# Qilin: A Multimodal Information Retrieval Dataset with APP-level User Sessions


Jia Chen[†]
Xiaohongshu Inc.
chenjia2@xiaohongshu.com

Qian Dong[†]
Tsinghua University
dq22@mails.tsinghua.edu.cn

Haitao Li
Tsinghua University
liht22@mails.tsinghua.edu.cn

Xiaohui He
Xiaohongshu Inc.
manyou@xiaohongshu.com

Yan Gao
Xiaohongshu Inc.
yadun@xiaohongshu.com

Shaosheng Cao
Xiaohongshu Inc.
shelsoncao@gmail.com

Yi Wu
Xiaohongshu Inc.
xiaohui@xiaohongshu.com

Ping Yang
Xiaohongshu Inc.
jiadi@xiaohongshu.com

Chen Xu
Xiaohongshu Inc.
chenlin1@xiaohongshu.com

Yao Hu
Xiaohongshu Inc.
yaoohu@gmail.com

Qingyao Ai
Tsinghua University
aiqy@tsinghua.edu.cn

Yiqun Liu
Tsinghua University
yiqunliu@tsinghua.edu.cn



## ABSTRACT

User-generated content (UGC) communities, especially those featuring multimodal content, improve user experiences by integrating visual and textual information into results (or items). The challenge of improving user experiences in complex systems with search and recommendation (S&R) services has drawn significant attention from both academia and industry these years. However, the lack of high-quality datasets has limited the research progress on multimodal S&R. To address the growing need for developing better S&R services, we present a novel multimodal information retrieval dataset in this paper, namely Qilin. The dataset is collected from *Xiaohongshu*, a popular social platform with over 300 million monthly active users and an average search penetration rate of over 70%. In contrast to existing datasets, Qilin offers a comprehensive collection of user sessions with heterogeneous results like image-text notes, video notes, commercial notes, and direct answers, facilitating the development of advanced multimodal neural retrieval models across diverse task settings. To better model user satisfaction and support the analysis of heterogeneous user behaviors, we also collect extensive APP-level contextual signals and genuine user feedback. Notably, Qilin contains user-favored answers and their referred results for search requests triggering the Deep Query Answering (DQA) module. This allows not only the training & evaluation of a Retrieval-augmented Generation (RAG) pipeline, but also the exploration of how such a module would affect users' search behavior. Through comprehensive analysis and experiments, we provide interesting findings and insights for further improving S&R systems. We hope that Qilin will significantly contribute to the advancement of multimodal content platforms with S&R services in the future.




*SIGIR '25, July 13–18, 2025, Padua, Italy.*



ACM ISBN xxxx...$15.00
https://doi.org/xxxx

**ACM Reference Format:**
Jia Chen[†], Qian Dong[†], Haitao Li, Xiaohui He, Yan Gao, Shaosheng Cao, Yi Wu, Ping Yang, Chen Xu, Yao Hu, Qingyao Ai, and Yiqun Liu. 2025. Qilin: A Multimodal Information Retrieval Dataset with APP-level User Sessions. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25), July 13–18, 2025, Padua, Italy*. ACM, New York, NY, USA, 11 pages. https://doi.org/xxxx


## 1 INTRODUCTION

Search engines and recommender systems play an essential role in online content platforms nowadays. Beyond textual content, mainstream User-Generated Content (UGC) communities usually provide illustrated texts or videos as results in either single-column or two-column display [19, 65]. These multimodal contents help users find desired information more conveniently, e.g., a note with images for each step of preparing a Tiramisu or even a video for the whole process is more intuitive for users to reproduce the flavor. Therefore, how to incorporate multimodal elements into the retriever is crucial for improving system effectiveness. Nevertheless, most existing datasets for general search or recommendation tasks mainly contain textual information or statistically dense features, which is deficient for investigating better multimodal search and recommendation (**S&R**) services.

To improve user satisfaction for a specific mobile application, in-depth user behavior analysis at the APP level can be crucial. As users' information needs become complicated, they may issue a series of related queries within a short interval to strive for better results, a.k.a., session search [2, 10, 11]. In other circumstances,

[†]These authors contributed equally to this work.

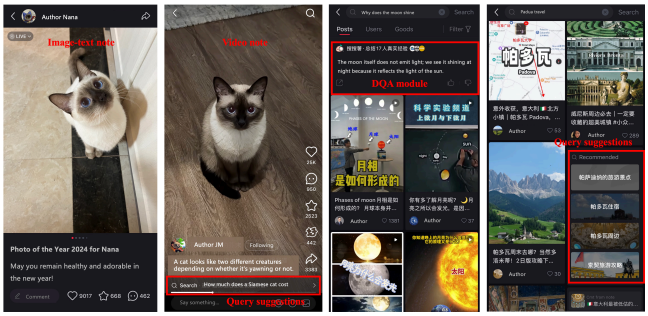



**Figure 1: Xiaohongshu leverages a two-column result list for S&R services, retrieving heterogeneous results like image-text, video, and commercial notes. The search system is equipped with a DQA module to provide direct answers for users. There are also various modules to stimulate users to search for any topics they might be interested in.**

users may be inspired by a recommended result or a query suggestion module [9, 42] to explore relevant topics via the search system. Users' APP-level tracks are very diverse, containing abundant contextual information. Some of these contexts/factors can be essential for estimating user satisfaction or long-term retention, e.g., Ozertem et al. [42] find that besides click signals, users' reformulating behavior can be regarded as a good proxy for modeling satisfaction. Therefore, beyond a single request, we should also focus on the session-level signals such as reformulating & revisiting actions, the source of user search intent, and the pattern of APP-level user transition behaviors (**S→R** or **R→S**) to better estimate long-term user satisfaction towards the whole application.

Due to the different forms of information displayed in multimodal scenarios, traditional behavioral analysis results may no longer be applicable. Predicting user satisfaction in a multimodal S&R system with heterogeneous functional modules is still challenging. For example, modern search engines usually incorporate a Deep Query Answering (DQA) module to provide users with succinct, direct answers. Typically, a DQA module operates through a Retrieval-Augmented Generation (RAG) pipeline [20, 52] which firstly retrieves relevant documents and then prompts a LLM for self-consistency check and summarization. When a direct answer is presented, user browsing behavior will be greatly impacted, e.g., they may focus more on the top results while interacting less with other organic ones [58]. De facto, the influence of the DQA module on user satisfaction and retention still remains largely under-investigated. To better evaluate the effect of the RAG module on users' perceived experience, the academia highlights the need for a dataset that includes both genuine user feedback on the DQA module and contextual user behaviors before and after they engage with such a module.

To shed light on the aforementioned issues, we present a novel multimodal dataset with large-scale APP-level user information discovery sessions in multiple scenarios (including search, DQA, and recommendation), namely Qilin. All user sessions are collected from *Xiaohongshu* [1], which is a popular social platform as well as the largest lifestyle search engine in China, with over 300 million monthly active users and a search penetration rate of over 70%. Different from other platforms, *Xiaohongshu* provides heterogeneous results such as image-text notes, video notes, commercial notes, and direct answers (as shown in Figure 1), posing challenges for optimizing both search and recommendation. To sum up, the main novelties of Qilin are listed as follows:

- To the best of our knowledge, Qilin is the first practical multimodal S&R dataset with heterogeneous results collected from a social media platform. Genuine user behaviors and multimodal content features facilitate the training of sophisticated neural architectures and even LLMs in various task settings.
- Besides side information like request and item features, we also contain abundant APP-level contextual signals (e.g., query sources, request history, timestamps, position, etc) and multiple user feedback toward the whole system. These signals are crucial for the in-depth investigation of user state transitions, revisits, and query reformulations to model user satisfaction or long-term retention.
- For the search requests triggering the DQA module, we further collect user-favored answers and their referred results as the positive instance under the specific query. To this end, Qilin can not only be taken as a benchmark for training or evaluating an RAG module but also be used for exploring the influence the module brings to user behavior.

To facilitate the reproducibility of this work, all resources for Qilin, code implementation, and related experiment details have been released in the repository below [2].

## 2 RELATED WORK

### 2.1 Multimodal Information Retrieval

Multimodal information retrieval has been extensively studied by Information Retrieval (IR), Computer Vision (CV), and Multimedia communities. To facilitate understanding, we systematically categorize related studies by approaches and task types, respectively.

Generally, multimodal retrieval approaches can be divided into three groups: 1) *representation learning* based approaches [6, 18, 28, 51, 56, 68], 2) *modality fusion or interaction* based approaches [23, 32, 61], and 3) *hybrid modeling* approaches [23, 33, 35, 44]. Representation learning-based approaches aim to map a modality (usually an image) into a binary Hamming space with hash functions [36, 68] or to encode it into a latent semantic space [6, 18]. Generally, hash-aware methods have low storage costs and are efficient in real-time infrastructure [64, 67]. In contrast, semantic-based ones focus more on modality understanding and cross-modality matching with deep neural networks (DNNs). Although DNNs show better generalization compared with traditional approaches [66], architectures with modality fusion and interaction usually achieve higher performance. To address intra-modal reasoning and cross-modal alignment, Qu et al. [44] develop a dynamic modality interaction network with a routing mechanism. Consequently, the hybrid modeling of combining modality representation with multimodal feature interaction tends to outperform single modeling approaches.

[1] Known in English as *rednote*, official site: www.xiaohongshu.com.

[2] https://github.com/RED-Search/Qilin

For task type, related literature can be classified into *cross-modal retrieval* [18, 37, 56] and *multi-modal retrieval* [6, 23, 28]. In broad terms, multimodal retrieval refers to requests containing queries or results with modalities beyond text (e.g., images, videos, or audio). Specially, cross-modal retrieval usually involves unimodal queries and results but with different modalities, e.g., text-to-image retrieval [18, 48, 54], text-to-video retrieval [32, 53], image-to-caption retrieval [18, 25, 37], etc. Balaneshin-kordan and Kotov [6] jointly model the text-image embedding space by cross-modal alignment and unified representation learning. Their approach can handle both cross-modal (text-to-image, T $\rightarrow$ I) and multimodal tasks (I $\rightarrow$ IT, T $\rightarrow$ IT). Besides *single query sessions*, researchers also aim at optimizing *multi-query sessions* [34, 61]. For example, Xie et al. [61] enhance image search by leveraging session contexts such as historical queries and engaged images. In addition, Liu et al. [34] utilize a heterogeneous graph neural network (HGN) to model intra-query, inter-query, and inter-modality information diffusion in multi-query product search. Experimental results have shown the usefulness of session-level contexts in user intent modeling and multimodal information matching.

## 2.2 Heterogeneous User Behavior Analysis

Analyzing heterogeneous user behavior and further exerting specific patterns in corresponding scenarios is essential for user satisfaction modeling and system optimization. To investigate user behavior in vertical search scenarios, researchers conduct in-depth log analysis for visual search [13] or eye-tracking study for image search [60], respectively. Their findings help improve the corresponding ranking algorithms and system layout. Besides, numerous studies have explored users' intent transition while interacting with a retrieval system [9, 46, 57]. For example, Chen et al. [9] investigate differences in user query reformulation behavior from delicate aspects such as the reformulation reason, interface, and the inspiration source via a field study. While directly modeling user retention is challenging in practice, Wu et al. [57] discover that frequent revisits could indicate a stronger user stickiness. Although these studies present valuable insights for optimizing the system, they mainly discuss a single component in the S&R system. As one service can significantly influence user behavior in another, e.g., a certain number of search queries rise from the suggestions in a recommended result, we collect APP-level user sessions for Qilin. These sessions contain user feedback on search, recommendation, and DQA services, along with contextual information such as session IDs, search sources, and timestamps, to better analyze user tracks within and across services.

## 2.3 Datasets for Search and Recommendation

Datasets are the foundation of both improving and evaluating retrieval systems. So far, most widely used datasets for search [41, 59], recommendation [19, 38], and S&R [4, 31, 49] usually only contain textual contents or value-based features for items, which is deficient for building better multimodal retrieval systems. One exception is the e-commerce scenario [4, 45], where the system needs to rank the products with both titles and images. In this respect, Miao et al. [40] and Gong et al. [22] present datasets containing short titles and images for Taobao products. Besides the mentioned

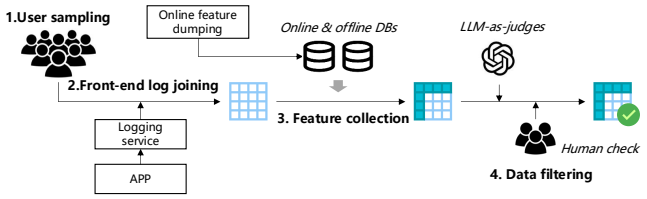



**Figure 2: The data construction process of Qilin. The front-end log is joined with sampled user IDs to obtain the dataset backbone. Then we collect features for the request, user, and note from various databases. Finally, all content features undergo rigorous filtering by LLMs and human experts.**

datasets, multimodal retrieval datasets such as UniIR [55] and Flickr30K [43] have also been developed. These datasets mainly contain factoid queries such as "find a picture of a person riding a horse" which are characterized by clear intent and typically have ground truth matches. However, in practical S&R scenarios, user intents tend to be ambiguous. Although these datasets are valuable for image-text modality alignment, they fall short when directly applied to the training of a general retrieval system. Such a system requires the ability to capture complex user intents behind both factoid and non-factoid queries and retrieve pertinent results correspondingly.

In this regard, the Qilin dataset features a substantial number of non-factoid queries paired with multimodal results, which is quite challenging. Moreover, collected user sessions contain abundant textual and image contents, facilitating the application of all aforementioned approach groups along with the investigation of complicated retrieval scenarios. This complexity poses challenges for not only multimodal intent understanding but session-level cross-modal matching as well.

# 3 THE QILIN DATASET

In this section, we delve into the details of the Qilin dataset. We first elaborate on the data preprocessing and construction pipeline. Then, a data schema of Qilin is presented to provide a clear glance at the information it includes. Finally, we discuss the potential research tasks that Qilin may support in various scenarios such as search, recommendation, retrieval-augmented generation, etc.

## 3.1 Data Construction

As shown in Figure 2, the overall data construction pipeline of Qilin includes several steps: user sampling, front-end log joining, feature collection, and data filtering.

**User sampling.** To contain as much APP-level contextual information as possible, we randomly sample from the raw data w.r.t. the user IDs rather than the page views (PVs). All users are categorized into an engagement level according to their interaction frequency. We first sample 15k non-spam users from the total pool of users who exhibited core interactive behaviors within the APP on November 27th, 2024. These users show diverse behavioral patterns while interacting with the APP. However, most of the above search requests did not trigger the DQA module. To support intensive investigations concerning DQA, we select about 3,000 users

who explicitly engaged with the module on the same date. Finally, we obtain a merged list with 17,576 user IDs.

**Front-end log joining.** Next, we join a specific user ID with the front-end log to obtain request-level information and user feedback. Here we only use the log on the same date to control the final data size. For each recommendation request, we collect all exposed results with their positions, timestamps, and five user feedback signals such as click, like, comment, etc. As partial results are video notes, we also record viewing time as an implicit feedback label. Besides the above information, a search request also contains a query and a source of the query keyword. The query source records the interface type from which the current query originates. For special requests with the exposure of a direct DQA answer, we reserve the answer content, the referred result note IDs, and four feedback labels (e.g., whether the answer is liked by the user).

All the requests can be identified by a session ID, a request ID, and a user ID, where a session contains all requests from the APP being opened to closed through a user. Based on these identifiers, interactive signals, interfaces, and timestamps, we can easily recover APP-level user tracks and further analyze their behavioral patterns across multiple services. Furthermore, contextual information such as exposure positions enables the investigation of unbiased learning to rank algorithms [3, 7] and click models [11, 12, 30].

**Feature collection.** Requests and feedback labels composite the backbone of our dataset. Next, we join the requests with note feature and user feature tables to support the training of sophisticated search or recommendation models. For each note, we contain fields such as title, content, cover, and non-cover image IDs. Note that there is a proportion of video notes (about 20~40%), we extract the key frames of a video as the image data to compress the final data size and to protect user privacy. In addition, we also include other basic information such as note type, video duration, and multi-granularity taxonomies, as well as 30 statistically dense features like impression count, click count, etc. As for all users, demographic features (e.g., gender, age) and 40 encrypted dense features are provided. To further enhance user modeling in session search or context-aware S&R ranking, we additionally collect 20 recently clicked note IDs before users initiate a specific request.

**Data Filtering.** To efficiently leverage our data while maintaining its security, the filtering algorithm should be capable of excluding clearly unsafe content while preserving as much valid information as possible. Therefore, we prompt LLMs with detailed instructions to classify document safety based on the title and the content. The models we applied for text and image filtering are Qwen2.5-14B-Instruct [63] and Qwen2-VL-7B-Instruct [62], respectively. The latter is a multimodal LLM that supports images as input and has achieved state-of-the-art performance on various visual understanding benchmarks. As shown in Table 1, we point out four types of textual content that should be rigorously excluded. The prompt used for image filtering is similar but with more rules. Besides the four previously mentioned types, we also instruct the LLM to identify whether an image contains real human faces. The LLM should not only provide a safety label but also a short textual description of the detected image. Firstly, we filter out all the images with one or more portraits according to the output labels. To further protect user privacy, we use another text-only LLM with 14B parameters to classify the remaining images based on the descriptions. Subsequently, we select a subset of 3,000 images and recruit three experts to double-check whether this subset includes any minor unsafe content. Such a process is repeated several times until the manual verification is passed. As a result, we obtain a corpus with 1,983,938 notes and 5,006,181 images.

**Table 1: Prompts used for filtering textual and image data.**

| **MODEL**: Qwen2.5-14B-Instruct (Qwen2-VL-7B-Instruct) |
|---|
| **PROMPT_PREFIX**:<br>Imagine you are a content safety reviewer for text (or images). Please examine the following text (or image) and assess whether it contains any illegal or sensitive information:<br>*1. Pornographic/Obscene*: Includes explicit sexual descriptions, obscene language, etc.;<br>*2. Violence*: Includes bloody, terrifying, extreme violence, etc.;<br>*3. Political Figures*: Mentions or discusses information related to political figures;<br>*4. Private Information of Ordinary Individuals*: Involves privacy such as names, identifiers, contact information, addresses, etc.;<br>*5. Portraits*: Includes any real human faces. Note that the face of a virtual character (e.g., animation, comic, or game roles) is not a portrait;<br>**Additional Notes**:<br>⋆ Advertising content does not count as illegal or sensitive information.<br>⋆ Mentioning public figures (such as actors, singers, entrepreneurs, etc.) does not count as illegal or sensitive information unless they are political figures.<br>⋆ Return True only if the text clearly contains any of the specified illegal or sensitive content; otherwise, return False.<br>⋆ Any emoji is not pornographic content.<br>⋆ Normal romantic-related content or sexual education content is not considered pornographic; but if the content is overly explicit or detailed, then it counts.<br>⋆ Expressions of negative emotions, when not coupled with violent content, should not be classified as acts of violence.<br>⋆ Do not be too strict, as minor issues being marked as illegal or sensitive can result in false positives. If you are uncertain, it is acceptable to refrain from labeling the content as illegal or sensitive. Only when you are very sure, mark it as a positive instance and return True.<br>Text (Image) to be checked: [ |
| **PROMPT_SUFFIX**:<br>] Please only return True or False without any explanations. |

## 3.2 Statistics & Data Schema

Generally, Qilin comprises APP-level sessions from 15,482 users. Comparison of basic statistics between Qilin and existing S&R datasets (Amazon [24, 39], JD Search [31], KuaiSAR [49]) is given in Table 2. Among these datasets, only Amazon can be marginally adopted for studying multimodal S&R systems. However, it only offers pseudo queries derived from the product metadata, which might compromise the reliability of experimental findings due to the absence of real user search behaviors. Moreover, product titles and images in the Amazon dataset need to be crawled additionally, increasing the experiment cost. On the other side, JD Search and KuaiSAR merely provide anonymized item contents (i.e., encrypted word ID sequence) which may cause difficulties in interpreting model effectiveness. Fortunately, *Xiaohongshu* maintains an open community with abundant user-generated content (UGC),

**Table 2: Comparison between Qilin and existing S&R datasets. Note that queries in Amazon are pseudo ones. JD Search and KuaiSAR only provide anonymized item content. Besides S&R, Qilin also include user actions on DQA.**

| Property | Amazon | JD Search | KuaiSAR | Qilin |
|---|---|---|---|---|
| # Users | 192,403 | 173,831 | 25,877 | 15,482 |
| # Items | 63,001 | 12,872,736 | 6,890,707 | 1,983,938 |
| # Queries | 3,221 | 171,728 | 453,667 | 57,188[1] |
| # Actions | 1,689,188 | 26,667,260 | 19,664,885 | 2,498,594 |
| # Images[2] | ? | – | – | 5,006,181 |
| DQA info | × | × | × | √ |
| Item text | title+review | anon'd | anon'd | title+body |

[1] Qilin provides a subset of 2,932 queries triggering the DQA module.
[2] Images of Amazon need to be crawled on demand, thus no statistics here.

**Table 3: Data schema of Qilin.**

| Table | Key | Fields |
|---|---|---|
| Search | search id | query, session id, user id, 20 recently clicked note ids, query source (1-8), search result details (list, each element is encapsulated as {*note id*, *position*, *timestamp*, *six engage labels*}, sorted by ascending timestamp), DQA details (if triggered); |
| Rec | request id | session id, user id, 20 recently clicked note ids, recommendation result details (a list of encapsulated elements, ditto); |
| DQA | search id | all search fields, answer content, referred note ids, four user engage labels; |
| User | user id | gender, platform, age, fan number, follow number, 40 encrypted dense features; |
| Note | note id | note type, note title, note content, image id list[1], video duration, video height, video width, image num, content length, commercial flag, 1/2/3-level taxonomy ids, 30 statistical features (e.g., the number of monthly impressions or clicks); |

[1] For video notes, we select images by equidistant sampling from key frames.

aiming at facilitating human connection and mutual assistance. To ensure the comprehensiveness of our dataset, we release the original note content (title + main body + images) after a thorough filtering process.

Qilin is organized into tabular format as shown in Table 3. The first three rows are the backbone tables, which contain the basic information for each search or recommendation request. Besides the user ID, we also collect 20 recently clicked note IDs to support short-term user modeling. Unlike existing missing-at-random datasets [19], we remain all exposed results at various positions for both search and recommendation. When we know whether an item has been examined, then there is no need to preserve a high density of user-item interaction. Binary feedback labels include click and four engaging actions: like (clicking the heart icon in Figure 1), collect (clicking the star icon in Figure 1), comment, and share. Since these engaging actions are sparse, we hereby consider the viewing time in seconds as a label for model training.

For a search request triggering the DQA module, we record the content of the answer, the referred notes, and corresponding user feedback. As direct answers are presented at the top of the result page, we assume all of them have been browsed by users. To further distinguish subtle differences in user satisfaction towards the answer, we collect four user actions on the DQA module: 1) whether the answer is liked by the user, 2) whether the user clicks the reference superscript, 3) whether the user clicks the answer body, and 4) whether the user clicks the aggregated experience tab (to unfold the relevant passages contributing to the final answer). Given referred notes and user-favored answers, we can easily train an LLM within an RAG pipeline and evaluate its performance. Furthermore, researchers can also investigate the differences of user behaviors in search scenarios with and without the DQA module.

To facilitate the training of various approaches, we collect rich content-based and ID-based side features for users and notes. Retrieval approaches in industry can be broadly divided into two groups. The first group of approaches usually has a pyramid architecture, with enormous sparse ID embedding parameters but light weights for feature interaction, e.g., CTR models such as DCN-V2 [50]. These models are usually fast in convergence and have high inference speed. However, they are poor in generalization due to the shallow semantic representation. On the contrary, content-based approaches only involve word embeddings and exploit deep networks for representation learning and feature interaction, e.g., pre-trained models [16]. They are more robust to unseen data while requiring huge online computation resources for efficient deployment. Considering the above issues, we emphasize the significance of combining content with sparse features to construct better retrieval systems. For all notes, we attach an image list in chronological order, i.e., the first image in the list is the cover, which has a great influence on user clicking behaviors. Since video contents are informationally redundant, we equidistantly sample images from key frames besides the cover. All images are compressed in WebP format, which typically achieves an average of 30% more compression than JPEG without loss of image quality.

**Table 4: Potential tasks that Qilin supports.**

| Scenario | Tasks |
|---|---|
| Search | CTR prediction, click simulation, unbiased learning to rank, multimodal search, context-aware ranking, session search, (post) pre-training for web search, enhancing search by recommendation, pre-training for RAG, evaluation for RAG, multimodal RAG, query performance prediction, etc; |
| Rec | content-based recommendation, multimodal recommendation, session-based recommendation, unbiased recommendation, query (intent) recommendation, enhancing recommendation by search, etc; |
| General | multi-task learning, multi-scenario learning, heterogeneous user behavior analysis, multimodal alignment, multimodal LLM-as-judges, scaling laws fitting, etc; |

## 3.3 Potential Tasks

Based on the features and information available, we present in Table 4 the range of tasks that Qilin potentially supports. Firstly, Qilin enables content-based retrieval and reranking for both search and recommendation. By incorporating images into user intent or note content encoding, there is further space to explore multimodal search and recommender systems. In heterogeneous result pages,

there tend to be more user behavioral biases compared to traditional text-only pages. As a result, unbiased learning to rank [3, 7] becomes crucial for improving system fairness and sustainability. Next, by leveraging multiple user feedback labels across various scenarios, it could be feasible to efficiently train multi-task or multi-scenario learning frameworks [2, 8, 46]. Besides joint learning, information from one scenario can be selectively applied to enhance another. For example, user actions in recommendation could be integrated into search ranking algorithms to better model user preferences [47]. As our dataset involves user engagements with different modules and services, there may be interesting heterogeneous user behavioral patterns that warrant further exploration. Some other tasks include post pre-training for retrieval-augmented generation (RAG) [20], query performance prediction (QPP) [5], query (intent) recommendation [42], multimodal alignment, scaling laws fitting [17], and using multimodal large language models as judges [29]. Concretely, the LLM-based data filtering process in Section §3.1 can be regarded as an example of LLM-as-judges.

Last but not least, Qilin can be taken as a semi-finished benchmark when augmented with precise human annotations to support more tasks like user satisfaction modeling, entity-enhanced search, multimodal RAG evaluation, etc.

## 4 PRIMARY DATA ANALYSIS

In this section, we present primary data analysis for Qilin, mainly including four parts: 1) demographics, 2) engagement & result distribution, 3) transitions across services, and 4) query analysis.

### 4.1 Demographics

According to IP locations, 15,482 users come from more than 87 countries or regions, mainly including China (84.32%), Andorra (1.45%), Australia (1.22%), Iceland (1.19%), Malaysia (1.14%), Japan, South Korea, United States, United Kingdom, Canada, Ireland, France, Singapore, Austria, and Germany. 76.53% of them are female, while others are male, with a ratio slightly higher than the average level of the APP. This may be because we retain a certain proportion of search requests that trigger the DQA module, and women users tend to engage more within these requests. Most users are young and middle-aged individuals, in the age range of 16 to 40 years old (74.5%). There are also small proportions of users below 16 (3.88%) and over 40 (7.27%). As for the platform, iOS and Android accounts for 53.56% and 38.79%, respectively. The rest of the users issue requests from desktop websites or other mobile operating systems such as Harmony.

### 4.2 Engagement & result distribution

Based on the Qilin dataset, in this section we aim to analyze 1) user engagements across scenarios, 2) user behavioral biases, and 3) result type distributions.

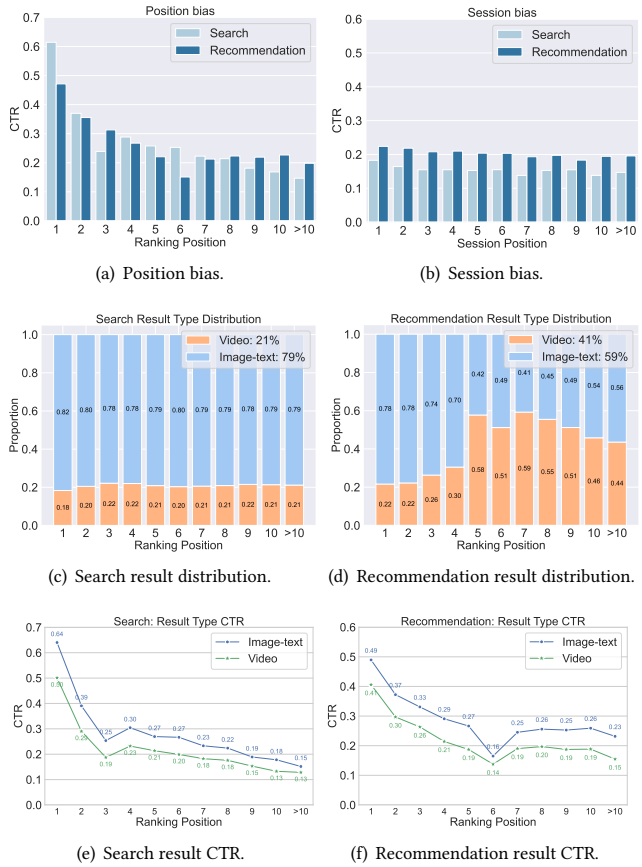
To explore user engagements, we consider two scenario pairs for comparison: S (Search) vs. R (Recommendation) and S+DQA (Search with DQA) vs. S-DQA (Search without DQA). As shown in Table 5, we have several findings. Firstly, compared to search, users click fewer results but have higher engagement rates except for the collecting action in the recommendation. With clearer intents, users may browse more results to find relevant or useful notes in the search service, resulting in a high browsing depth and click number. Once users find a helpful note, they may save it to the private collection for future demands. Overall, in the recommendation process, user intents are diverse and ambiguous. They may just click on several funny notes to kill time and usually scroll from one video to another. Therefore, recommendation users may browse fewer results on the original result page while browsing and liking more results in the embedded screaming. By comparing the right two columns, we find that the DQA module can greatly impact user behaviors. When provided with the direct answer, users only click several high-related notes at the top (i.e., with a lower first click position) and engage with significantly fewer notes. There could be two reasons: 1) Queries that trigger the DQA module tend to be factoid questions, where user information needs can be easily satisfied by a small number of notes. 2) By providing a direct answer with reference notes, the DQA module can save user effort in searching and summarizing useful information. Given this significant difference in user behavior, further exploration is required on how to rank the original results and design the result page layout to enhance users' search experience.

**Table 5: User engagements across scenarios, where S-DQA and S+DQA denote search requests without or with triggering the DQA module. Note that engagement rates such as like rate and collect rate are calculated based on the conditional probability $P(engage = 1|click = 1)$.**

| | S | R | S-DQA | S+DQA |
|---|---|---|---|---|
| *Avg. browsing depth* | 22.75 | 18.80 | 23.41 | 10.61 |
| *Avg. first click rank* | 3.01 | 4.97 | 3.03 | 2.50 |
| *Avg. click num* | 3.88 | 3.67 | 3.99 | 2.50 |
| *Click-through rate* | 21.01% | 24.13% | 21.02% | 20.73% |
| *Like rate* | 4.11% | 7.07% | 4.19% | 1.29% |
| *Collect rate* | 1.87% | 1.47% | 1.88% | 1.26% |
| *Share rate* | 0.57% | 0.63% | 0.57% | 0.52% |
| *Comment rate* | 0.32% | 0.99% | 0.32% | 0.21% |
| *# Samples* | 57,188 | 94,552 | 54,256 | 2,932 |

Next, we plot CTR across ranking positions and session positions in Figure 3(a) and Figure 3(b), respectively. For both search and recommendation, there is a decay in CTR when the ranking position increases, which is often referred to as the position bias. The decay is relatively slower in recommendation, suggesting that user browsing behavior may be more random compared to search, where users are more likely to click on top results. It is quite intriguing to observe a dip in the distribution at the third and the sixth positions for search and recommendation. We guess these positions have higher probabilities of exposing commercial notes [3], thus users are not very inclined to click on them. For the session dimension, we find that there also exists a decaying click rate when users issue more requests within an APP-level session. However, this decay is more gradual at the session scale. As search users are in a trial-and-error process, the corresponding CTR curve decays faster at the end of a session compared to the recommendation. Factors such as user fatigue or satisfaction may contribute to this session bias, i.e., users' accumulated cost or gain have reached the

[3]A commercial note is an image-text note or a video note with a product or advertisement, only accounting for a tiny proportion of the data.




(a) Position bias. (b) Session bias.

(c) Search result distribution. (d) Recommendation result distribution.

(e) Search result CTR. (f) Recommendation result CTR.

**Figure 3: Position bias, session bias, result type distribution, and CTR for two result types w.r.t. ranking positions in search and recommendation scenarios.**

upper limit, so they are more reluctant to click on a result in the later stage of sessions.

For result types, we only consider image-text and video notes. We first calculate the distribution proportion of each note type across ranking positions. As revealed in Figure 3(c) and Figure 3(d), the proportion of video notes in both S&R scenarios initially rises and then levels off. By comparing the two figures, it is obvious video notes are exposed more in the recommendation. This phenomenon may be caused by prolonged user-system in-loop interaction. Firstly, users are more inclined to view videos to kill time when using recommendation services. After some time, the recommender system may have learned this pattern and automatically ranks more videos at the top positions. Users continue to click on these video notes, resulting in a huge difference in the exposure structure between the search and recommendation scenarios. Generally, users have a higher CTR on image-text notes in both scenarios (see Figure 3(e) and 3(f)). We guess there are two main causes: the difference in browsing effort and the limitation of the user environment. Video notes typically require more time to complete viewing, contain less textual information per screen, and may produce noise in the surrounding environment. These limitations

**Table 6: User transition analysis. S and R stand for search and recommendation, respectively.**

| | S→S | R→S | R→R | S→R |
|---|---|---|---|---|
| *Proportion* | 34.88% | 2.95% | 59.32% | 2.84% |
| *Avg. click num* | 4.0014 | 3.9782 | 2.9745 | 4.4466 |
| *Click-through rate* | 20.18% | 22.98% | 23.20% | 25.01% |
| # Samples | 29,295 | 2,476 | 49,815 | 2,387 |

**Table 7: Query length distribution and corresponding user engagement. Note that a deeper color denotes a higher value of CTR or average click number.**

| | Proportion | CTR | Avg. click num |
|---|---|---|---|
| *Length<=3* | 0.32% | 0.1967 | 2.9892 |
| *Length=4* | 13.64% | 0.1962 | 4.0276 |
| *Length=5* | 13.94% | 0.2047 | 3.9303 |
| *Length=6* | 17.27% | 0.2051 | 4.0338 |
| *Length=7* | 14.07% | 0.2111 | 3.9446 |
| *Length=8* | 12.42% | 0.2150 | 3.8180 |
| *Length=9* | 9.54% | 0.2199 | 3.6971 |
| *Length=10* | 6.89% | 0.2272 | 3.7497 |
| *Length>10* | 11.91% | 0.2216 | 3.7010 |

make users more inclined to click on image-text notes in most circumstances. However, video notes can offer greater advantages in certain situations, such as for entertainment or when there is a need for detailed moment-level information. To this end, more investigation should be conducted on selecting the appropriate note type for specific user intents or request conditions to improve user satisfaction in the future.

## 4.3 Transitions across services

This section presents user transitions across S&R services. As APP-level sessions may contain noises in user intent transformation, we redefine a session for more accurate transition analysis. If the beginning timestamps of two requests (search or recommendation) from one user are within the 30-minute gap, then they are considered to belong to one session. As shown in Table 6, users mainly transfer within search or recommendation services. There is a certain proportion of transition from R→S (2.95%) and S→R (2.84%). When transferred from recommendation, users tend to click search results with a higher click-through rate compared to the general search condition. Similarly, users also have higher engagements with exposed results when transferring from search to recommendation, indicating the potential of considering these transition behaviors for better user preference modeling [46].

## 4.4 Query analysis

Following, we analyze search queries in the Qilin dataset, including character-based query length, query reformulation types, and search sources. Firstly, we calculate the proportion of queries with different lengths and further compare user engagement under these query groups. From Table 7, we can observe that most queries contain 4-8 characters, under which users tend to click on more results (i.e., the middle-attention of the rightmost column). These

**Table 8: Examples of different query reformulation types. The total number of reformulating actions is 29,875.**

| Type | Prop. | Examples |
|---|---|---|
| *Add* | 13.57% | 广州批发 → 广州批发市场 (Guangzhou wholesale→Guangzhou wholesale market)<br>五大文明 → 五大文明发源地 (the five great civilizations→the birthplaces of the five great civilizations) |
| *Delete* | 2.53% | ip 设计插件 →ip 设计 (intellectual property designing plugin→intellectual property design)<br>农业经营模式分析 → 农业经营模式 (analysis of agricultural business modes→agricultural business modes) |
| *Change* | 47.16% | 文案破碎感 → 文案自由松弛感 (writing sense with vulnerability→writing sense with freedom and relaxation)<br>将数据分三分位 → 四分位间距 (dividing data into tertiles→Interquartile Range, IQR) |
| *Repeat* | 7.57% | usmtekun 美食 →usmtekun 美食 (delicacy of usm tekun cafe, which refers to a cafe in Malaysia)<br>脸部画法 → 脸部画法 (techniques of drawing a face) |
| *Others* | 29.17% | 相机镜头进灰 → 适马 56 (the camera len has dust in it→SIGMA 56mm camera lens)<br>挡脸怎么弄好看 → 文案配照片 (cover the face in a visually appealing way→match captions with photos) |

queries clearly express user intentions and are not too difficult, allowing the search system to return more relevant results. Intriguingly, the click-through rate keeps rising as the query length increases. Apart from the fact that longer queries tend to have clearer intent and better results to be matched, users are generally more eager to obtain satisfactory answers when issuing longer queries, e.g., task-oriented ones. Therefore, they are more likely to click a search result under longer queries.

Secondly, we aim to explore users' query reformulation actions. Following [9, 27], we define five types of reformulating behaviors based on syntax relationships between two consecutive queries $q_{t-1}$ and $q_t$. Given $+q_t = \{w|w \in W(q_t), w \notin W(q_{q-1})\}$, $-q_t = \{w|w \notin W(q_t), w \in W(q_{t-1})\}$, and $\cap q_t = \{w|w \in W(q_t), w \in W(q_{t-1})\}$ where $w$ and $W(\cdot)$ denote a word [4] and the word set, we normalize the expression of Add, Delete, Change, Repeat, and Other reformulations as follows:

$$Add : +\Delta q_t \neq \emptyset, -\Delta q_t = \emptyset; \quad (1)$$

$$Delete : +\Delta q_t = \emptyset, -\Delta q_t \neq \emptyset; \quad (2)$$

$$Change : +\Delta q_t \neq \emptyset, -\Delta q_t \neq \emptyset, \cap q_t \neq \emptyset; \quad (3)$$

$$Repeat : +\Delta q_t = \emptyset, -\Delta q_t = \emptyset, \cap q_t \neq \emptyset; \quad (4)$$

$$Others : +\Delta q_t \neq \emptyset, -\Delta q_t \neq \emptyset, \cap q_t = \emptyset; \quad (5)$$

Based on the above definitions and the S→S sessions obtained from Section §4.3, we present the proportions of various syntactic query reformulation types and their corresponding examples in Table 8. Type "Change", "Add" and "Others" account for about 90% of all samples, which is consistent with previous findings in general search [9]. The balanced distribution of all types implies that users' search propensity in *Xiaohongshu* is quite complex. When user search intent is specialized or generalized, they may add or delete keywords based on the last query. For example, having issued a query "the five great civilizations", the user may be curious about the origin of these civilizations and subsequently submit a second query. When users are immersed in a spontaneous information-seeking flow, their trust and stickiness toward the system can be substantial. Therefore, it will be crucial to improve context-aware user search experiences for the whole S&R system.

Besides query reformulations, query sources can also be used to model user state change. According to the interface functionality, tens of entries can be broadly categorized into eight types, listed in Table 9. We find that about half of the queries are from the auto-completion module, which is quite different from the general search engines such as Google and Baidu where users mostly issue a query via the input box by themselves [9]. About 6% of queries come from history, indicating users may have similar search intents within a time interval. Additionally, users will also click query suggestions presented in result pages or notes. Modeling future intents and then recommending potential queries for users at an appropriate time can further improve user engagement.

[4] As Qilin is Chinese-centric, we count a character as a word.

**Table 9: Query source definition and analysis.**

| | Name | Definition | Prop. |
|---|---|---|---|
| 1 | *Active search* | using the search input box | 33.70% |
| 2 | *Auto-completion* | using the auto-completion | 50.97% |
| 3 | *Search history* | revisiting search history | 5.78% |
| 4 | *Suggestion I* | clicking result page suggestions | 3.97% |
| 5 | *Suggestion II* | clicking suggestions in a result | 3.02% |
| 6 | *Search filter* | using search filters[†] | 1.37% |
| 7 | *Hotlist* | clicking on a hotlist | 0.63% |
| 8 | *Others* | all other types | 0.55% |

[†] Including filters such as "most recent notes", "video notes only", and etc.

## 5 EXPERIMENTS

In this section, we report the performance of baseline approaches on search, recommendation, and DQA tasks, respectively. Due to the page limit, we do not consider jointly optimizing multiple tasks and leave it as future work.

### 5.1 Search and Recommendation

We sort all exposed samples in chronological order and split them into training and testing sets with a ratio of 11:1 in hours, i.e., samples after 22:00 pm will be taken as the testing ones. Baselines include BM25, BERT [14] bi-encoder, BERT cross-encoder, DCN-V2 [50], and Visual Language Model (VLM) [62]. In recommendation, we use the concatenated titles of recently clicked notes as the pseudo query to model user preference. For BM25, BERT bi-encoder, and BERT cross-encoder, we consider query text, note title and note content as input. Additionally, VLM further integrates cover images into note encoding. As for DCN-V2, besides the query and note (title + note) embeddings generated by pre-trained BERT bi-encoders (mean-pooled hidden states), we also

**Table 10: Comparison of search and recommendation performances on various approaches.**

| Model | Search | | | |
|---|---|---|---|---|
| | MRR@10 | MRR@100 | MAP@10 | MAP@100 |
| BM25 | 0.3388 | 0.3467 | 0.2399 | 0.3139 |
| $\text{BERT}_{bi}$ | 0.5320 | 0.5359 | 0.3855 | 0.4536 |
| $\text{BERT}_{cross}$ | 0.5336 | 0.5386 | 0.3848 | 0.4533 |
| DCN-V2 | **0.5600** | **0.5653** | **0.4014** | **0.4683** |
| VLM | 0.5523 | 0.5563 | 0.3937 | 0.4601 |
| **Model** | **Recommendation** | | | |
| | MRR@10 | MRR@100 | MAP@10 | MAP@100 |
| BM25 | 0.5379 | 0.5418 | 0.3933 | 0.4634 |
| $\text{BERT}_{bi}$ | 0.6067 | 0.6087 | 0.4548 | 0.5183 |
| $\text{BERT}_{cross}$ | 0.6346 | 0.6362 | 0.4786 | 0.5394 |
| DCN-V2 | 0.6307 | 0.6321 | 0.4651 | 0.5278 |
| VLM | **0.6394** | **0.6409** | **0.4890** | **0.5477** |

utilize various dense features and contextual signals such as recently clicked notes provided in Qilin. We scale up DCN-V2's trainable parameters to 0.13B, which is comparable with the BERT-base-chinese model (0.1B). Except for the VLM, all these models are trained in an end-to-end manner. We choose Qwen2-VL-7B-Instruct [62] with 4-bit quantization as the backbone of VLM and train it with LoRA [26], which is a parameter-efficient fine-tuning approach. By setting lora_rank to 16, only about 10M VLM parameters are tuned to ensure a fair comparison. All the experimental details can be found in the released repository hereinabove.

For evaluation metrics, we choose Mean Reciprocal Rank (MRR) and Mean Average Precision (MAP), which can be formulated as:

$$MRR@K = \frac{1}{N}\sum_{i}^{N}\frac{\mathbb{I}(r_i \le K)}{r_i} \tag{6}$$

$$MAP@K = \frac{1}{N}\sum_{i}^{N}\frac{1}{R_i}\sum_{k}^{K} P_{i,k}\cdot\mathbb{I}(rel_{i,k}=1) \tag{7}$$

where $N$ denotes the total number of testing instances, $r_i$ represents the rank of the first positive result in the reranked list. $R_i$ and $P_{i,k}$ are the number of relevant results and the precision at $k$ metric for the $i$-th instance, respectively. As for $rel_{i,k}$, we use clicks as binary relevance labels for simplicity.

All results are shown in Table 10. For both search and recommendation, the BERT cross-encoder outperforms the bi-encoder, which is consistent with the findings of existing work. Explicit query and document interaction help the model better learn the relevance matching. By considering visual information, VLM further achieves better performance on two tasks, indicating the effectiveness of considering note images in both user modeling and note representing. Compared with these pre-trained models, DCN-V2 shows competitive performance in both tasks. It combines user history, sparse ID-based features, dense features, and pre-trained semantic embeddings altogether, thereby performing the best in search ranking. However, the advantage of DCN-V2 is relatively smaller in the recommendation. There may be two main reasons: 1) In our setting, the pseudo query we use in recommendation already summarizes user preference. Therefore, the margin between DCN-V2 and other approaches is narrowed. 2) Recommendation requires a higher model robustness to deal with the out-of-distribution problem. As DCN-V2 heavily depends on sparse features and lacks deep modeling of semantic signal matching, it may fall short in the recommendation. Overall, the experimental results have shown the great potential of considering multimodal features and rich contextual signals to optimize the retrieval system.

**Table 11: Comparison of Retrieval-augmented generation (RAG) performance across various LLMs.**

| Generation\Retrieval | | NA | BM25 | $\text{BERT}_{bi}$ | Oracle |
|---|---|---|---|---|---|
| ROUGE-L | GPT3.5 | 0.332 | 0.388 | 0.390 | 0.424 |
| | GPT4o-mini | 0.316 | 0.384 | 0.380 | 0.429 |
| | Qwen2.5-72B | **0.342** | **0.396** | **0.396** | **0.436** |
| | Llama3.3-70B | 0.290 | 0.349 | 0.347 | 0.396 |
| | GLM4 | 0.318 | 0.363 | 0.366 | 0.401 |
| BERTScore | GPT3.5 | 0.708 | 0.725 | 0.722 | 0.748 |
| | GPT4o-mini | 0.710 | 0.729 | **0.728** | 0.754 |
| | Qwen2.5-72B | **0.714** | 0.725 | 0.724 | 0.752 |
| | Llama3.3-70B | 0.699 | **0.730** | 0.727 | **0.760** |
| | GLM4 | 0.701 | 0.715 | 0.713 | 0.739 |

## 5.2 DQA

For the DQA task, all user-engaged answers are considered the standard answer for evaluation. We then adopt five popular LLMs (GPT3.5 [1], GPT4o-mini [1], Qwen2.5-72B-Instruct [63], Llama3.3-70B-Instruct [15], GLM4 [21]) to generate answers directly via a vanilla RAG pipeline in a zero-shot setting and evaluate their quality based on the standard answers. As represented in Table 11, we test the generation performance of these LLMs in four conditions: 1) using no reference document (**NA**), 2) using top five documents retrieved by BM25, 3) using top five documents retrieved by BERT bi-encoders, 4) using reference documents provided by Qilin (**Oracle**). Considering the answer quality at both syntactic and semantic levels, we use ROUGE-L and BERTScore (F1) as the evaluation metrics. From the table, we have several observations. Firstly, the generation performance increases in the order: NA < BM25 < $\text{BERT}_{bi}$ < Oracle, which is consistent with our expectation. Through comparison, we find Qwen2.5 achieves the highest ROUGE-L score while Llama outperforms other models in terms of semantic matching, indicating that different LLMs may have different capability focuses.

# 6 CONCLUSION

In this work, we have presented a novel multimodal S&R dataset Qilin. Comprising APP-level sessions from 15,482 users, Qilin provides both textual and image content for heterogeneous results. Besides, we have also collected abundant contextual signals such as query sources, multiple user feedback, and deep query answering (DQA) details to facilitate the investigation of various IR-related tasks. Through comprehensive data analysis covering demographics, user engagement, result distribution, and query patterns, we present multi-faceted insights for enhancing S&R systems. To better instantiate its application, we conduct preliminary experiments in search, recommendation, and deep query answering on Qilin. We believe these findings and insights will be valuable in developing more advanced multimodal retrieval systems.

# REFERENCES


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

[2] Wasi Uddin Ahmad, Kai-Wei Chang, and Hongning Wang. 2019. Context attentive document ranking and query suggestion. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 385–394.

[3] Qingyao Ai, Keping Bi, Cheng Luo, Jiafeng Guo, and W Bruce Croft. 2018. Unbiased learning to rank with unbiased propensity estimation. In *The 41st international ACM SIGIR conference on research & development in information retrieval*. 385–394.

[4] 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*. 645–654.

[5] Negar Arabzadeh, Maryam Khodabakhsh, and Ebrahim Bagheri. 2021. BERT-QPP: contextualized pre-trained transformers for query performance prediction. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 2857–2861.

[6] Saeid Balaneshin-kordan and Alexander Kotov. 2018. Deep neural architecture for multi-modal retrieval based on joint embedding space for text and images. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. 28–36.

[7] Jia Chen, Haitao Li, Weihang Su, Qingyao Ai, and Yiqun Liu. 2023. THUIR at WSDM Cup 2023 Task 1: Unbiased Learning to Rank. *arXiv preprint arXiv:2304.12650* (2023).

[8] Jia Chen, Jiaxin Mao, Yiqun Liu, Ziyi Ye, Weizhi Ma, Chao Wang, Min Zhang, and Shaoping Ma. 2021. A hybrid framework for session context modeling. *ACM Transactions on Information Systems (TOIS)* 39, 3 (2021), 1–35.

[9] Jia Chen, Jiaxin Mao, Yiqun Liu, Fan Zhang, Min Zhang, and Shaoping Ma. 2021. Towards a better understanding of query reformulation behavior in web search. In *Proceedings of the web conference 2021*. 743–755.

[10] Jia Chen, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2019. TianGong-ST: A new dataset with large-scale refined real-world web search sessions. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 2485–2488.

[11] Jia Chen, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2020. A context-aware click model for web search. In *Proceedings of the 13th International Conference on Web Search and Data Mining*. 88–96.

[12] Aleksandr Chuklin, Ilya Markov, and Maarten De Rijke. 2022. *Click models for web search*. Springer Nature.

[13] Arnon Dagan, Ido Guy, and Slava Novgorodov. 2021. An image is worth a thousand terms? analysis of visual e-commerce search. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 102–112.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:cs.CL/1810.04805

[15] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).

[16] Yixing Fan, Xiaohui Xie, Yinqiong Cai, Jia Chen, Xinyu Ma, Xiangsheng Li, Ruqing Zhang, Jiafeng Guo, et al. 2022. Pre-training methods in information retrieval. *Foundations and Trends® in Information Retrieval* 16, 3 (2022), 178–317.

[17] Yan Fang, Jingtao Zhan, Qingyao Ai, Jiaxin Mao, Weihang Su, Jia Chen, and Yiqun Liu. 2024. Scaling laws for dense retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1339–1349.

[18] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. 2013. Devise: A deep visual-semantic embedding model. *Advances in neural information processing systems* 26 (2013).

[19] Chongming Gao, Shijun Li, Wenqiang Lei, Jiawei Chen, Biao Li, Peng Jiang, Xiangnan He, Jiaxin Mao, and Tat-Seng Chua. 2022. KuaiRec: A fully-observed dataset and insights for evaluating recommender systems. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 540–550.

[20] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997* (2023).

[21] Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793* (2024).

[22] Yu Gong, Xusheng Luo, Kenny Q Zhu, Wenwu Ou, Zhao Li, and Lu Duan. 2019. Automatic generation of chinese short product titles for mobile display. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 9460–9465.

[23] Yangyang Guo, Zhiyong Cheng, Liqiang Nie, Xin-Shun Xu, and Mohan Kankanhalli. 2018. Multi-modal preference modeling for product search. In *Proceedings of the 26th ACM international conference on Multimedia*. 1865–1873.

[24] Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*. 507–517.

[25] Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research* 47 (2013), 853–899.

[26] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).

[27] Jeff Huang and Efthimis N Efthimiadis. 2009. Analyzing and evaluating query reformulation strategies in web search logs. In *Proceedings of the 18th ACM conference on Information and knowledge management*. 77–86.

[28] Katrien Laenen, Susana Zoghbi, and Marie-Francine Moens. 2018. Web search of fashion items with multimodal querying. In *Proceedings of the eleventh ACM international conference on web search and data mining*. 342–350.

[29] Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. LLMs-as-Judges: A Comprehensive Survey on LLM-based Evaluation Methods. *arXiv preprint arXiv:2412.05579* (2024).

[30] Jianghao Lin, Bo Chen, Hangyu Wang, Yunjia Xi, Yanru Qu, Xinyi Dai, Kangning Zhang, Ruiming Tang, Yong Yu, and Weinan Zhang. 2024. ClickPrompt: CTR Models are Strong Prompt Generators for Adapting Language Models to CTR Prediction. In *Proceedings of the ACM on Web Conference 2024*. 3319–3330.

[31] Jiongnan Liu, Zhicheng Dou, Guoyu Tang, and Sulong Xu. 2023. Jdsearch: A personalized product search dataset with real queries and full interactions. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2945–2952.

[32] Meng Liu, Xiang Wang, Liqiang Nie, Xiangnan He, Baoquan Chen, and Tat-Seng Chua. 2018. Attentive moment retrieval in videos. In *The 41st international ACM SIGIR conference on research & development in information retrieval*. 15–24.

[33] Shilei Liu, Lin Li, Jun Song, Yonghua Yang, and Xiaoyi Zeng. 2023. Multimodal pre-training with self-distillation for product understanding in e-commerce. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*. 1039–1047.

[34] Xinyi Liu, Wanxian Guan, Lianyun Li, Hui Li, Chen Lin, Xubin Li, Si Chen, Jian Xu, Hongbo Deng, and Bo Zheng. 2022. Pretraining Representations of Multimodal Multi-query E-commerce Search. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 3429–3437.

[35] Yiqun Liu, Kaushik Rangadurai, Yunzhong He, Siddarth Malreddy, Xunlong Gui, Xiaoyi Liu, and Fedor Borisyuk. 2021. Que2search: fast and accurate query and document understanding for search at facebook. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 3376–3384.

[36] Xin Luo, Ye Wu, and Xin-Shun Xu. 2018. Scalable supervised discrete hashing for large-scale search. In *Proceedings of the 2018 World Wide Web Conference*. 1603–1612.

[37] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. 2015. Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN). arXiv:cs.CV/1412.6632 https://arxiv.org/abs/1412.6632

[38] Benjamin M Marlin and Richard S Zemel. 2009. Collaborative prediction and ranking with non-random missing data. In *Proceedings of the third ACM conference on Recommender systems*. 5–12.

[39] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*. 43–52.

[40] Lianhai Miao, Da Cao, Juntao Li, and Weili Guan. 2020. Multi-modal product title compression. *Information Processing & Management* 57, 1 (2020), 102123.

[41] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human-generated machine reading comprehension dataset. (2016).

[42] Umut Ozertem, Olivier Chapelle, Pinar Donmez, and Emre Velipasaoglu. 2012. Learning to suggest: a machine learning framework for ranking query suggestions. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. 25–34.

[43] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*. 2641–2649.

[44] Leigang Qu, Meng Liu, Jianlong Wu, Zan Gao, and Liqiang Nie. 2021. Dynamic modality interaction modeling for image-text retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1104–1113.

[45] Zhaochun Ren, Xiangnan He, Dawei Yin, Maarten de Rijke, et al. 2024. Information Discovery in E-commerce. *Foundations and Trends® in Information Retrieval*

18, 4-5 (2024), 417–690.

[46] 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*. 1029–1039.

[47] 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*. 1313–1323.

[48] Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. 2014. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics* 2 (2014), 207–218.

[49] Zhongxiang Sun, Zihua Si, Xiaoxue Zang, Dewei Leng, Yanan Niu, Yang Song, Xiao Zhang, and Jun Xu. 2023. KuaiSAR: A Unified Search And Recommendation Dataset. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 5407–5411.

[50] Ruoxi Wang, Rakesh Shivanna, Derek Cheng, Sagar Jain, Dong Lin, Lichan Hong, and Ed Chi. 2021. Dcn v2: Improved deep & cross network and practical lessons for web-scale learning to rank systems. In *Proceedings of the web conference 2021*. 1785–1797.

[51] Wei Wang, Xiaoyan Yang, Beng Chin Ooi, Dongxiang Zhang, and Yueting Zhuang. 2016. Effective deep learning-based multi-modal retrieval. *The VLDB Journal* 25 (2016), 79–101.

[52] Xiaohua Wang, Zhenghua Wang, Xuan Gao, Feiran Zhang, Yixin Wu, Zhibo Xu, Tianyuan Shi, Zhengyuan Wang, Shizheng Li, Qi Qian, et al. 2024. Searching for best practices in retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 17716–17736.

[53] Yunxiao Wang, Meng Liu, Yinwei Wei, Zhiyong Cheng, Yinglong Wang, and Liqiang Nie. 2022. Siamese alignment network for weakly supervised video moment retrieval. *IEEE Transactions on Multimedia* 25 (2022), 3921–3933.

[54] Zihao Wang, Xihui Liu, Hongsheng Li, Lu Sheng, Junjie Yan, Xiaogang Wang, and Jing Shao. 2019. Camp: Cross-modal adaptive message passing for text-image retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*. 5764–5773.

[55] Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. 2025. Uniir: Training and benchmarking universal multimodal information retrievers. In *European Conference on Computer Vision*. Springer, 387–404.

[56] Jiwei Wei, Yang Yang, Xing Xu, Xiaofeng Zhu, and Heng Tao Shen. 2021. Universal weighting metric learning for cross-modal retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 10 (2021), 6534–6545.

[57] Qingyun Wu, Hongning Wang, Liangjie Hong, and Yue Shi. 2017. Returning is believing: Optimizing long-term user engagement in recommender systems. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 1927–1936.

[58] Zhijing Wu, Mark Sanderson, B Barla Cambazoglu, W Bruce Croft, and Falk Scholer. 2020. Providing direct answers in search results: A study of user behavior. In *Proceedings of the 29th acm international conference on information & knowledge management*. 1635–1644.

[59] Xiaohui Xie, Qian Dong, Bingning Wang, Feiyang Lv, Ting Yao, Weinan Gan, Zhijing Wu, Xiangsheng Li, Haitao Li, Yiqun Liu, et al. 2023. T2ranking: A large-scale chinese benchmark for passage ranking. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2681–2690.

[60] Xiaohui Xie, Yiqun Liu, Xiaochuan Wang, Meng Wang, Zhijing Wu, Yingying Wu, Min Zhang, and Shaoping Ma. 2017. Investigating examination behavior of image search users. In *Proceedings of the 40th international acm sigir conference on research and development in information retrieval*. 275–284.

[61] Xiaohui Xie, Jiaxin Mao, Yiqun Liu, Maarten de Rijke, Qingyao Ai, Yufei Huang, Min Zhang, and Shaoping Ma. 2019. Improving web image search with contextual information. In *Proceedings of the 28th ACM international conference on information and knowledge management*. 1683–1692.

[62] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, et al. 2024. Qwen2 Technical Report. arXiv:cs.CL/2407.10671 https://arxiv.org/abs/2407.10671

[63] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115* (2024).

[64] Fan Yang, Ajinkya Kale, Yury Bubnov, Leon Stein, Qiaosong Wang, Hadi Kiapour, and Robinson Piramuthu. 2017. Visual search at ebay. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. 2101–2110.

[65] Chao Zhang, Shiwei Wu, Haoxin Zhang, Tong Xu, Yan Gao, Yao Hu, and Enhong Chen. 2024. NoteLLM: A Retrievable Large Language Model for Note Recommendation. In *Companion Proceedings of the ACM on Web Conference 2024*. 170–179.

[66] Hanwang Zhang, Zheng-Jun Zha, Yang Yang, Shuicheng Yan, Yue Gao, and Tat-Seng Chua. 2013. Attribute-augmented semantic hierarchy: towards bridging semantic gap and intention gap in image retrieval. In *Proceedings of the 21st ACM international conference on Multimedia*. 33–42.

[67] Yanhao Zhang, Pan Pan, Yun Zheng, Kang Zhao, Yingya Zhang, Xiaofeng Ren, and Rong Jin. 2018. Visual search at alibaba. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 993–1001.

[68] Lei Zhu, Xu Lu, Zhiyong Cheng, Jingjing Li, and Huaxiang Zhang. 2020. Deep collaborative multi-view hashing for large-scale image search. *IEEE Transactions on Image Processing* 29 (2020), 4643–4655.