models. The obvious enhancement among all metrics for all existing models verifies the generalization of HMPPS.

To verify the superiority of HMPPS in PPS content understanding, we compare it with other content-based solutions to rerank the candidate products filtered by UniSAR. As Table 2 shows, HMPPS performs better in content-based reranking compared to conventional and LLM-based solutions, proving the advantage of HMPPS in content understanding for PPS. It is worth noting that InstructRec is trained on the full dataset based on a 3B language model while HMPPS is implemented based on a 1B model and trained for merely 10% training data. The positive experimental results demonstrate that, except for the accuracy improvement, HMPPS can effectively alleviate the severe dependency on data and lead to an efficient training procedure.

In addition, as shown in Table 3, HMPPS outperforms QIN even though the basic retriever exhibits inferiority, which validates the effectiveness of HMPPS in capturing valid user history with the two-stage training paradigm.

5.3 Ablation Study (RQ3)

5.3.1 MLLM-based PPS. To verify the advantage of utilizing MLLM for PPS based on multimodal contents, we compare the results of taking only textual information as input with those including visual figures. To reduce distraction, we keep the original textual descriptions without summarization and train HMPPS only for the first stage in this experiment.

Table 3: Comparison results of PPS performance of HMPPS *vs.* user history selection PPS method, QIN, on Beauty dataset. Here, HMPPS* stands for HMPPS trained on the full dataset.

Model	MRR@8	NDCG@4	Recall@4
QIN	0.222	0.231	0.321
HMPPS	0.234	0.238	0.315
HMPPS*	**0.257**	**0.263**	**0.340**

Table 4: Ablation results of HMPPS applying contents of different modalities as input on Office Products dataset.

Content Type	Training	M@8	N@4	R@4	R@1
Text	Zero-shot	0.203	0.189	0.287	0.074
Text + Vision	Zero-shot	0.220	0.208	0.310	0.087
Text	Finetuned	0.376	0.391	0.514	0.231
Text + Vision	Finetuned	0.386	0.401	0.524	0.240

As shown in Table 4, no matter whether the backbone model is finetuned or not, it is superior to apply the multimodal combination of textual and visual contents as HMPPS input, proving that multimodal contents contribute to search information complementary for PPS. In addition, finetuned models consistently perform better than those without training, which validates that it is necessary to finetune MLLM on specific search corpus since PPS is a challenging task that requires knowledge migration.

5.3.2 Perspective-guided Description Summarization. To explore the most appropriate summarization strategy, we design the other two types of prompts for comparison: 1) direct summarization prompt that instructs LLM to generate description summary directly without any demonstration and reasoning request; and 2) reasoning-based summarization prompt that requires LLM to reason before generating the final summary with one-shot demonstration, which can be seen as a simple version of perspective-guided summarization without an explicit declaration for specific perspectives. The comparison results among the original and all three types of summarized descriptions are shown in Table 5.

The experimental results lead to three observations: 1) utilizing LLM to refine product descriptions into decreased words will not negatively influence or even boost the PPS performance of HMPPS, which proves that product descriptions indeed contain redundancy and noise; 2) even though reasoning-based summarization prompt achieves the smallest average word count, its summarization result is not stable which results in the largest max word count and worst PPS performance. Comparing to direct summarization prompt, the reasoning instruction increases the generation complexity for LLM while it lacks explicit perspectives that constrain the LLM output in perspective-guided summarization; and 3) perspective-guided summarization prompt achieves the best PPS performance due to the reasonable summarization procedure around search-relevant perspectives with an acceptable processing cost, which becomes the final choice for description summarization.

Table 5: Ablation results of HMPPS applying different description summarization strategies on Office Products dataset. Here, WC_p denotes the description word count of one product and WC_h denotes the user history word count of one sample composed of multiple products.

Sum Prompt	WC_p		WC_h		N@4	R@4
	Avg	Max	Avg	Max		
None	106	3538	651	5507	0.401	0.524
Direct	38	**88**	281	**557**	0.401	0.525
Reasoning-based	**34**	1513	**271**	2103	0.400	0.522
Perspective-guided	41	513	287	932	**0.403**	**0.529**

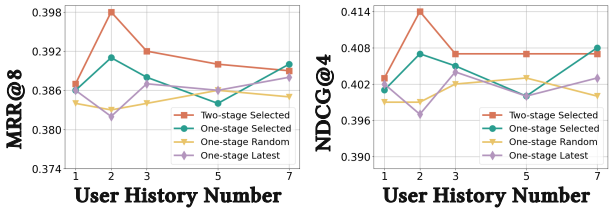

Figure 3: Ablation results of HMPPS with different types of user history on Office Products dataset.

5.3.3 User History Selection by Two-stage Training Paradigm. To demonstrate the effectiveness of the two-stage training paradigm for user history selection, we compare the search results of four variants: 1) **One-stage Latest** takes the latest purchased products as user history input to train HMPPS for only one stage; 2) **One-stage Random** randomly selects purchased products as user history; 3) **One-stage Selected** utilizes the finetuned MLLM of One-stage Random to select user history related to the query and candidate product and then train HMPPS based on the selected user history only for one stage, similar to the previous two variants; and 4) **Two-stage Selected** applies the selected user history extracted by the finetuned MLLM of One-stage Random and further trains the MLLM for the second stage of HMPPS.

From the experimental results shown in Figure 3, we can obtain the following observations: 1) even though without the first-stage learned parameters, training HMPPS with only two selected historical products can surpass both the best results of five random historical products and nine latest historical products. This demonstrates that the finetuned MLLM is effective for capturing exact user history relevant to query and candidate product. It not only enhances the PPS performance but also reduces the computation burden of the second-stage training, where the inference speed quadruples with decreased history size; and 2) further training HMPPS with two selected historical products based on the first-stage learned parameters achieves the best performance. This not only verifies the history selection accuracy of the finetuned MLLM but also the effectiveness of the two-stage training paradigm in enhancing the robustness of HMPPS, due to valid user history selection, to some extent, can be regarded as data augmentation.

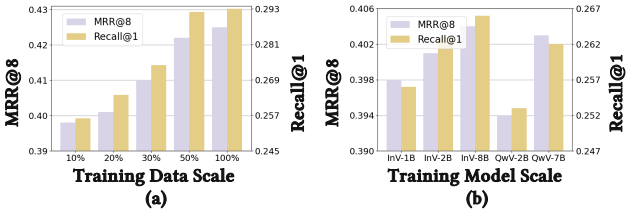

Figure 4: Ablation results of HMPPS trained on different scales of training data and models on Office Products dataset. Here, InV and QwV refer to MLLMs of InternVL2 and Qwen2VL.

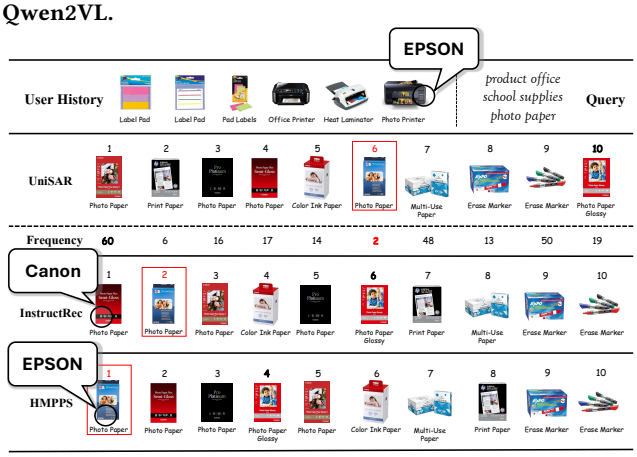

Figure 5: Example PPS results of UniSAR employing HMPPS *vs.* InstructRec to rerank its search results on Office Products dataset. Here, Frequency denotes the occurance frequency of the corresponding product in the training dataset and the target product is labeled by a red box.

5.4 Training Scale (RQ4)

5.4.1 Training Data Scale. To further explore the potential of HMPPS, we experiment on different data scales for HMPPS training, which are 10%, 20%, 30%, 50% and 100%. From the experimental results in Figure 4 (a), we can observe that HMPPS can definitely perform better by training on more data for almost all metrics. There exists significant improvement from using 10% to 100% training data, which reveals the enormous potential of HMPPS for boosting PPS.

5.4.2 Training Model Scale. One important practice of migrating MLLM to downstream tasks is to implement the proposed approaches with MLLM of different model scales. Larger-scale MLLM always stands for more powerful understanding, reasoning and generation capabilities that contribute to performance improvement. To explore the potential of HMPPS with MLLMs of different model scales, we experiment on applying different scales of InternVL2 as the MLLM backbone of HMPPS. We additionally take Qwen2VL [32] as the HMPPS MLLM backbone to validate the university of HMPPS. The evaluation result is shown in Figure 4 (b), which uncovers that larger-scale MLLMs, no matter which types they are, can indeed lead to improvement for HMPPS due to their remarkable progress in multimodal content understanding and reasoning with more well-learned parameters.

5.5 Case Study (RQ5)

We provide case study in Figure 5 to qualitatively illustrate the advantage of HMPPS. To demonstrate that HMPPS can effectively make up for the limitations of ID-based approaches and outperforms LLM-based methods in content understanding, we compare HMPPS results with those of ID-based UniSAR and LLM-based InstructRec and we can observe that: 1) UniSAR ranks the most frequent product as the first while the target one with the lowest frequency are ranked sixth. However, HMPPS can successfully capture the target product which validates that HMPPS can relieve the data bias problem of ID-based methods; and 2) for hard cases like similar candidates, HMPPS performs better than InstructRec since it can find out fine-grained differences according to multimodal contents, *e.g.,* brand information absent in textual contents but showed in visual figures.

5.6 Online Evaluation (RQ6)

To demonstrate the practicability of HMPPS, we conduct A/B testing on an online search system that boasts billion-level daily active users. This specific practice involves the following three procedures: 1) we firstly train HMPPS based on the offline data with a powerful large-scale MLLM; 2) the trained model is then distilled into a small-scale variant for online application; and 3) since the online system is composed of multiple processing modules, we utilize the distilled model to predict the search probability, serving as one of the factors of the final search result.

During A/B testing, we replace the search probability predicted by a conventional multimodal transformer, which resembles Bert4Rec [29], with that extracted by HMPPS. The online experiment, conducted over a span of 14 days, yields a 0.53% gain for query-ctr and a 0.77% increase in efficient click count with p-value = 1.16%, which demonstrates a significant improvement for highly-optimized real-world systems. Here, query-ctr assesses whether users click on the items returned by the online system in response to their search queries. Meanwhile, the efficient click count quantifies the number of items that users have viewed for more than 5 seconds.

The inference time of ranking 10 candidate items for a query is 22 microseconds in average. Since HMPPS is targeted for reranking, which just ranks a handful of retrieved candidates for more accurate and finegrained search result, its latency and computation expense can be acceptable for online application.

6 Conclusion

In this paper, to address the limitations of LLM-based approaches in PPS reranking, we proposed a novel method, HMPPS, harnessing pre-trained MLLMs to deal with PPS based on multimodal contents. Except for adapting MLLM to PPS by converting the search task into a multimodal language understanding problem, we designed two query-aware refinement modules to reduce the redundancy in PPS input, which is a perspective-guided summarization module for product description refinement and a two-stage training paradigm for user history selection. Both of these two modules improve the prediction accuracy and reduce the computation cost of HMPPS. Extensive experiments were conducted on four datasets to demonstrate the effectiveness of HMPPS and the evident gain in online A/B testing also validated the practicability of HMPPS.

Acknowledgments

This work is supported by the National Science Foundation of China (62072232), the Collaborative Innovation Center of Novel Software Technology and Industrialization, and the Young Elite Scientists Sponsorship Program by CAST (2023QNRC001).

References

[1] Qingyao Ai, Daniel N Hill, SVN Vishwanathan, and W Bruce Croft. 2019. A zero attention model for personalized product search. In *Proceedings of the ACM International Conference on Information and Knowledge Management*. 379–388.

[2] Qingyao Ai and Lakshmi Narayanan. R. 2021. Model-agnostic vs. model-intrinsic interpretability for explainable product search. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 5–15.

[3] Qingyao Ai, Yongfeng Zhang, Keping Bi, Xu Chen, and W Bruce Croft. 2017. Learning a hierarchical embedding model for personalized product search. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 645–654.

[4] Qingyao Ai, Yongfeng Zhang, Keping Bi, and W Bruce Croft. 2019. Explainable product search with a dynamic relation embedding model. In *ACM Transactions on Information Systems*, Vol. 38. 1–29.

[5] Keping Bi, Qingyao Ai, and W Bruce Croft. 2020. A transformer-based embedding model for personalized product search. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1521–1524.

[6] Keping Bi, Qingyao Ai, and W Bruce Croft. 2021. Learning a fine-grained review-based transformer model for personalized product search. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 123–132.

[7] Alexander Brinkmann, Roee Shraga, Reng Chiz Der, and Christian Bizer. 2023. Product Information Extraction using ChatGPT. In *arXiv preprint arXiv:2306.14921*.

[8] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. 2023. Bias and debias in recommender system: A survey and future directions. In *ACM Transactions on Information Systems*, Vol. 41. 1–39.

[9] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 24185–24198.

[10] Dian Cheng, Jiawei Chen, Wenjun Peng, Wenqin Ye, Fuyu Lv, Tao Zhuang, Xiaoyi Zeng, and Xiangnan He. 2022. Ihgnn: Interactive hypergraph neural network for personalized product search. In *Proceedings of the ACM Web Conference*. 256–265.

[11] Shitong Dai, Jiongnan Liu, Zhicheng Dou, Haonan Wang, Lin Liu, Bo Long, and Ji-Rong Wen. 2023. Contrastive Learning for User Sequence Representation in Personalized Product Search. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 380–389.

[12] Lu Fan, Qimai Li, Bo Liu, Xiao-Ming Wu, Xiaotong Zhang, Fuyu Lv, Guli Lin, Sen Li, Taiwei Jin, and Keping Yang. 2022. Modeling user behavior with graph convolution for personalized product search. In *Proceedings of the ACM Web Conference*. 203–212.

[13] Tong Guo, Xuanping Li, Haitao Yang, Xiao Liang, Yong Yuan, Jingyou Hou, Bingqing Ke, Chao Zhang, Junlin He, Shunyu Zhang, et al. 2023. Query-dominant User Interest Network for Large-Scale Search Ranking. In *Proceedings of the ACM International Conference on Information and Knowledge Management*. 629–638.

[14] Yangyang Guo, Zhiyong Cheng, Liqiang Nie, Yinglong Wang, Jun Ma, and Mohan Kankanhalli. [n. d.]. Attentive long short-term preference modeling for personalized product search. In *ACM Transactions on Information Systems*, Vol. 37. 1–27.

[15] Yangyang Guo, Zhiyong Cheng, Liqiang Nie, Xin-Shun Xu, and Mohan Kankanhalli. 2018. Multi-modal preference modeling for product search. In *Proceedings of the ACM international conference on Multimedia*. 1865–1873.

[16] Jyun-Yu Jiang, Tao Wu, Georgios Roumpos, Heng-Tze Cheng, Xinyang Yi, Ed Chi, Harish Ganapathy, Nitin Jindal, Pei Cao, and Wei Wang. 2020. End-to-end deep attentive personalized item retrieval for online content-sharing platforms. In *Proceedings of the ACM Web Conference*. 2870–2877.

[17] Jiongnan Liu, Zhicheng Dou, Qiannan Zhu, and Ji-Rong Wen. 2022. A category-aware multi-interest model for personalized product search. In *Proceedings of the ACM Web Conference*. 360–368.

[18] Shang Liu, Wanli Gu, Gao Cong, and Fuzheng Zhang. 2020. Structural relationship representation learning with graph embedding for personalized product search. In *Proceedings of the ACM International Conference on Information and Knowledge Management*. 915–924.

[19] Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2024. Fine-Tuning LLaMA for Multi-Stage Text Retrieval. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2421–2425.

[20] Julian McAuley, Rahul Pandey, and Jure Leskovec. 2015. Inferring networks of substitutable and complementary products. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 785–794.

[21] Yaoxin Pan, Shangsong Liang, Jiaxin Ren, Zaiqiao Meng, and Qiang Zhang. 2021. Personalized, sequential, attentive, metric-aware product search. *ACM Transactions on Information Systems* 40, 2, 1–29.

[22] Qi Pi, Weijie Bian, Guorui Zhou, Xiaoqiang Zhu, and Kun Gai. 2019. Practice on long sequential user behavior modeling for click-through rate prediction. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2671–2679.

[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 *Proceedings of the ACM International Conference on Information and Knowledge Management*. 2685–2692.

[24] Kan Ren, Jiarui Qin, Yuchen Fang, Weinan Zhang, Lei Zheng, Weijie Bian, Guorui Zhou, Jian Xu, Yong Yu, Xiaoqiang Zhu, et al. 2019. Lifelong sequential modeling with personalized memorization for user response prediction. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 565–574.

[25] 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*. 3464–3475.

[26] Kaiming Shen, Xichen Ding, Zixiang Zheng, Yuqi Gong, Qianqian Li, Zhongyi Liu, and Guannan Zhang. 2024. SEMINAR: Search Enhanced Multi-modal Interest Network and Approximate Retrieval for Lifelong Sequential Recommendation. In *arXiv preprint arXiv:2407.10714*.

[27] Qijie Shen, Hong Wen, Jing Zhang, and Qi Rao. 2022. Hierarchically fusing long and short-term user interests for click-through rate prediction in product search. In *Proceedings of the ACM International Conference on Information and Knowledge Management*. 1767–1776.

[28] Teng Shi, Zihua Si, Jun Xu, Xiao Zhang, Xiaoxue Zang, Kai Zheng, Dewei Leng, Yanan Niu, and Yang Song. 2024. UniSAR: Modeling User Transition Behaviors between Search and Recommendation. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1029–1039.

[29] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. BERT4Rec: Sequential Recommendation with Bidirectional Encoder Representations from Transformer. In *Proceedings of the ACM International Conference on Information and Knowledge Management*. 1441–1450.

[30] Thibaut Thonet, Jean-Michel Renders, Mario Choi, and Jinho Kim. 2022. Joint Personalized Search and Recommendation with Hypergraph Convolutional Networks. *Advances in Information Retrieval* 13185, 443–456.

[31] Ghazaleh Haratinezhad Torbati, Anna Tigunova, Andrew Yates, and Gerhard Weikum. 2023. Recommendations by Concise User Profiles from Review Text. In *arXiv preprint arXiv:2311.01314*.

[32] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. In *arXiv preprint arXiv:2409.12191*.

[33] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, Vol. 35. 24824–24837.

[34] Tianxin Wei, Bowen Jin, Ruirui Li, Hansi Zeng, Zhengyang Wang, Jianhui Sun, Qingyu Yin, Hanqing Lu, Suhang Wang, Jingrui He, and Xianfeng Tang. 2024. Towards Unified Multi-Modal Personalization: Large Vision-Language Models for Generative Recommendation and Beyond. In *Proceedings of the International Conference on Learning Representations*. 1–19.

[35] Bin Wu, Zaiqiao Meng, and Shangsong Liang. 2023. Dynamic Bayesian Contrastive Predictive Coding Model for Personalized Product Search. In *ACM Transactions on the Web*, Vol. 17. 1–31.

[36] Bin Wu, Zaiqiao Meng, Qiang Zhang, and Shangsong Liang. 2022. Meta-Learning Helps Personalized Product Search. In *Proceedings of the ACM Web Conference*. 2277–2287.

[37] Jiahao Wu, Qijiong Liu, Hengchang Hu, Wenqi Fan, Shengcai Liu, Qing Li, Xiao-Ming Wu, and Ke Tang. 2025. TF-DCon: Leveraging Large Language Models (LLMs) to Empower Training-Free Dataset Condensation for Content-Based Recommendation. In *arXiv preprint arXiv:2310.09874*.

[38] 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 ACM Conference on Recommender Systems*. 12–22.

[39] Teng Xiao, Jiaxin Ren, Zaiqiao Meng, Huan Sun, and Shangsong Liang. 2019. Dynamic bayesian metric learning for personalized product search. In *Proceedings of the ACM International Conference on Information and Knowledge Management*.

1693–1702.

[40] An Yang, Baosong Yang, Beichen Zhang, and et al. 2024. Qwen2.5 Technical Report. In *arXiv preprint arXiv:2412.15115*.

[41] Zheng Yuan, Fajie Yuan, Yu Song, Youhua Li, Junchen Fu, Fei Yang, Yunzhu Pan, and Yongxin Ni. 2023. Where to go next for recommender systems? id-vs. modality-based recommender models revisited. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2639–2649.

[42] Junjie Zhang, Ruobing Xie, Yupeng Hou, Xin Zhao, Leyu Lin, and Ji-Rong Wen. 2024. Recommendation as instruction following: A large language model empowered recommendation approach. *ACM Transactions on Information Systems* (2024).

[43] 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 *Proceedings of the AAAI conference on Artificial Intelligence*. 5941–5948.

[44] Guorui Zhou, Xiaoqiang Zhu, Chenru 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 *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1059–1068.

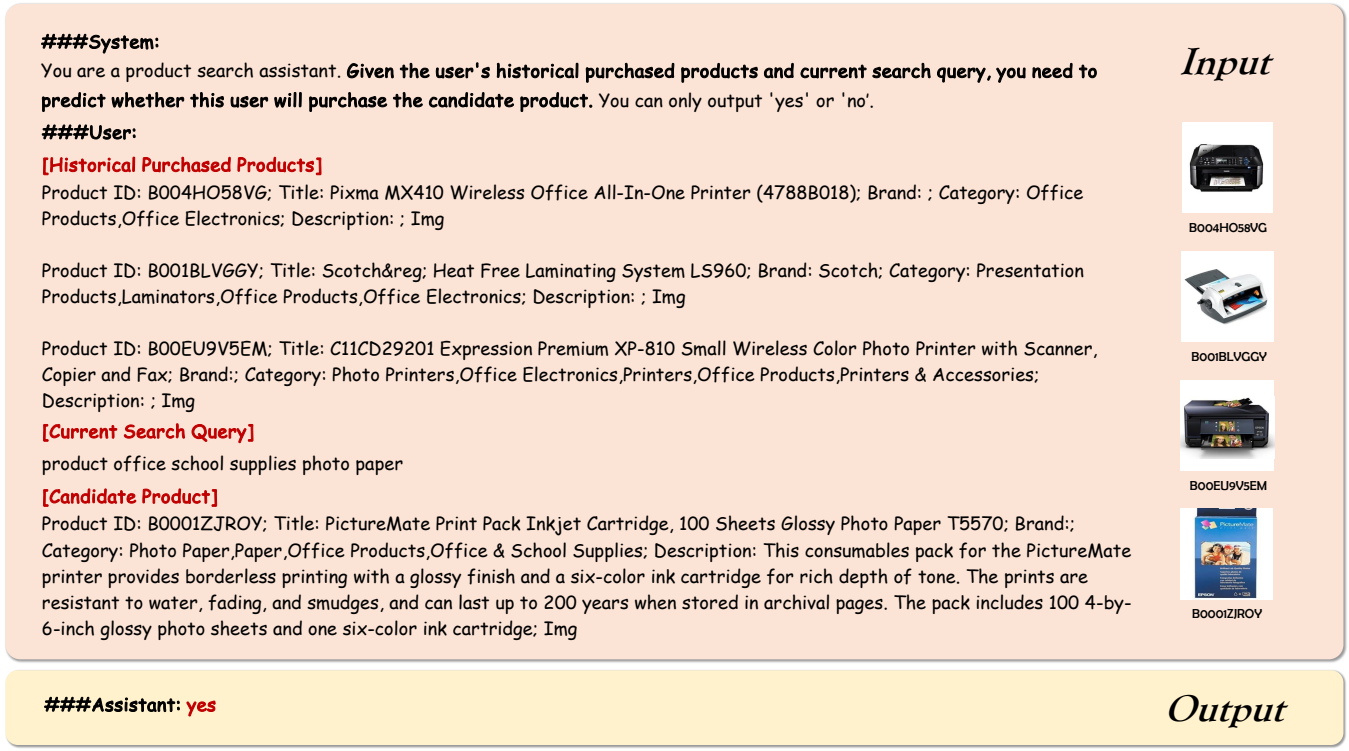

Figure 6: An example of MLLM-based personalized product search on Office Products dataset.

A Method Details

A.1 Example of HMPPS

Figure 6 showcases an example of the input and output of HMPPS, which applies MLLM to deal with PPS. The task instruction at the beginning of the input prompt requires the MLLM to understand the relationship among user historical products, query and candidate product and output the search decision. Each product involves multimodal contents where product ID, title, brand, category and description compose textual contents and visual displays make up for visual contents. The output result demonstrates that HMPPS can generate precise search decision that relates to the user history and query via conducting analysis based on multimodal contents of query and product.

A.2 Example of Perspective-guided Description Summarization

Figure 7 showcases an example of the perspective extraction, which is the first step in perspective-guided description summarization module. As the instruction requires extracting information perspectives of product description with the guide of user query, it can capture perspectives that relates to customer search preferences. Accurate perspectives are obtained with the help of the demonstration in prompt. After collecting perspectives from all product descriptions, we retain a fixed number of perspectives as core information perspectives according to their frequency.

Figure 8 showcases an example of the summary generation, which is the second step in perspective-guided description summarization module. Core perspectives obtained in the previous step are leveraged to guide the process of summarization generation. It can be observed that irrelevant noise, *e.g.*, text font information, in the original description have been filtered out after the summarization. Moreover, the generated summary not only includes precise product information but also meets the search preferences of customers as the summarization result is around the information perspectives that customers care about during product search.

A.3 Online Search System Implementation

Due to the computation cost and inference efficiency of MLLM, it is impractical for an online search system to leverage a large-scale powerful MLLM to obtain the search result. Therefore, we conduct the training process of HMPPS off-line with a large-scale MLLM of 72 billion parameters and then distill it to a small-scale variant of 300 million parameters. Since there are many other valuable modules contributing to the large and complicated online search system, it is insufficient to merely depend on HMPPS to make the search decision. As a result, we leverage the distilled model to obtain the search probability, which serves as a component participating in the final search result fusion for online system. It is noteworthy that we apply the proposed model only in reranking stage of the online search system that involves a handful of candidates filtered by the prepositive retrievers, for which the computation expense is acceptable.

To obtain a powerful search model, we collect 20 million samples from real user search logs of three months, covering 2 million active users and 8 million items, for HMPPS training. To ensure that MLLM can effectively adapt to PPS domain, we additionally

###System:
You are a product search assistant. To **better understand customers' search preferences**, you are required to find out which information perspectives a customer will care about when deciding to buy a product for a query. The input will be the **search query** and the purchased **product description**. And you need to output **information perspectives** of the description that leads to the buying.
For example, given {"Search Query": "product office electronics", "Product Description": "The BA II Plus Professional calculator features all the great features of its predecessor while packing in even more time-saving functions to make short work of complex equations..."}, you should output: {"Perspectives": "product function, product feature, target audience, professional use, build quality, additional accessories"}
###User:
{**"Search Query"**: "product office school envelope shipping supplies transit sack mail bags", **"Product Description":** "Perfect for packing a small apartment or dorm room! Convenient, all-in-one moving essentials kit including 8 small and 6 medium basic moving boxes, 4 rolls of bubble wrap, 55 yds. of packing tape and 1 permanent black marker. Tape required for box assembly "}

Input

###Assistant:
{**"Perspectives"**: "product function, product feature, product size, product material, product color "}

Output

Figure 7: An example of perspective extraction in perspective-guided description summarization from Office Products dataset.

###System:
You are a product search assistant. To **capture exact and useful information to match customers' search queries**, you are required to summarize product descriptions. The input will be the **description** of a product and key **perspectives** you need to focus on when summarizing the description. You need to output the **reasoning** for summarizing the product description from the aspect of key perspectives and the final **summary**.
For example, given {"Product Description": "The BA II Plus Professional calculator features all the great features of its predecessor while packing in even more time-saving functions to make short work of complex equations...", "Perspectives": "product feature, product function, ..., product safety"}, you should output: {"Reasoning": "The summary focuses on key features and functions of the BA II Plus Professional calculator, highlighting its...", "Summary": "The BA II Plus Professional calculator is designed for both entry-level and advanced users in finance, accounting, and related fields..."}
###User:
{**"Product Description"**: "{color:#676767;font-family:Arial,Verdana,Helvetica,sans-serif;font-size:9px;font-weight:normal;text-align:left;line-height:12px;} HP 75 TRI-COLOR INKJET PRINT CARTRIDGEIDEAL FOR:Everyday color printingPhotos that last for generations using HP Premium Plus Photo Paper1Yield2:170 pagesSTOCK UP ON HP 74/75 INKS:#1 BRAND IN INK IS ALSO THE #1 VALUEPRINT MORE. REPRINT LESS. Original HP Ink is proven to be more reliablePrint up to 65% more pages than bargain inks3More than 33% of tested refilled ink cartridges failed during use or right out of the ...", **"Perspectives":** "product feature, product function, ..., product safety "}

Input

Output

###Assistant:
{**"Reasoning"**: ".The summary focuses on the key features and functions of the HP 75 Tri-Color Inkjet Cartridge, highlighting its suitability for everyday color printing and its compatibility with HP Premium Plus Photo Paper. It emphasizes the cartridge's yield, stock options, and reliability, which are crucial for potential buyers. Additionally, it mentions the HP Planet Partners Recycling Program as an additional accessory, which adds value. The target audience is clearly defined, and the summary captures the product's durability and design aspects, which are important for users in demanding environments", **"Summary"**: " The HP 75 Tri-Color Inkjet Cartridge is designed for everyday color printing and is compatible with HP Premium Plus Photo Paper. It offers a yield of 170 pages and is available in stock options. The cartridge is proven to be more reliable than bargain inks and is designed with the environment in mind. It is part of the HP Planet Partners Recycling Program."}

Figure 8: An example of summary generation in perspective-guided description summarization on Office Products dataset.

conduct specific data mining based on the relevance between user history and query to improve the data quality.

B Experimental Details

B.1 Dataset details

We take 5-core Amazon product search dataset [20] as our experimental corpus, which contains product metadata and multimodal search logs from the Amazon website from May, 1996 to July, 2014. To verify that HMPPS can adapt to diverse search scenarios, we select four typical subsets of the Amazon dataset which are *Office Products* (Office), *Cell Phones & Accessories* (Cell), *Beauty* and *Sports & Outdoors* (Sports). Statistics of these datasets are shown in Table 6. Following previous PPS works [1, 3, 4], we split each dataset into train and test sets with a ratio of 7:3 and extract search queries using the same strategy as them.

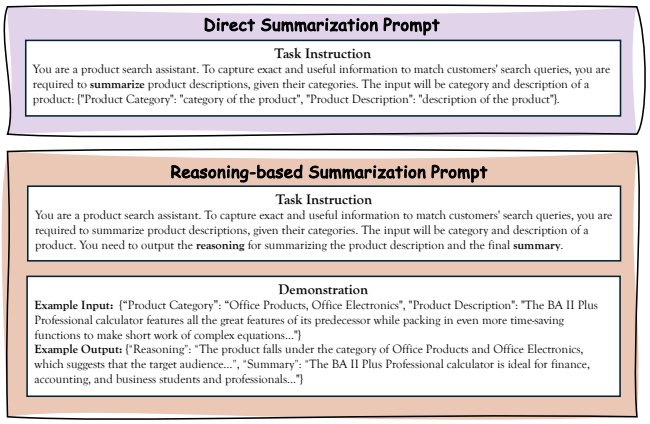

Figure 9: Example prompts for direct and reasoning-based description summarization.

Table 6: Data statistics of the four datasets used in all experiments.

Dataset	Office	Cell	Beauty	Sports
#users	4,905	27,879	22,363	35,598
#products	2,420	10,429	12,101	18,357
#queries	290	165	249	1,543
#interactions	53,258	194,439	198,502	296,337

B.2 Evaluation Metrics.

To evaluate the PPS performance of HMPPS, we utilize three typical search metrics which are Mean Reciprocal Rank (MRR), Normalized Discounted Cumulative Gain (NDCG) and Recall. MRR and NDCG denotes search precision, which is calculated based on the position of positive products in the result list, evaluating ranking abilities of algorithms. Recall corresponds to search coverage, concentrating on retrieval performance by calculating the ratio of positive products appearing in the returned products of search systems. In this paper, we report MRR at position 8, NDCG at position 4 and Recall at positions {4,1}. The higher values indicate better performance for all the metrics.

B.3 Existing PPS Models

HMPPS is aimed at reranking a limited number of candidate products filtered by existing PPS solutions. To verify the generalization of HMPPS, we choose six different, representative ID-based PPS approaches: 1) **HEM** [3] jointly learns distributed embeddings for queries, products and users with a deep neural network; 2) **ZAM** [1] utilizes an attention function over user purchased products to build user embeddings for product search; 3) **DREM** [4] constructs a knowledge graph to model the dynamic relationships between users and products in the latent space; 4) **DREM-HGN** [2] proposes to construct and train an intrinsic-explainable model for PPS with user-interaction data and knowledge graph; 5) **CAMI** [17] proposes a category-aware multi-interest model that learns multiple interest embeddings to encode diverse user preferences; and 6) **UniSAR** [28] is trained jointly on personalized search and recommendation tasks to enhance the user preference comprehension for PPS. We reproduce these methods based on their official codes to obtain the search results for HMPPS reranking.

To validate the superiority of HMPPS in content understanding, we compare it with multiple content-based approaches for PPS reranking. **RTM** [6] utilizes a transformer to encode query, user and item reviews to obtain fine-grained interactions for PPS. **InstructRec** [42] leverages LLM to reason the relationship between textual user history, query and candidate item. Since InstructRec aims to construct a general model for multiple recommendation and search tasks, it is trained on large amounts of synthesized data and evaluated underlying zero-shot setting. To keep fair comparison with HMPPS, which is trained with supervised setting, we reproduce the LLM-based PPS framework of InstructRec with a 3B language model and train the model on specific PPS datasets with its instruction tuning settings in its official report.

To validate the effectiveness of the two-stage training paradigm of HMPPS in user history selection, we also compare HMPPS with **QIN** [13], which utilizes recommendation behaviors to augment search history and design a cascaded strategy to select user history. It is noteworthy that, to keep fair comparison with QIN reports, we apply *leave-one-out* strategy to split data in the experiment.

B.4 Implementation Details.

We take InternVL2-1B [9] as the MLLM backbone of HMPPS for most of our experiments. We utilize Qwen2.5-14B [40] to conduct description summarization and the number of core perspectives K_d is set as 20 for all datasets. The ID-based method M_{ID}, which generates the basic search results for HMPPS reranking, is UniSAR for all experiments, except for that in Table 1 which validates that HMPPS can adapt to any ID-based solutions for performance improvement. The size of reranking candidate product set K_p is set to 10. We sample 5 negative samples for training, where the number of simple negative samples K_s^n is 2 and the number of hard negatives K_h^n is 3. The product number K_{s_1} of randomly selected user history for the first-stage of HMPPS is set as {5, 7, 7, 7} for Office, Cell, Beauty and Sports datasets, respectively. The product number K_{s_2} of user history selected based on the relevance to query and candidate product is set as {2, 3, 3, 3} for the second-stage of HMPPS.

Low-rank adaption (LoRA) is adopted for parameter-efficient finetuning of all components of MLLM, including the vision module, language module and the MLP layer. Benefitting from the remarkable generalization of MLLM, we train HMPPS on merely 10% training data with only 1 epoch, batch size 1 and learning rate 0.0001 using AdamW optimizer on all datasets.

B.5 Description Summarization Strategies

The specific prompts of the other two description summarization strategies in ablation study 5.3.2 are shown in Figure 9. The direct summarization prompt instructs LLM to generate description summary directly without any demonstration and reasoning request. The reasoning-based summarization prompt requires LLM to reason before generating the final summary with one-shot

demonstration, which can be seen as a simple version of perspective-guided summarization without an explicit declaration for specific perspectives.

C Other Related Work

User History Selection. Modeling user preference from their historical products helps improve product search and recommendation performance [22]. However, redundancy and noise in user history cause accuracy decrease and latency burden of online systems. Therefore, various works focus on efficient user history selection. [22, 24] leverage updatable, limited-size memory to store refined user history. [14, 43] apply sequence modeling networks like GRU and LSTM to summarize user interest from user history. [1, 13, 23, 44] calculate attention to restrain the influence of query-irrelevant products.

HMPPS essentially also calculates attention to suppress redundancy and noise. Nevertheless, via the two-stage training paradigm, HMPPS obtains a migrated MLLM to extract robust multimodal representations for a more comprehensive matching between query and historical products, which is superior to those relying on ID-based embeddings and frozen semantic representations [1, 13].