Similarity = Value? Consultation Value Assessment and Alignment for Personalized Search

Weicong Qin[1], Yi Xu[1], Weijie Yu[2]*, Teng Shi[1], Chenglei Shen[1],
Ming He[3]*, Jianping Fan[3], Xiao Zhang[1], Jun Xu[1]
[1]Gaoling School of Artificial Intelligence, Renmin University of China, China
[2]School of Information Technology and Management,
University of International Business and Economics, China
[3]AI Lab at Lenovo Research, Lenovo Group Limited, China
`{qwc, yixu00, shiteng, chengleishen9, zhangx89, junxu}@ruc.edu.cn,`
`yu@uibe.edu.cn, {heming4, jfan1}@lenovo.com`

Abstract

Personalized search systems in e-commerce platforms increasingly involve user interactions with AI assistants, where users consult about products, usage scenarios, and more. Leveraging consultation to personalize search services is trending. Existing methods typically rely on semantic similarity to align historical consultations with current queries due to the absence of 'value' labels, but we observe that semantic similarity alone often fails to capture the true value of consultation for personalization. To address this, we propose a consultation value assessment framework that evaluates historical consultations from three novel perspectives: (1) Scenario Scope Value, (2) Posterior Action Value, and (3) Time Decay Value. Based on this, we introduce VAPS, a value-aware personalized search model that selectively incorporates high-value consultations through a consultation–user action interaction module and an explicit objective that aligns consultations with user actions. Experiments on both public and commercial datasets show that VAPS consistently outperforms baselines in both retrieval and ranking tasks. Codes are available at `https://github.com/E-qin/VAPS`.

1 Introduction

Personalized product search (Ai et al., 2019a; Shi et al., 2024, 2025) is key for e-commerce and search engines, delivering tailored results by leveraging user data. As AI assistants become more common in e-commerce, users increasingly rely on bots for product queries, comparisons, and availability checks. These consultation logs provide valuable insights into user intent beyond explicit searches, helping improve personalization and search accuracy (Qin et al., 2025).

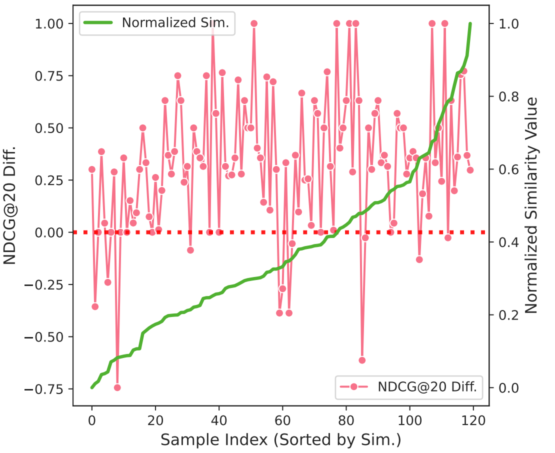


Figure 1: NDCG@20 difference with/without consultations vs. normalized semantic similarity, using MAPS (Qin et al., 2025) on Amazon dataset. Semantic similarity does not consistently reflect the value of consultations for improving search.

Recent studies can be broadly divided into two directions. The first mainly extracts user-specific preferences from user interactions (across multiple scenarios) to personalize search results (Bi et al., 2020), but lacks exploration of needs in user consultations. The second direction, represented by Motivation-Aware Personalized Search model (MAPS) (Qin et al., 2025), while utilizing user interaction data, primarily attempts to align current search queries with consultation histories via semantic similarity to tap into search needs within them. Due to the absence of 'value' labels, these methods assume that semantically related past consultations inherently benefit current searches. However, as shown in Fig. 1, our analysis of MAPS' search performance differences and consultation-query semantic similarity reveals: semantic similarity cannot fully reliably reflect search utility. Consultations with high similarity scores may harm ranking performance, while some with lower similarity can improve it, even though an overall positive correlation exists. This suggests that semantic alignment is an inadequate proxy for identifying

---

*Corresponding authors.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 9839–9852
November 4-9, 2025 ©2025 Association for Computational Linguistics

valuable consultations.

Inspired by the consumer journey (Hamilton and Price, 2019) in e-commerce, we argue that the value of a historical consultation should be assessed in the context of the user's evolving goals and actions. There are distinct consumer consultation-action patterns for users as illustrated in Fig. 2. (1) Out-of-scope consultations, such as those related to politics, are regarded as noise. (2) Out-of-date offers such as the 2023 spring sale discount code are no longer useful. (3) Interest in products like the color options of the "[Pad Tab] S7" lacks action verification as no related actions follow. In contrast, inquiries about the "[Laptop OG] G14" are accompanied by related actions, showing verified interest.

We identify three key factors affecting a consultation's value for personalized search: (1) Scenario Scope Value: whether the consultation pertains to the platform's product or service domain; (2) Time Decay Value: whether the consultation has expired, and (3) Posterior Action Value: whether the consultation is followed by user relevant actions (e.g., click, search, purchase) indicating its influence. For example, a timely inquiry that leads to a purchase conveys stronger search-relevant intent than an outdated or out-of-scope inquiry. However, it still faces the challenge of missing explicit labels.

To operationalize these insights and address the challenge, we propose a Consultation Value Assessment Framework that quantifies the value of historical consultations with empirical function-based assessment. Based on this, we introduce VAPS (Value-Aware Consultation-Enhanced Personalized Search) model that selectively incorporates high-value consultations with a consultation–action interaction module and explicitly aligns consultations with consumer action signals. We further design tailored loss functions to ensure our methods capturing the value-rich consultation-action signals. Extensive experiments on real and synthetic data show VAPS significantly outperforms baselines in both retrieval and ranking tasks.

To summarize, our contributions are as follows:

- We pioneer the value assessment of user-AI consultations in personalized search, exposing flaws in current semantic-similarity-only method.
- We propose a consultation value assessment framework with three novel dimensions (Scenario Scope, Time Decay, and Posterior Action) tailored for personalized search, addressing gaps in prior work.
- We introduce VAPS, a value-aware consultation-enhanced personalized search model designed to align with value assessment signals
- Extensive experiments across retrieval and ranking stages, conducted on both real-world commercial and synthetic datasets, demonstrate that VAPS outperforms existing personalized search methods, traditional/conversational retrieval approaches, and multi-scenario models.

2 Related Work

Personalized search provides relevant items based on user queries (Shi et al., 2024). Traditional methods like BM25 (Robertson et al., 2009) focus on word frequency, while dense retrieval (e.g., BGE-M3 (Chen et al., 2024)) uses embeddings. Conversational methods like CHIQ (Mo et al., 2024) improve accuracy using search history, but lack personalization.

Recent work includes QEM (Ai et al., 2019a) and DREM (Ai et al., 2019b) for query-item similarity, while HEM (Ai et al., 2017), AEM (Ai et al., 2019a), ZAM (Ai et al., 2019a), and TEM (Bi et al., 2020) incorporate user data. Multi-scenario methods combine search and recommendation: SESRec (Si et al., 2023) uses contrastive learning, UnifiedSSR (Xie et al., 2023) has dual-branch networks, and UniSAR (Shi et al., 2024) employs transformers. However, they ignores the value of consultation in enchancing search intent.

In e-commerce, Zeng et al. (2025) uses conversation history and product knowledge to improve responses. Ferreira et al. (2023) integrated conversational features with behavioral signals for dynamic weighting. Recently, consultation-enhanced personalized search methods have emerged, with MAPS (Qin et al., 2025) as a representative approach that mines search motivations from consultation history to enhance current search queries. However, these methods do not take into account the value of consultations to users in personalized scenarios.

3 Problem Formulation

For each user $u \in \mathcal{U}$, the corresponding chronologically stored user history $\mathcal{H}_u = \{\mathcal{S}_u, \mathcal{C}_u, \mathcal{D}_u\} \subseteq \mathcal{H}$ includes: (1) search session history $\mathcal{S}_u$, (2) consultation history $\mathcal{C}_u$, and (3) interaction history

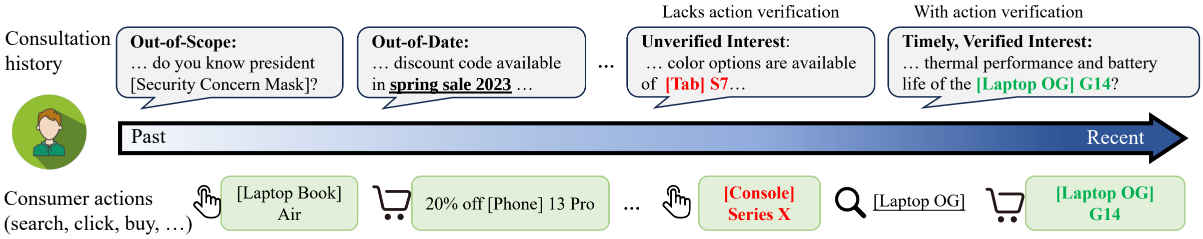

Lacks action verification
With action verification
Consultation history
Out-of-Scope:
… do you know president [Security Concern Mask]?
Out-of-Date:
… discount code available in spring sale 2023 …
…
Unverified Interest:
… color options are available of [Tab] S7…
Timely, Verified Interest:
… thermal performance and battery life of the [Laptop OG] G14?
Past
Recent
Consumer actions (search, click, buy, …)
[Laptop Book] Air
20% off [Phone] 13 Pro
…
[Console] Series X
[Laptop OG]
[Laptop OG] G14

Figure 2: Illustration of consumer consultation patterns and action verification in E-commerce platform. The bold and underlined text indicates outdated information; red text means no action verified for the interest; green text shows verified needs with actions.

$\mathcal{D}_u$. Specifically, let $\mathcal{S}_u = \{s_u^{(1)}, \dots, s_u^{(N)}\} \subseteq \mathcal{S}$, where $s_u^{(i)} = (q_u^{(i)}, a_{qi})$ denotes the i-th search session. Let $q_u^{(i)}$ denotes the i-th search query, and a_{qi} denotes the i-th search action. Similarly, $\mathcal{C}_u = \{c_u^{(1)}, \dots, c_u^{(M)}\} \subseteq \mathcal{C}$ represents M consultation sessions for user u. The interaction history $\mathcal{D}_u = \mathcal{S}_u \bigcup \{(v_u^{(1)}, a_{v1}), \dots, (v_u^{(K)}, a_{vK})\} \subseteq \mathcal{D}$, which contains interacted search session history $\mathcal{S}_u$ and item $v \in \mathcal{V}$ and corresponding interaction action categories $a \in \mathcal{A}$.

The task of personalized search is: Given $\mathcal{H}_u$, a new query $q_u^{(N+1)}$, and a candidate item list $\mathcal{V}_u' \subseteq \mathcal{V}$, the model assigning a ranking probability score $p\left(v' | \mathcal{H}_u, q_u^{(N+1)}, \mathcal{V}_u'\right)$ to each candidate item $v' \in \mathcal{V}_u'$.

4 Methodology

In consultation value assessment, considering that (1) value lacks explicit labels (making it difficult for model learning), and (2) consultation value is not fixed but relative (the same consultation may have different value for different searches), we therefore introduce empirical functions for assess multi-value on the data side, followed by alignment on the model side, rather than direct model learning.

The overview of our VAPS can be found in Fig. 3. We will introduce the main methods of VAPS in the next two sections: (1) Data-side Consultation Value Assessment and (2) Model-side Consultation Value Alignment.

4.1 Consultation Value Assessment

In response to the consultation value mentioned in Sec. 1, we first define the consultation value as $\mathcal{O}$.

For $\forall s \in \mathcal{S}_u$, there exist a consultation set $\mathcal{C}_u^{(t_s)} \subseteq \mathcal{C}_u$ that occurred before s and an item interaction history $\mathcal{D}_u^{(t_s)} \subseteq \mathcal{D}_u$ that occurred after s, where t_s is the timestamp of s. For $\forall c \in \mathcal{C}_u^{(t_s)}$, based on the target search s, we will present different value definitions for consultation c in this section, including **time decay value**, **scenario scope value**, and **posterior action value**, all in the form of functions. Finally, the value of consultation c for user u and search session s can be defined as $\mathcal{O}_u^{s,c}$.

4.1.1 Time Decay Value

Inspired by forgetting curves (Rubin and Wenzel, 1996), the **time decay value** $\mathcal{O}_{\text{time}}^{s,c}$ measures the freshness value of a consultation for a user conducting a current search query. This value models the intuition that recent consultations leave a stronger impression on users and reflect their immediate needs, whereas older consultations are less relevant to their current interests. Specifically, $\mathcal{O}_{\text{time}}^{s,c}$ depends on the time interval between the consultation and the current search:

$$\mathcal{O}_{\text{time}}^{s,c} = \alpha^{t_s - t_c},$$

where t_s and t_c are the hour-level timestamps of s and c, respectively. α is a temporal decay factor[1].

We observe that directly using raw timestamps leads to severe sparsity issues, as each timestamp is often unique and difficult for the model to generalize across similar temporal patterns. Moreover, the time intervals between consecutive user interactions vary significantly across users, making it challenging to capture consistent temporal dependencies. Inspired by the work (Li et al., 2020), we discretize the time intervals into b coarse-grained buckets, which helps alleviate the sparsity problem and enables the model to learn user behavior dynamics in a more robust and generalizable manner.

4.1.2 Scenario Scope Value

An observation is that out-of-scope consultations (e.g., discussions about politics) provide little value

---

[1] Based on empirical experience, for scenarios where consultations older than 30 days are deemed irrelevant, setting $\alpha = 0.99$ results in $\mathcal{O}_{\text{time}}^{s,c} = 7 \times 10^{-3} \approx 0$.

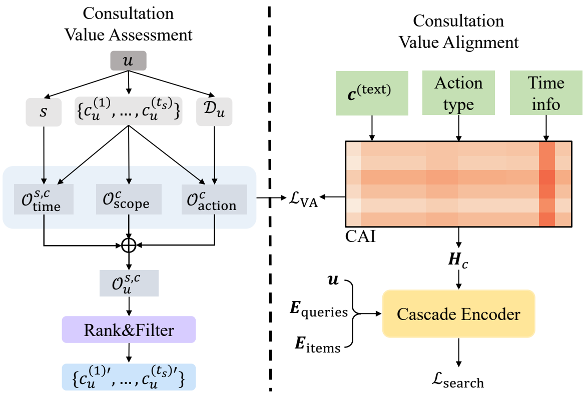

Consultation
Value Assessment
u
s
$\{c_u^{(1)}, \ldots, c_u^{(t_s)}\}$
$\mathcal{D}_u$
$\mathcal{O}_{\text{time}}^{s,c}$
$\mathcal{O}_{\text{scope}}^{c}$
$\mathcal{O}_{\text{action}}^{c}$
$\mathcal{O}_u^{s,c}$
Rank&Filter
$\{c_u^{(1)\prime}, \ldots, c_u^{(t_s)\prime}\}$
Consultation
Value Alignment
$\boldsymbol{c}^{(\text{text})}$
Action type
Time info
$\mathcal{L}_{\text{VA}}$
CAI
$\boldsymbol{H}_c$
$\boldsymbol{u}$
$\boldsymbol{E}_{\text{queries}}$
$\boldsymbol{E}_{\text{items}}$
Cascade Encoder
$\mathcal{L}_{\text{search}}$

Figure 3: Overview of VAPS.

to e-commerce platforms, while consultations with scenario-specific terms are more likely to accurately represent user needs. To operationalize this insight, we quantify **scenario scope value** $\mathcal{O}_{\text{scope}}^{c}$ with scenario-specific terms.

Terms can be directly obtained from the inverted index $\mathcal{I}$ that e-commerce search systems always maintain (Catena et al., 2014). The inverted index $\mathcal{I}_{\text{invert}} = \{term : [v_1, v_2, ..., v_{N_t}]\}$ establishes a mapping from scenario-specific terms *term* (e.g., “2023 Sales Event” and “Folding Phone”) to their corresponding item lists, where N_{term} is the total number of items linked to term *term*.

As our purpose is to eliminate out-of-scope consultations rather than to discriminate between ‘marginally relevant’ and ‘highly relevant’ cases, we apply a thresholding approach to compute $\mathcal{O}_{\text{scope}}^{c}$:

$$\mathcal{O}_{\text{scope}}^{c} = f_{\text{scope}}(|\mathcal{I}_c|), \quad \mathcal{I}_c \subseteq \mathcal{I}_{\text{invert}},$$
$$f_{\text{scope}}(x) = \begin{cases} x/\lambda_{\text{thresh}}, & \text{if } x < \lambda_{\text{thresh}}; \\ 1.0, & \text{otherwise.} \end{cases}$$

4.1.3 Posterior Action Value

To distinguish between casual consultation and those that users genuinely care about, we measure the **posterior action value** $\mathcal{O}_{\text{action}}^{c}$ of the target consultation c through relevant consumer actions.

Here, we consider the three most typical consumer actions[2] in e-commerce search systems: buy, click, and search, i.e., $\mathcal{A} = \{\text{search, click, buy}\}$. Here, we employ offline inverted retrieval to annotate subsequent actions associated with each consultation, with details in App. A.1.

The relative importance of consumer actions varies depending on platform strategies. Based on related studies (Verhallen, 1982; Teubner and Graul, 2020) and practical experience, we adhere to a common principle: Scarcer consumer actions generally hold greater value. For instance, relatively scarce purchases typically prove more valuable than relatively common clicks. The specific formula is as follows:

$$\mathcal{O}_{\text{action}}^{c} = \sum_{a} \gamma_a R_c^{T_a}, \quad \gamma_a = \frac{1}{|T_a|} \cdot \frac{1}{\sum_j \frac{1}{|T_j|}},$$

where γ_a represents the unit scarcity value of action type a, $a \in \mathcal{A}$. $T_a \subseteq \mathcal{D}_u^{(t_s)}$ is the subset of interaction histories in $\mathcal{D}_u^{(t_s)}$ with action type a. $R_c^{T_a}$ is the action frequency of action type a corresponding to the current consultation. Details can be found in App. A.2.

4.1.4 Aggregated Consultation Value

To ensure interpretability, avoid over-penalizing low scores, and mitigate delayed feedback issue[3] (Joulani et al., 2013), we use weighted summation. The aggregated consultation value score is computed as:

$$\begin{aligned} \mathcal{O}_u^{s,c} =& (1-\lambda_1)\mathcal{O}_{\text{time}}^{s,c} + \\ & \lambda_1(\lambda_2 \mathcal{O}_{\text{scope}}^{c} + (1-\lambda_2)\mathcal{O}_{\text{action}}^{c}), \end{aligned} \tag{1}$$

where λ_1 and λ_2 are hyper-parameters controlling the weight (analysis in Sec. 5.6).

Considering the maximum input length L_{seq} for model (Qin et al., 2025; Shi et al., 2024), we can rank and filter the consultation history for each search session s of the user to get rid of useless consultation:

$$\mathcal{C}_u' = \text{Rank\&Filter}_{\mathcal{O}_u^{s,c}}(\mathcal{C}_u, L_{\text{seq}}, s, u).$$

4.2 Model-side Consultation Value Alignment

Previous works only consider the similarity between consultation history and the current query, without integrating interactions with user consumption actions on the model side. Based on the data-side consultation value assessment in Sec. 4.1, we expect the model to understand consultation needs not only by measuring the semantic similarity between consultations and the current query but also by modeling the interactions between consultations

---

[2] This paper only considers three typical actions. Note that other actions such as “favorites” and “cart additions” are also compatible.

[3] For example, when posterior actions from recent consultations have not yet been counted, freshness value (time-decay value) is used as a compensatory measure.

and user consumption actions to assess whether and to what extent consultations are valued and recognized by users. Therefore, to enable the model to understand the correlation between user consumption actions and consultation history, we introduce **Consultation-Action Interaction (CAI)**, aiming to help the model learn whether (and how strongly) consultations are supported by corresponding user consumption actions, thereby inferring the degree of user recognition for consultations.

4.2.1 Embedding Construction

First, based on the same dimensions d, we construct a token embedding layer TokenEmb() for the text vocabulary to model textual semantics, along with item embedding and user embedding layers (ItemEmb() and UserEmb()) to store collaborative information of items and users, respectively.

For user u, the user embedding $\boldsymbol{u} = \text{UserEmb}(u)$. Given the u's ranked and filtered consultation history $\mathcal{C}'_u = \{c_u^{(1)}, ..., c_u^{(M')}\}$, we can obtain the text embedding for consultations through MoAE (Mixture of Attention Expert) text encoder (Qin et al., 2025):

$$\boldsymbol{c}_j^{\text{(text)}} = \text{MoAE}(\text{TokenEmb}(tok_{j,1}, \ldots, tok_{j,N_{tok}})),$$

where $tok_{j,k}$ is the k-th token of the text $c_u^{(j)}$, $j = 1, 2, \ldots, M'$. We obtain the user consultation text sequence: $[c_j^{\text{(text)}}; \ldots; c_M^{\text{(text)}}]$. Similarly, we can obtain their corresponding text embeddings $q_j^{\text{(text)}}, j = 1, 2, \ldots, N$ and $v_j^{\text{(text)}}, j = 1, 2, \ldots, K$.

4.2.2 Consultation-Action Interaction

Inspired by Lin et al. (2022), we propose Consultation-Action Interaction (CAI), which employs cross-attention between consumer actions (as keys/values) and consultations (as queries). The purpose is to identify posterior support in consumer actions for consultations and capture the underlying relationships between consultations and actions.

Here, we introduce (1) discrete time interval embeddings TimeEmb() to align with the temporal awareness in Sec. 4.1.1, and (2) action-type embeddings ActionEmb() to distinguish between different action types as mentinoed in Sec. 4.1.3. For a_i in $\mathcal{D}_u$, the embedding is:

$$\boldsymbol{a}_i = \begin{cases} \text{ActionEmb}(a_i) + \boldsymbol{v}_{a_i} & \text{if } a \text{ in } \{\text{buy, click}\}; \\ \text{ActionEmb}(a_i) + \boldsymbol{q}_{a_i} & \text{if } a \text{ in } \{\text{search}\} \end{cases}.$$

We further incorporate temporal information to form the attention inputs:

$$\boldsymbol{e}_i^{\text{(attn-q)}} = \boldsymbol{c}_i^{\text{(text)}} + \text{TimeEmb}(c_i),$$
$$\boldsymbol{e}_j^{\text{(attn-k)}} = \boldsymbol{e}_j^{\text{(attn-v)}} = \boldsymbol{a}_i + \text{TimeEmb}(a_j).$$

Subsequently, the cross-attention mechanism derives posterior support from actions for c_i, while preserving the original text semantics through skip connections with hyperparameter λ_3:

$$\boldsymbol{h}_{c_i} = \boldsymbol{c}_i^{\text{(text)}} + \lambda_3 \text{Attn}(\boldsymbol{e}_i^{\text{(attn-q)}}, \boldsymbol{E}^{\text{(attn-k)}}, \boldsymbol{E}^{\text{(attn-v)}}),$$

where $\boldsymbol{E}^{\text{(attn-k)}} = [\boldsymbol{e}_1^{\text{(attn-k)}}; ...; \boldsymbol{e}_K^{\text{(attn-k)}}]$, $\boldsymbol{E}^{\text{(attn-v)}} = [\boldsymbol{e}_1^{\text{(attn-v)}}; ...; \boldsymbol{e}_K^{\text{(attn-v)}}]$.

4.2.3 Value-Assessment Alignment

To ensure alignment with value assessment in Sec. 4.1, besides ranking and filtering consultations with $\mathcal{O}_u^{s,c}$, we propose a value-assessment alignment loss $\mathcal{L}_{\text{VA}}$ to supervise the cross-attention scores of CAI.

Specifically, we obtain consultation-action pairs for the consultation-action mapping in Sec. 4.1.3. For each pair (c, a), we compute the weight-projected embeddings embeddings: $\boldsymbol{e}^{\text{(attn-q)}'} = \boldsymbol{e}^{\text{(attn-q)}} \boldsymbol{W}^{\text{(attn-q)}}$, $\boldsymbol{e}^{\text{(attn-k)}'} = \boldsymbol{e}^{\text{(attn-k)}} \boldsymbol{W}^{\text{(attn-k)}}$. Considering that the cross attention scores are obtained through dot product and softmax, and inspired by MAPS's alignment, we employ softmax-based contrastive learning to supervise the attention:

$$\text{sim}(c, a) = \boldsymbol{e}^{\text{(attn-q)}'} \cdot \boldsymbol{e}^{\text{(attn-k)}'}$$

$$\mathcal{L}_{\text{VA}} = -\sum_{(c,a)} \log \frac{\exp(\text{sim}(c, a)/\tau_1)}{\sum_{a^- \in \mathcal{D}_u \backslash a} \exp(\text{sim}(c, a^-)/\tau_1)},$$

where τ_1 is temperature parameter used to control the sharpness of the softmax distribution (Hinton, 2015).

4.2.4 Personalized Search Learning

Through the cascaded bidirectional attention encoder (Qin et al., 2025), we obtain the final query embedding.

$$\boldsymbol{e}_q^{\text{(final)}} = \text{Cascaded-Encoder}(\boldsymbol{H}_c, \boldsymbol{E}_{\text{items}}, \boldsymbol{E}_{\text{queries}}, \boldsymbol{u}),$$

where $\boldsymbol{H}_c = [\boldsymbol{h}_{c_1}; \ldots; \boldsymbol{h}_{c_M}]$, $\boldsymbol{E}_{\text{items}} = [\boldsymbol{v}_1; \ldots; \boldsymbol{v}_K]$, $\boldsymbol{E}_{\text{queries}} = [\boldsymbol{q}_1; \ldots; \boldsymbol{q}_N]$. For inference, candidate item $\boldsymbol{v}'$ can be ranked based on similarity-derived probability scores:

$$p(v' | \mathcal{H}_u, s_u^{(N+1)}, \mathcal{V}'_u) = \text{sim}(\boldsymbol{e}_q^{\text{(final)}}, \boldsymbol{v}').$$

For optimization, following previous works (Bi et al., 2020; Ai et al., 2017; Shi et al., 2024), the objective is to increase the similarity scores of ground-truth given user history. The personalized alignment loss $\mathcal{L}_{\text{search}}$ can be formulated as:

$$\mathcal{L}_{\text{search}} = \sum_{(u,v,q)} \log \frac{\exp(\text{sim}(\boldsymbol{e}_q^{(\text{final})}, \boldsymbol{v})/\tau_2)}{\sum_{v' \in \mathcal{V}'_u} \exp(\text{sim}(\boldsymbol{e}_q^{(\text{final})}, \boldsymbol{v}')/\tau_2)}.$$

Following existing works (Ai et al., 2019b; Shi et al., 2024; Zhang et al., 2025a), we employ negative sampling (Le and Mikolov, 2014). The overall loss $\mathcal{L}_{\text{final}}$ is:

$$\mathcal{L}_{\text{final}} = \mathcal{L}_{\text{search}} + \lambda_3 \mathcal{L}_{\text{VA}} + \lambda_4 ||\Theta||_2,$$

where λ_3 and λ_4 are hyper-parameters, and Θ denotes the model parameters.

5 Experiments

We answer the following research questions with experiments: **RQ1:** How does VAPS rank compared to existing baselines? **RQ2:** How effective is VAPS in retrieval? **RQ3:** How does VAPS compare to multi-scenario methods? **RQ4:** How effective are VAPS's individual modules? **RQ5:** How reliably does VAPS assess consultation value? **RQ6:** How do hyper-parameters influence consultation value assessment in VAPS? **RQ7:** What is the time complexity of VAPS?

5.1 Experimental Settings

5.1.1 Datasets

To validate VAPS's effectiveness, experiments are conducted on two datasets. **Commercial dataset** is a real user interaction dataset from an e-commerce platform with AI consulting services (Qin et al., 2025). We follow the original setup by using the last two days of data for validation and testing. **Amazon dataset** is derived from the Amazon Reviews dataset (Ni et al., 2019), subsequently enhanced by PersonalWAB (Cai et al., 2024) and MAPS (Qin et al., 2025), containing user profiles as well as interaction behaviors such as searches and consultations. We adopt the dataset version used in MAPS and strictly follow its data processing. The statistics of these datasets are shown in Tab. 1

5.1.2 Baselines

For **ranking** evaluation, we adopt these personalized search baselines: **ZAM** (Ai et al., 2019a), **HEM** (Ai et al., 2017), **AEM** (Ai et al., 2019a), **QEM** (Ai et al., 2019a), **TEM** (Bi et al., 2020),**CoPPS** (Dai et al., 2023), and MAPS (Qin et al., 2025). For **retrieval** performance, we additionally introduce traditional, dense, and conversational retrieval methods: **BM25** (Robertson et al., 2009), **BGE-M3** (Chen et al., 2024), and **CHIQ** (Mo et al., 2024). Furthermore, we also consider include multi-scenario methods, including **SESRec** (Si et al., 2023), **UnifiedSSR** (Xie et al., 2023), and **UniSAR** (Shi et al., 2024) . For more model settings and implementation details, see App. B.1.

Dataset	#Users	#Items	#Inters
Commercial	2096	2691	24662 / 18774
Amazon	967	35772	7263 / 40567

Table 1: Statistics of the 2 pre-processed datasets. In "#Inters", searches are shown on the left of '/', and consultations are shown on the right.

5.1.3 Metrics and Implementation Details

Following (Shi et al., 2024; Zhang et al., 2024b,a), we adopt **Hit Ratio** (HR@k) and **Normalized Discounted Cumulative Gain** (NDCG@k or N@k) for ranking, and **Mean Reciprocal Rank** (MRR@k) for retrieval. Following (Qin et al., 2025; Shi et al., 2025; Zhang et al., 2024c), each ground-truth item is paired with 99 negatives, evaluating HR/NDCG at $\{5, 10, 20, 50\}$. For retrieval, all items are candidates with MRR reported at $\{10, 20, 50\}$. Details appear in App. B.2.

5.2 Overall Performance

In a retrieval-then-ranking e-commerce system, the primary objective of personalized product search is personalized ranking performance. To answer **RQ1**, **RQ2**, and **RQ3**, we first evaluate the ranking performance in Tab. 2 and Tab. 4, followed by a comparison of the retrieval performance in Tab. 3.

Regarding ranking, VAPS outperforms all other personalized product search methods and search-integrated multi-scenario approaches. The improvements are significant across most metrics, particularly on the Amazon dataset. We also observed relatively smaller gains (approximately 2%) on Commercial. We attribute it to the inherently limited quantity of ground-truth items on Commercial and data saturation effect. Concerning retrieval, VAPS surpasses all personalized product search methods and traditional, dense, and conversational retrieval approaches. This fully shows VAPS' effectiveness and superiority in ranking and retrieval tasks, high-

Model	HR@5	HR@10	HR@20	HR@50	NDCG@5	NDCG@10	NDCG@20	NDCG@50
				Commercial				
ZAM	0.3680	0.5247	0.6810	0.8203	0.2491	0.2988	0.3378	0.3659
HEM	0.3487	0.4911	0.6375	0.8036	0.2359	0.2811	0.3183	0.3522
AEM	0.3892	0.5372	0.6721	0.8255	0.2648	0.3127	0.3474	0.3790
QEM	0.3991	0.5468	0.6732	0.8438	0.2675	0.3148	0.3461	0.3803
TEM	0.4062	0.5681	0.7199	0.8742	0.2869	0.3405	0.3761	0.4053
CoPPS	0.4057	0.5632	0.7178	0.8656	0.2829	0.3345	0.3735	0.4033
MAPS	0.5276	0.7064	0.8321	0.9323	0.3762	0.4360	0.4639	0.4871
VAPS	**0.5565**†	**0.7145**†	**0.8398**	**0.9422**†	**0.3884**†	**0.4424**†	**0.4726**†	**0.4931**†
				Amazon				
ZAM	0.3100	0.4487	0.5433	0.7302	0.1826	0.2115	0.2498	0.2782
HEM	0.2736	0.4192	0.5412	0.7458	0.1984	0.2173	0.2595	0.2975
AEM	0.3184	0.4559	0.5366	0.7247	0.1861	0.2128	0.2470	0.2769
QEM	0.2832	0.3879	0.5283	0.7664	0.1900	0.2122	0.2273	0.2918
TEM	0.4028	0.4813	0.7201	0.8051	0.2965	0.3123	0.3416	0.3540
CoPPS	0.3871	0.4862	0.7289	0.8013	0.2784	0.3299	0.3435	0.3696
MAPS	0.6062	0.7835	0.8990	0.9702	0.4237	0.4717	0.5001	0.5189
VAPS	**0.6418**†	**0.8019**†	**0.9101**†	**0.9748**	**0.4903**†	**0.5213**†	**0.5556**†	**0.5665**†

Table 2: Search ranking performance compared with personalized search baselines. The best results are shown in bold. '†' indicates the model significantly outperforms all baseline models with paired t-tests at $p < 0.05$ level.

Method	MRR@5	MRR@10	MRR@20	MRR@50
ZAM	0.2211	0.2539	0.2628	0.2702
HEM	0.2251	0.2524	0.2739	0.2872
AEM	0.2132	0.2424	0.2630	0.2705
QEM	0.2264	0.2540	0.2655	0.2808
TEM	0.2597	0.2729	0.3089	0.3239
CoPPS	0.2517	0.2806	0.3178	0.3371
BM25	0.2780	0.2870	0.2917	0.2997
BGE-M3	0.3408	0.3540	0.3614	0.3651
CHIQ	0.3526	0.3691	0.3883	0.4047
MAPS	0.4119	0.4324	0.4397	0.4523
VAPS	**0.4559**†	**0.4749**†	**0.4815**†	**0.4944**†

Table 3: Retrieval performance on the Amazon dataset.

Method	HR@10	HR@20	N@10	N@20
SESRec	0.5601	0.7189	0.3464	0.3788
UnifiedSSR	0.5709	0.7082	0.3598	0.3793
UniSAR	0.5837	0.7295	0.3605	0.3894
MAPS	0.7064	0.8321	0.4360	0.4639
VAPS	**0.7145**†	**0.8398**	**0.4424**†	**0.4726**†

Table 4: Search ranking performance compared with multi-scenario baselines on the Commercial dataset.

lighting its ability to boost search performance on e-commerce platforms.

5.3 Ablation Study

Ablation	HR@5	HR@10	HR@20	N@10	N@20
VAPS	**0.6418**	**0.8019**	**0.9101**	**0.5213**	**0.5556**
w/o $\mathcal{O}^c_{\text{time}}$	0.6133	0.7784	0.9011	0.4783	0.5106
w/o $\mathcal{O}^c_{\text{scope}}$	0.6189	0.7853	0.9036	0.4897	0.5202
w/o $\mathcal{O}^c_{\text{action}}$	0.6128	0.7792	0.9010	0.4795	0.5135
w/o $\mathcal{L}_{\text{VA}}$	0.6306	0.7909	0.9051	0.5062	0.5419
w/o CAI	0.6334	0.7867	0.9049	0.4928	0.5258

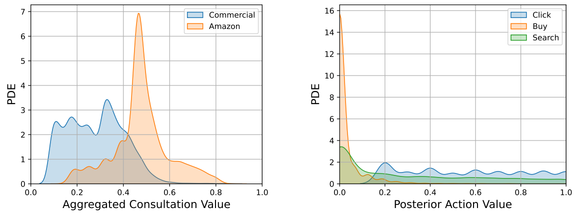
Table 5: Ablation study of VAPS on Amazon.

In this section, we discuss the specific roles of each module in VAPS to answer **RQ4**. As shown in Table 5, removing $\mathcal{O}^c_{\text{time}}$ from the aggregated consultation value yields the most significant performance drop, followed by removing $\mathcal{O}^c_{\text{action}}$. $\mathcal{O}^c_{\text{time}}$ directly measures the temporal difference between consultations and the current search. Through a decay function, it effectively captures the user's evolving search-consultation interests over time. When $\mathcal{O}^c_{\text{time}}$ is removed, the aggregated consultation value loses temporal context, failing to distinguish the similarity of old versus new consultations to the current search. This results in personalized search returning outdated or irrelevant content. Conversely, removing $\mathcal{O}^c_{\text{action}}$ causes the aggregated consultation value to overlook validation from actual user actions in later interations. Relying solely on semantic matching of search queries and temporal information, the filtered consultations lack real-world posterior support, leading personalized search astray from genuine user needs.

5.4 Consultation Value Distribution Analysis

To answer **RQ5**, we present the assessed values in a distribution format for both datasets.

Fig. 4a reveals that the commercial dataset ex-



(a) $\mathcal{O}_u^{s,c}$ on the Amazon and Commercial datasets.

(b) $\mathcal{O}^c_{\text{action}}$ of each a on Commercial.

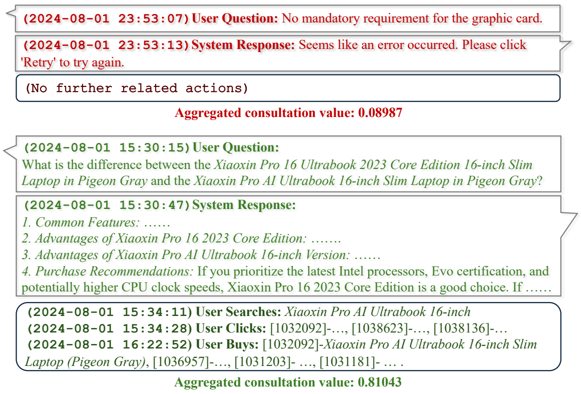
Figure 4: Value distribution visualization.

hibits lower normalized scores overall, attributable to its sparser user interactions, consistent with the fact that its primary product categories are non-daily necessities. In Fig. 4b, we find the value score distribution of clicks is very flat, reflecting their high prevalence as a user action. In contrast, searches are relatively concentrated near 0, while purchases exhibit the sharpest distribution, reflecting increasing scarcity. As the “action cost” progressively increases, action become increasingly scarce and consequently more valuable (indicating stronger user intent).

5.5 Case Study

In this section, we analyse discrete cases for consultation value scores in both commercial and amazon datasets.

As shown in Fig. 5, the consultation in the first case provide no useful information about the graphic card requirement, reflected by the low value of 0.08987. In contrast, the second case involved a user query about the differences between two laptops. The system give a comprehensive response covering common features, advantages of each model, and purchase recommendations. Moreover, the consultation is also in-scope and timely, verified by subsequent user actions such as related searches and purchases, resulting in a high-valued consultation (0.81043).


`(2024-08-01 23:53:07)` **User Question:** No mandatory requirement for the graphic card.

`(2024-08-01 23:53:13)` **System Response:** Seems like an error occurred. Please click 'Retry' to try again.

`(No further related actions)`

Aggregated consultation value: 0.08987

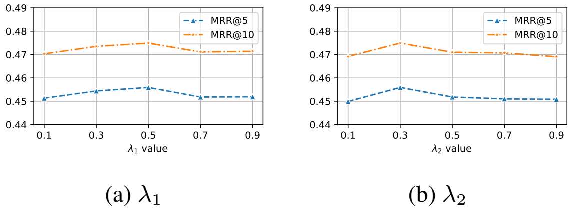
`(2024-08-01 15:30:15)` **User Question:**
What is the difference between the *Xiaoxin Pro 16 Ultrabook 2023 Core Edition 16-inch Slim Laptop in Pigeon Gray* and the *Xiaoxin Pro AI Ultrabook 16-inch Slim Laptop in Pigeon Gray*?

`(2024-08-01 15:30:47)` **System Response:**
1. Common Features: ……
2. Advantages of Xiaoxin Pro 16 2023 Core Edition: …….
3. Advantages of Xiaoxin Pro AI Ultrabook 16-inch Version: ……
4. Purchase Recommendations: If you prioritize the latest Intel processors, Evo certification, and potentially higher CPU clock speeds, Xiaoxin Pro 16 2023 Core Edition is a good choice. If ……

`(2024-08-01 15:34:11)` **User Searches:** *Xiaoxin Pro AI Ultrabook 16-inch*
`(2024-08-01 15:34:28)` **User Clicks:** [1032092]-…, [1038623]-…, [1038136]-…
`(2024-08-01 16:22:52)` **User Buys:** [1032092]-*Xiaoxin Pro AI Ultrabook 16-inch Slim Laptop (Pigeon Gray)*, [1036957]-…, [1031203]- …, [1031181]- … .

Aggregated consultation value: 0.81043

Figure 5: Consultation value case study with actions on Commercial.



(a) λ_1

(b) λ_2

Figure 6: Impact of different hyperparameters on Amazon. Default setting is $\lambda_1 = 0.5, \lambda_2 = 0.3$.

5.6 Hyperparameter Analysis

In this section, we investigate the impact of changing hyper-parameters λ_1 and λ_2 to answer **RQ6**. According to Eq. (1), the values of λ_1 and λ_2 should be within the range of [0, 1]. We performed parameter tuning for λ_1 and λ_2 separately. By freezing one parameter at a time, we test multiple weight values between 0 and 1 for the other parameter. The optimal λ_1 is found to be 0.5, which highlights the necessity of analyzing the temporal impact of consultations independently. The optimal λ_2 is 0.3, indicating that posterior actions are more critical than scenario scopes for the aggregated consultation value and have a greater influence on user behavior. Meanwhile, we assume that the scope value (based on inverted indices) might be overly restrictive, and more effective alternatives should be explored in future work.

These conclusions are consistent with those in Sec. 5.3.

5.7 Complexity Analysis

To address **RQ7**, we conduct complexity analysis in this section. VAPS has an overall time complexity of $O(N'^2 d)$ ($N' = \max(M, N, K)$; see App. B.3 for details), matching state-of-the-art methods like UniSAR and MAPS. Notably, VAPS outperforms these methods under the same complexity constraints. While optimizing computational efficiency for large-scale platforms matters, we defer this to future work.

6 Conclusion

We propose the VAPS method to enhance e-commerce personalized search by leveraging user consultations. Existing semantic-only methods fail to capture consultation value. To address this gap, we propose a consultation value assessment framework comprising three dimensions: Scenario Scope, Posterior Action, and Time Decay Value. Based on this framework, we develop VAPS, a value-aware personalized search model that incorporates high-value consultations through a consultation–user action interaction module and an objective function explicitly aligning consultations with user actions.

7 Limitations

The VAPS framework advances personalized search through its novel consultation value assessment and alignment mechanisms, though there remains some limitations and room for improvement in certain aspects to further enhance its research contributions. First, while its time complexity aligns with state-of-the-art methods, further optimizing the computation for ultra-large-scale environments could enhance real-time deployment flexibility. Regarding hyperparameter tuning, while the optimal values of hyper-parameters are validated in experimental settings, developing automated tuning strategies could reduce dependency on manual adjustments across diverse business scenarios. Furthermore, the model's reliance on historical interaction data raises considerations for cold-start scenarios. Integrating content-based features or transfer learning could mitigate this challenge. Lastly, while the current design focuses on interaction data, enriching the framework with external knowledge graphs or large language model insights could enhance its ability to parse complex consultative intents in specialized domains. These observations highlight opportunities for future research to further strengthen VAPS's scalability, adaptability, and semantic modeling capabilities.

Acknowledgement

This work was funded by the National Natural Science Foundation of China (62502091), the Beijing Social Science Foundation Planning Project (Grant No. 24GLC041), the Fundamental Research Funds for the Central Universities in UIBE (Grant No. 24QN06, 24PYTS22). The work was done while Weicong Qin, Yi Xu and Chenglei Shen were interns at Lenovo.

References

Qingyao Ai, Daniel N Hill, SVN Vishwanathan, and W Bruce Croft. 2019a. A zero attention model for personalized product search. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 379–388.

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 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 645–654.

Qingyao Ai, Yongfeng Zhang, Keping Bi, and W Bruce Croft. 2019b. Explainable product search with a dynamic relation embedding model. *ACM Transactions on Information Systems (TOIS)*, 38(1):1–29.

Keping Bi, Qingyao Ai, and W Bruce Croft. 2020. A transformer-based embedding model for personalized product search. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1521–1524.

Hongru Cai, Yongqi Li, Wenjie Wang, Fengbin Zhu, Xiaoyu Shen, Wenjie Li, and Tat-Seng Chua. 2024. Large language models empowered personalized web agents. *Preprint*, arXiv:2410.17236.

Jianneng Cao, Panagiotis Karras, Panos Kalnis, and Kian-Lee Tan. 2011. Sabre: a sensitive attribute bucketization and redistribution framework for t-closeness. *The VLDB Journal*, 20:59–81.

Matteo Catena, Craig Macdonald, and Iadh Ounis. 2014. On inverted index compression for search engine efficiency. In *Advances in Information Retrieval: 36th European Conference on IR Research, ECIR 2014, Amsterdam, The Netherlands, April 13-16, 2014. Proceedings 36*, pages 359–371. Springer.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Jiayi Chen, Wen Wu, Liye Shi, Wei Zheng, and Liang He. 2023. Long-tail session-based recommendation from calibration. *Applied Intelligence*, 53(4):4685–4702.

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 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2023, Long Beach, CA, USA, August 6-10, 2023*, pages 380–389. ACM.

Rafael Ferreira, David Semedo, and João Magalhães. 2023. Rating prediction in conversational task assistants with behavioral and conversational-flow features. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2314–2318.

Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. 2017. Deepfm: a factorization-machine based neural network for ctr prediction. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 1725–1731.

Rebecca Hamilton and Linda L Price. 2019. Consumer journeys: Developing consumer-based strategy.

Henderi Henderi, Tri Wahyuningsih, and Efana Rahwanto. 2021. Comparison of min-max normalization and z-score normalization in the k-nearest neighbor (knn) algorithm to test the accuracy of types of breast cancer. *International Journal of Informatics and Information Systems*, 4(1):13–20.

Geoffrey Hinton. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Michael Jahrer, Andreas Töscher, and Robert Legenstein. 2010. Combining predictions for accurate recommender systems. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 693–702.

Chen Jin, Luyi Yang, and Kartik Hosanagar. 2023. To brush or not to brush: Product rankings, consumer search, and fake orders. *Information Systems Research*, 34(2):532–552.

Pooria Joulani, Andras Gyorgy, and Csaba Szepesvári. 2013. Online learning under delayed feedback. In *International conference on machine learning*, pages 1453–1461. PMLR.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR.

Jiacheng Li, Yujie Wang, and Julian McAuley. 2020. Time interval aware self-attention for sequential recommendation. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, WSDM '20, page 322–330, New York, NY, USA. Association for Computing Machinery.

Hezheng Lin, Xing Cheng, Xiangyu Wu, and Dong Shen. 2022. Cat: Cross attention in vision transformer. In *2022 IEEE international conference on multimedia and expo (ICME)*, pages 1–6. IEEE.

Fengran Mo, Abbas Ghaddar, Kelong Mao, Mehdi Rezagholizadeh, Boxing Chen, Qun Liu, and Jian-Yun Nie. 2024. CHIQ: Contextual history enhancement for improving query rewriting in conversational search. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2268.

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *EMNLP-IJCNLP*.

Weicong Qin, Yi Xu, Weijie Yu, Chenglei Shen, Ming He, Jianping Fan, Xiao Zhang, and Jun Xu. 2025. Maps: Motivation-aware personalized search via llm-driven consultation alignment. *Preprint*, arXiv:2503.01711.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

David C Rubin and Amy E Wenzel. 1996. One hundred years of forgetting: A quantitative description of retention. *Psychological review*, 103(4):734.

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 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1029–1039.

Teng Shi, Jun Xu, Xiao Zhang, Xiaoxue Zang, Kai Zheng, Yang Song, and Enyun Yu. 2025. Unified generative search and recommendation. *arXiv preprint arXiv:2504.05730*.

Zihua Si, Zhongxiang Sun, Xiao Zhang, Jun Xu, Xiaoxue Zang, Yang Song, Kun Gai, and Ji-Rong Wen. 2023. When search meets recommendation: Learning disentangled search representation for recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, pages 1313–1323. ACM.

Timm Teubner and Antje Graul. 2020. Only one room left! how scarcity cues affect booking intentions on hospitality platforms. *Electronic Commerce Research and Applications*, 39:100910.

Theo MM Verhallen. 1982. Scarcity and consumer choice behavior. *Journal of Economic Psychology*, 2(4):299–322.

Jiayi Xie, Shang Liu, Gao Cong, and Zhenzhong Chen. 2023. Unifiedssr: A unified framework of sequential search and recommendation. *arXiv preprint arXiv:2310.13921*.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. 2024. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. *Preprint*, arXiv:2406.12045.

Jingying Zeng, Hui Liu, Zhenwei Dai, Xianfeng Tang, Chen Luo, Samarth Varshney, Zhen Li, and Qi He. 2025. Cite before you speak: Enhancing context-response grounding in e-commerce conversational llm-agents. *arXiv preprint arXiv:2503.04830*.

Changshuo Zhang, Teng Shi, Xiao Zhang, Qi Liu, Ruobing Xie, Jun Xu, and Ji-Rong Wen. 2024a. Modeling domain and feedback transitions for cross-domain sequential recommendation. *arXiv preprint arXiv:2408.08209*.

Changshuo Zhang, Teng Shi, Xiao Zhang, Yanping Zheng, Ruobing Xie, Qi Liu, Jun Xu, and Ji-Rong Wen. 2024b. Qagcf: Graph collaborative filtering for q&a recommendation. *arXiv preprint arXiv:2406.04828*.

Kepu Zhang, Teng Shi, Sunhao Dai, Xiao Zhang, Yinfeng Li, Jing Lu, Xiaoxue Zang, Yang Song, and Jun Xu. 2024c. Saqrec: Aligning recommender systems to user satisfaction via questionnaire feedback. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 3165–3175.

Kepu Zhang, Zhongxiang Sun, Weijie Yu, Xiaoxue Zang, Kai Zheng, Yang Song, Han Li, and Jun Xu. 2025a. Qe-rag: A robust retrieval-augmented generation benchmark for query entry errors. *arXiv preprint arXiv:2504.04062*.

Kepu Zhang, Zhongxiang Sun, Xiao Zhang, Xiaoxue Zang, Kai Zheng, Yang Song, and Jun Xu. 2025b. Trigger3: Refining query correction via adaptive model selector. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 13260–13268.

A Appendix: Methodology Details

A.1 Appendix: Details of Related Actions for Consultations

Inspired by Qin et al. (2025), to build an **offline inverted retrieval** for recording the relevant consumer actions corresponding to each consultation, we set the following conditions and rules: 1. For each consumer action $a \in \mathcal{A}$ of a user, we search for related consultations $c \in \mathcal{C}$ within a certain time period before the action occurs (time ranges considered: 3, 5, 7, 14 days; due to the sparsity of the dataset, we set the time range to 14 days), and construct an action-related consultation table $x = \{a : \mathcal{C}_{\text{related}}\}$. Specifically, if the text information TI corresponding to the consumer action a (e.g., search query; item name and text features for clicks/purchases) appears completely at least once in consultation c, or more than half of the item contents in TI appears in c as text, or more than half of the query terms in TI appear in c as text, then consultation c is added to $\mathcal{C}_{\text{related}}$. 2. Reverse each key-value pair in the action-related consultation table x to obtain a consultation-related action inverted table $y = \{c : \mathcal{A}_{\text{related}}\}$, which constitutes the final content of the offline inverted retrieval.

A.2 Appendix: Details of the Usage of Bucketization and $R_c^{T_a}$ for Posterior Action Values

To enhance the robustness of subsequent model training, the calculated consultation value has been normalized to [0, 1]. Consequently, each component of the consultation value should also be normalized to [0, 1]. While the normalization of time decay and scenario scope value is relatively straightforward, typically using direct exponential or reciprocal-based methods, the normalization of posterior action value needs additional consideration.

In e-commerce platforms, consumer action data (e.g., user clicks, purchases) often exhibit long-tailed distributions, where a few extreme values (Chen et al., 2023; Jin et al., 2023) (e.g., a hacker artificially generating over 10k clicks in short intervals through network attacks) can significantly distort normalization methods sensitive to boundary conditions, such as Min-Max and Z-Score (Henderi et al., 2021). Bucketization, however, discretizes continuous values into fixed intervals, effectively mitigating the impact of outliers (Jahrer et al., 2010; Cao et al., 2011).

Therefore, we introduce bucketization in posterior action value part of VAPS to convert raw action counts into relative frequency values $R_c^{T_a}$), which includes:

1. **Construct Action Frequency Table** $\text{freq}_{\mathcal{A}}$: Using the consultation-related action inverted table $y = \{c : \mathcal{A}_{\text{related}}\}$ from offline inverted retrieval, we count the occurrences of each action type for every consultation c, aggregating them into a consultation-action frequency table $\text{freq}_{\mathcal{A}}$.

2. **Generate Eleven Equal Quantiles**: For each action type in $\text{freq}_{\mathcal{A}}$, compute the eleven equal quantiles (i.e., dividing the data into 11 groups with equal probability mass). These quantiles define 11 buckets (No.0 to No.10) for each action type.

3. **Map Frequencies to Relative Values**: For a given consultation c and its corresponding posterior action frequency T_a, determine the bucket to which T_a belongs. The relative frequency value $R_c^{T_a}$ is then calculated as the bucket index divided by 10 (e.g., bucket No.5 maps to $R_c^{T_a} = 0.5$).

B Appendix: Experiment Details

B.1 Baseline Details

We initiate our comparisons by evaluating our method against various baselines for ranking and retrieval tasks. For personalized search models, we consider the following:

- **AEM** (Ai et al., 2019a), an attention-based personalized model that merges the user's historical interaction items with the current query, enabling a more context-aware search experience.
- **QEM** (Ai et al., 2019a), which solely focuses on the matching scores between items and queries, providing a more straightforward but query-centric ranking approach.
- **HEM** (Ai et al., 2017), a latent vector-based personalized model that captures user preferences through hidden representations.
- **ZAM** (Ai et al., 2019a), an enhanced version of AEM that appends a zero vector to the item list, aiming to improve the model's performance in certain scenarios.
- **TEM** (Bi et al., 2020), which replaces the attention layer of AEM with a transformer encoder, leveraging the powerful sequential modeling capabilities of transformers.
- **CoPPS** (Dai et al., 2023), a model that harnesses contrastive learning techniques to enhance personalized search performance.

In addition, we benchmark our method against **MAPS** (Qin et al., 2025), the pioneering model that utilizes consultation information for personalized search. Next, we compare our approach with multi-scenario methods that integrate search and recommendation interactions:

- **SESRec** (Si et al., 2023) employs contrastive learning to learn disentangled search representations, facilitating more effective recommendations.
- **UnifiedSSR** (Xie et al., 2023) jointly models user behavior history across search and recommendation scenarios, capturing the intricate relationships between the two.
- **UniSAR** (Shi et al., 2024) effectively models diverse fine-grained behavior transitions using two distinct transformers and implements a cross-attention mechanism for enhanced interaction modeling.

For the retrieval task, we contrast VAPS with traditional, deep learning-based, and conversational-based baselines:

- **BM25** relies on word frequency to identify and rank relevant retrieval candidates, a classic approach in information retrieval.
- **BGE-M3** incorporates concepts such as embedding to boost the performance of retrieval tasks, leveraging deep learning techniques.
- **CHIQ** endeavors to integrate world knowledge from large language models (LLMs) into the search process, aiming to enhance retrieval effectiveness.

B.2 Implementation Details

All hyperparameters of the baseline are searched according to the settings in the original paper. Following related work (Shi et al., 2024), we set d to 64, and the maximum length L_{seq} of the user history sequence to 30. Fully following Qin et al. (2025), we filter out users with fewer than 5 interactions, use 'tanh' as the activation function, set the number of layers in the transformer encoder to 1, batch size to 72, and the number of negative samples for each positive sample for $\mathcal{L}_{\text{search}}$ to 10. For $\mathcal{L}_{\text{VA}}$, the random sampled negative is adopted, with the corresponding batch size searched among {64,128,256}. λ_{thresh} is set to 4. τ_1, τ_2, λ_1, λ_2, λ_3 and λ_4 is tuned in {0: 0.1: 1}. α is tuned in {0.98,0.99,0.999,0.9999}. We train all models for 100 epochs with early stopping (5 epochs) to avoid overfitting and optimize using Adam (Kingma and Ba, 2014). The learning rate is adjusted among {1e-3, 5e-4, 1e-4, 5e-5, 1e-5}. All experiments were completed on an A800 (80GB) GPU.

B.3 Appendix: Complexity Analysis Details

In this section, we give detailed complexity analysis of VAPS. We analyse that time complextity of VAPS includes: (1) construction of the consultation-related action inverted table $y = \{c : \mathcal{A}_{\text{related}}\}$ in offline inverted retrieval: $O(M * (N + K))$. (2) get all three parts of consultation value and aggregate the $\mathcal{O}_u^{s,c}$: $O(N * M) + O(M^2) + O(M^2)$ (3) consultation value alignment needs to compute CAI and use cascade encoder to compute $\mathcal{L}_{\text{VA}}$ and $\mathcal{L}_{\text{search}}$: $O(N'^2 d)$, where d is the encoding dimension, and $N' = \max(M, N, K)$.

Among all three parts, former two parts can be computed offline, and has lower complexity than

$O(N'^2 d)$, i.e. $O(M*(N+K)) + O(N*M) + O(M^2) + O(M^2) < O(N'^2) < O(N'^2 d)$, so final time complexity of VAPS is $O(N'^2 d)$.

C Appendix: Dataset License

Following Qin et al. (2025), we provide details of dataset license used in this paper. The Amazon dataset (based on PersonalWAB (Cai et al., 2024)) is released under the CC BY-NC 4.0 License. Its benchmark implementation, which is built upon the MIT-licensed tau-bench (Yao et al., 2024), incorporates significant modifications and enhancements tailored to the project's requirements. For the derived components, the implementation strictly adheres to and complies with the licensing terms of tau-bench.

D Appendix: Dataset Details

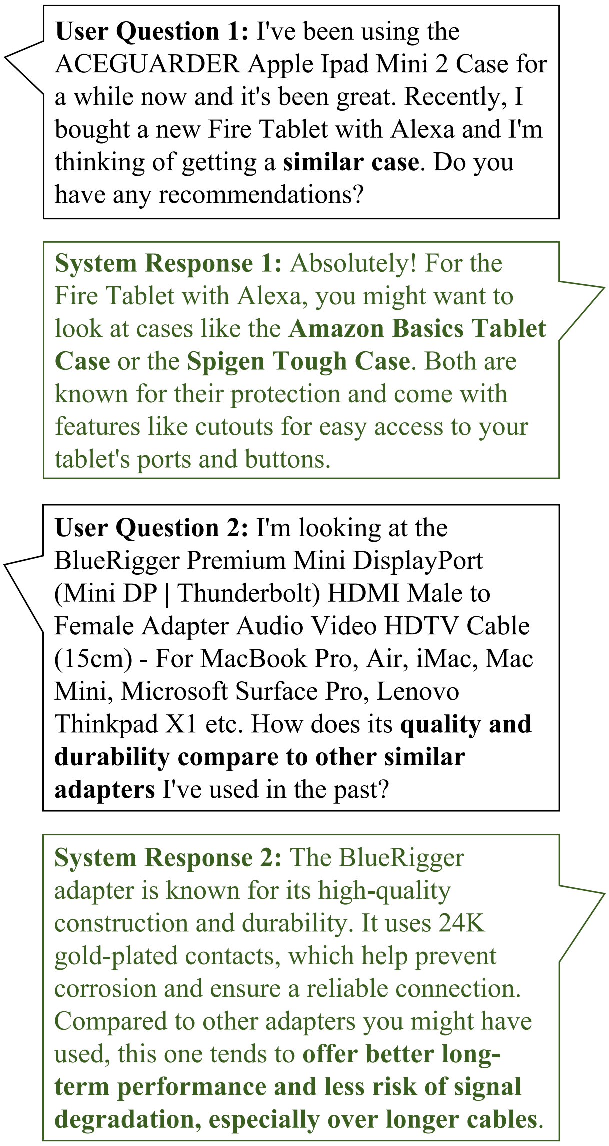
D.1 Dataset Repository

We publicly disclose the dataset used in this paper at the following link: `https://anonymous.4open.science/r/VAPS-to-go`. Note that we only disclose the **Amazon** dataset, since **Commercial** dataset is currently not available for publicity because of policy and law restriction. Also, we give the dataset example as follows (same as (Qin et al., 2025), due to we use the same datasets as theirs).

E Appendix: Discussion

The proposed Consultation Value Assessment Framework and VAPS model address critical gaps in consultation-enhanced personalized search by moving beyond semantic similarity to evaluate consultations through three novel dimensions: Scenario Scope, Posterior Action, and Time Decay Value. While these design choices are empirically justified, some aspects may invite scrutiny—particularly the heuristic nature of value functions, hyperparameter sensitivity, and dataset constraints. Here, we discuss the rationale behind these decisions and their implications.

First, the time decay value ($\mathcal{O}^{s,c}_{\text{time}}$) employs an exponential decay to model diminishing consultation similarity over time. Although alternative formulations (e.g., linear or step-wise decay) could be considered, exponential decay that we adopted aligns with prior work on temporal dynamics in user behavior (Rubin and Wenzel, 1996) and effectively captures the intuition that recent interactions better reflect immediate intent and even recent but irrelevant consultations leave some impression on users with a certain value. The choice of α as a hyperparameter ensures simplicity, though adaptive decay rates (e.g., user-specific or activity-dependent) could further enhance personalization.



User Question 1: I've been using the ACEGUARDER Apple Ipad Mini 2 Case for a while now and it's been great. Recently, I bought a new Fire Tablet with Alexa and I'm thinking of getting a **similar case**. Do you have any recommendations?

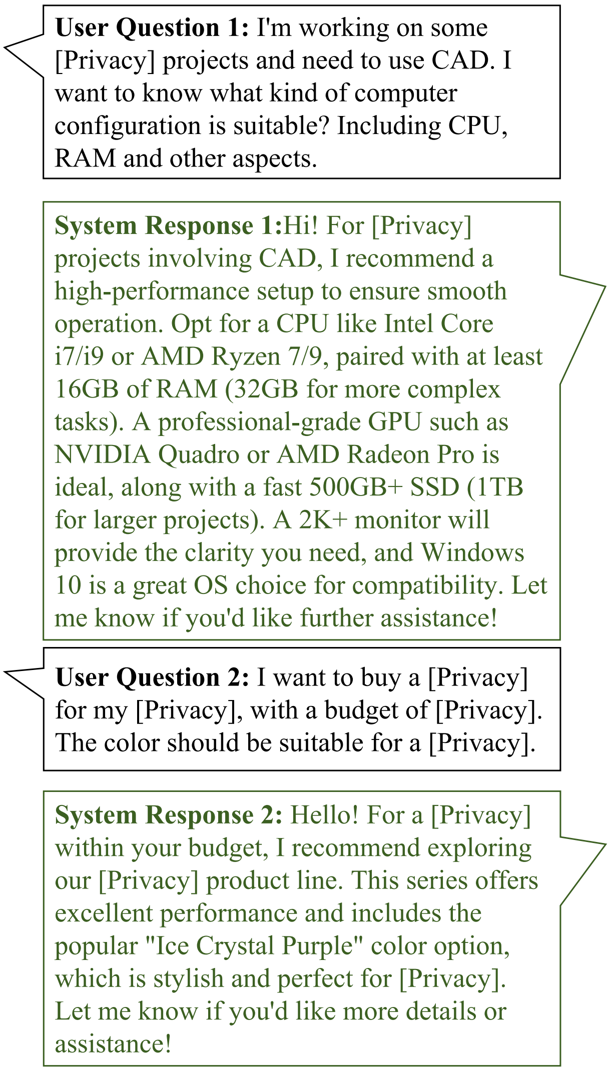
System Response 1: Absolutely! For the Fire Tablet with Alexa, you might want to look at cases like the **Amazon Basics Tablet Case** or the **Spigen Tough Case**. Both are known for their protection and come with features like cutouts for easy access to your tablet's ports and buttons.

User Question 2: I'm looking at the BlueRigger Premium Mini DisplayPort (Mini DP | Thunderbolt) HDMI Male to Female Adapter Audio Video HDTV Cable (15cm) - For MacBook Pro, Air, iMac, Mac Mini, Microsoft Surface Pro, Lenovo Thinkpad X1 etc. How does its **quality and durability compare to other similar adapters** I've used in the past?

System Response 2: The BlueRigger adapter is known for its high-quality construction and durability. It uses 24K gold-plated contacts, which help prevent corrosion and ensure a reliable connection. Compared to other adapters you might have used, this one tends to **offer better long-term performance and less risk of signal degradation, especially over longer cables**.

Figure 7: Examples of consultations on the **Amazon** dataset.

Second, the scenario scope value ($\mathcal{O}^{c}_{\text{scope}}$) uses a necessary thresholding approach to filter out-of-domain consultation noise (e.g., politics discussions). Inverted indices provide a effective and practical way to identify in-scope terms, the threshold was selected empirically to balance precision and recall. A stricter threshold might exclude marginally relevant consultations, whereas a lenient one risks noise inclusion. Future work could explore dynamic thresholds based on consultation



User Question 1: I'm working on some [Privacy] projects and need to use CAD. I want to know what kind of computer configuration is suitable? Including CPU, RAM and other aspects.

System Response 1: Hi! For [Privacy] projects involving CAD, I recommend a high-performance setup to ensure smooth operation. Opt for a CPU like Intel Core i7/i9 or AMD Ryzen 7/9, paired with at least 16GB of RAM (32GB for more complex tasks). A professional-grade GPU such as NVIDIA Quadro or AMD Radeon Pro is ideal, along with a fast 500GB+ SSD (1TB for larger projects). A 2K+ monitor will provide the clarity you need, and Windows 10 is a great OS choice for compatibility. Let me know if you'd like further assistance!

User Question 2: I want to buy a [Privacy] for my [Privacy], with a budget of [Privacy]. The color should be suitable for a [Privacy].

System Response 2: Hello! For a [Privacy] within your budget, I recommend exploring our [Privacy] product line. This series offers excellent performance and includes the popular "Ice Crystal Purple" color option, which is stylish and perfect for [Privacy]. Let me know if you'd like more details or assistance!

Figure 8: Examples of consultations on the **Commercial** dataset.

length or domain specificity.

Third, the posterior action value ($\mathcal{O}^{c}_{\text{action}}$) quantifies consultation utility through subsequent user actions (e.g., clicks, purchases, searches(Zhang et al., 2025b)). The bucketing of action frequencies (see App. A.2) is a efficient and stable method to mitigate long-tail distribution issues and ensures robust normalization, providing sufficient granularity without overfitting to sparse data (Guo et al., 2017). The relative weighting of actions reflects their inherent scarcity (e.g., purchases > clicks), a principle supported by consumer behavior studies (Teubner and Graul, 2020).

The aggregated value ($\mathcal{O}^{s,c}_{u}$) combines these dimensions via weighted summation, ensuring interpretability and mitigating the sparsity issue of value scores. Ablation studies (Sec. 5.3) confirm that all components contribute to performance. The alignment loss $\mathcal{L}_{\text{VA}}$ further ensures that high-value consultation-action are captured by the model during training, though the temperature parameter τ_1 could be adapted per dataset to sharpen or soften attention supervision.

Dataset limitations (e.g., sparsity in Commercial data) are acknowledged, yet results generalize well to the larger Amazon dataset. Since we consider an online personalized search scenario, LLM-based baselines are not included due to their excessive inference and training time costs. In fact, VAPS builds upon MAPS's MoAE (Mixture of Attention Experts) to integrate textual semantics from a frozen LLM embedding layer, ensuring efficiency while maintaining strong performance. Besides, VAPS's modular design (e.g., CAI) could integrate LLM-enhanced consultations in future work.

In summary, while certain design choices (e.g., fixed decay rates, bucketing) are simplified for reproducibility, they are grounded in empirical evidence and previous works. The framework's flexibility allows for incremental refinements—such as dynamic thresholds or adaptive weighting—without undermining its core contributions. VAPS advances personalized search by assessing consultation value to users and aligning consultations with consumer actions, offering a scalable foundation for future enhancements.