NavigationIR: An Interactive Image Retrieval System with Vision-Language Models and Guided Query Navigation

Anonymous ACL submission

Abstract

This paper introduces NavigationIR, an advanced interactive image retrieval system that enhances user query interpretation through guided navigation. Traditional methods often convert queries and images into embedding spaces to select relevant images based on similarity scores or extend user queries with simulated feedback. However, these approaches face two main issues: (1) ineffective exploration of user intent and (2) inefficient utilization of user feedback. NavigationIR proactively analyzes user queries and employs iterative guided navigation to refine search results leveraging a Vision-Language Model (VLM). Our system crafts contextually informed questions that delve into specific image attributes, drawing insights from initial retrieval candidates and user responses. Unlike previous methods, NavigationIR synthesizes these interactions into a cohesive, information-rich query representation that guides subsequent searches. Extensive evaluations on academic benchmarks including VisDial, MS COCO, Flickr30K, and a real-world industrial dataset demonstrate that NavigationIR significantly improves retrieval accuracy and efficiency compared to existing methods. Our demo can be visited by the following link: https://www.navigationllava.com/.

1 Introduction

Image Retrieval (IR) plays a crucial role across various domains, from internet search engines and catalog management to medical diagnostics and social media sharing (Cao et al., 2022; Zhang et al., 2024a; Chaudhuri et al., 2018; Jush et al., 2024; Nakata et al., 2024). However, interpreting textual queries remains a significant challenge due to the ambiguity and brevity of user inputs. Most existing methods focus on improving the extraction and matching of text and image features (Radford et al., 2021; Li et al., 2022; Zhai et al., 2023; Beyer et al., 2024). Yet, when confronted with limited information from brief queries, these methods can lead to a mismatch between the results retrieved and the actual intentions of the users (Zhang et al., 2024b; Barz and Denzler, 2017). Interactive Image Retrieval (IIR) has been developed to enhance retrieval by integrating simulated user feedback in a dialogue-based format, thus expanding textual queries and improving search accuracy (Guo et al., 2018; Huang and Zhang, 2023; Levy et al., 2023; Lee et al., 2024a). Despite progress, IIR still faces challenges when users struggle to express their needs precisely, such as in e-commerce environments where product varieties are wide and descriptions are complex.

- Ineffective exploration of user intent: Existing IIR systems often fail to proactively identify the most informative attributes from user queries and current image candidates, lacking the ability to prioritize relevant details crucial for accurate search refinement. (Fu et al., 2024; Mindlin et al., 2024)
- Inefficient utilization of user feedback: Many IIR methods augment the original query by directly concatenating it with question-feedback pairs, introducing irrelevant content and resulting in cumbersome queries. (Qazanfari et al., 2023; Zhu et al., 2024)

To address these challenges, we propose NavigationIR, a novel interactive image retrieval system that utilizes Vision-Language Models and guided query navigation to refine the retrieval process iteratively. NavigationIR constructs the retrieval as a series of guided dialogues between the user and the system, effectively narrowing down target images based on user feedback. By integrating a Vision-Language Model (VLM) (Yin et al., 2023), NavigationIR analyzes attributes from initial image candidates to generate contextually appropriate and informative questions using image sampling and attribute ranking strategies.

Through multi-turn interactions, the system summarizes user feedback, gradually converging on the user's intent and iteratively refining the search space. It implements a feedback loop that records addressed attributes and responses, guiding the generation of new, pertinent questions to maintain diversity and minimize redundancy. Three interconnected components—Chat, Search, and Question—compose the proposed system. Consequently, NavigationIR avoids static query expansion by dynamically synthesizing each interaction into a progressively refined query, enhancing retrieval accuracy and overall effectiveness.

Comprehensive experiments on benchmarks such as VisDial (Das et al., 2017a), MS COCO (Lin et al., 2014), Flickr30K (Young et al., 2014), and a real-world industrial dataset validate the superior performance and faster convergence of NavigationIR compared to typical IIR systems.

In summary, our contributions are as follows:

- We introduce NavigationIR, a groundbreaking interactive image retrieval system that employs guided query navigation using Vision-Language Models to generate context-aware and varied questions.
- NavigationIR implements an iterative strategy of synthesizing user feedback into progressively refined queries in each interaction, improving the focus and context of subsequent retrieval rounds.
- Extensive experiments on well-established academic benchmarks and a proprietary industrial dataset demonstrate the effectiveness and efficiency of NavigationIR compared to representative existing systems.

2 Related Works

Image Retrieval Image retrieval (Cao et al., 2022) has evolved from the use of hand-made features such as texture, color, and shape to deep learning techniques, with convolutional neural networks (CNN) significantly improving feature extraction (Radenović et al., 2016). Transformer-based models and frameworks like CLIP (Radford et al., 2021) and BLIP (Li et al., 2022) have further enhanced text-image alignment using contrastive learning, but often struggle with ambiguous queries (Zhang et al., 2024b; Barz and Denzler, 2017). Interactive Image Retrieval (IIR) improves query refinement by engaging users in iterative dialogues, better capturing user intent (Guo et al., 2018; Huang and Zhang, 2023; Levy et al., 2023; Lee et al., 2024a). Systems that incorporate guided query navigation have shown promise in effectively exploring user intent by proactively identifying informative attributes. Integrating Vision Transformers (ViTs) with Large Language Models (LLMs) enhances the understanding of complex images, allowing adaptive feedback through binary relevance or natural language input (Levy et al., 2023; Lee et al., 2024a). These developments underscore the value of interactive retrieval systems that utilize guided navigation to refine searches.

Visual Question Answering and Conversations Visual Question Answering (VQA) has progressed from simple one-step interactions to complex dialogues (Wu et al., 2017; Antol et al., 2015a). Initially, systems responded directly to user queries with images or answers (Antol et al., 2015b; Goyal et al., 2017; Wang et al., 2022; Lu et al., 2019; Zeng et al., 2022). Now, Visual Dialog involves progressive exchanges where users ask questions and systems provide contextually rich responses (Murahari et al., 2020), as seen in systems like VisualChatGPT (Wu et al., 2023), which integrate vision and language models for advanced interactions. Further advancements include Generative Visual Dialogues (Das et al., 2017b; Murahari et al., 2019; Zheng et al., 2021; Xu et al., 2021), where a Questioner-Bot (Q-bot) and an Answerer-Bot (A-bot) engage in natural conversations about images using reinforcement learning to aid in image identification. Various prompting strategies (e.g., chain-of-thought, ReAct, reflection) enhance the cognitive abilities of language models. However, efficiently combining large language models with vision models remains a challenge.

The most related works to us are (Zhu et al., 2024) and (Lee et al., 2024b). Different from our approach, (Zhu et al., 2024) iteratively rewrites the query to enhance the quality of retrieval, but our method employs a multi-turn vision-based dialogue mechanism that iteratively refines user queries with questions. (Lee et al., 2024b) encodes the image candidates into text embeddings and aligns them with the queries within the textual embedding space. In contrast, we directly utilize the VLLM for more accurate cross-modality alignment, which reduces the information loss during image-to-text transformation.

To summarize, our work builds on these technological advancements by incorporating Large

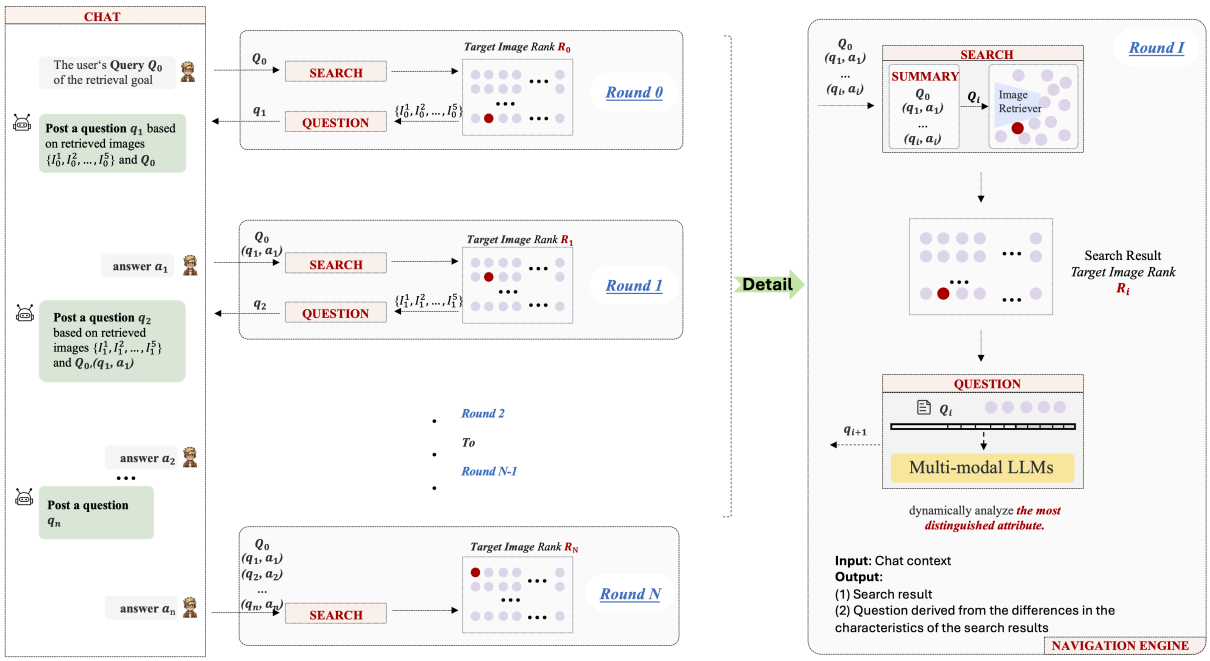


Figure 1: The left panel shows the overall workflow of the QuesIR interactive image retrieval, where the system refines the user's query through multi-turn interactions. The right panel details the Question Engine, which uses multi-modal LLMs to dynamically update the query based on user feedback, improving retrieval accuracy by focusing on key image features.

Language Models (LLMs) and V&L foundation models into the interactive image retrieval system. Unlike traditional VQA, which typically resolves queries in a single interaction, QuesIR iteratively refines user queries with the help of a multi-turn vision-based dialogue mechanism. This method not only significantly improves the accuracy of the retrieval, but also ensures that the system responses are more closely aligned with the user's intentions, representing a key evolution in the approach to interactive image retrieval.

3 Method

3.1 Problem Definition

The goal is to retrieve a target image I^* from an image database $\mathcal{D}$ based on user interactions, minimizing the number of iterations required to achieve a satisfactory rank for I^* in the retrieval results. Furthermore, the system is designed to exhibit strong generalization performance across different datasets.

3.2 System Framework

QuesIR operates through three interconnected components: **Chat**, **Search**, and **Question**, which interact over multiple dialogue rounds to approximate the user's target image iteratively.

3.2.1 Chat Component

The **Chat** component manages the multi-turn dialogue with the user, collecting initial queries and subsequent responses. It orchestrates the interaction, enabling iterative refinement of the search query through dialogue.

At the initiation round ($n = 0$), the user provides an initial textual description $Q^{(0)}$ of the target image I^*. This initial query is passed to the **Search** component, and the **Question** component returns a question $q^{(1)}$ based on the initial retrieval results. The user receives this question, initiating the next round of interaction.

In each subsequent iteration ($n \geq 1$), the user answers $a^{(n)}$ to the question $q^{(n)}$. The dialogue history is updated with the new question-answer pair $(q^{(n)}, a^{(n)})$. Modifiers $\delta^{(n)}$ are extracted from this pair to refine the search query, updating it as:

$$Q^{(n)} = Q^{(n-1)} + \delta^{(n)}. \quad (1)$$

3.2.2 Search Component

The **Search** component performs image retrieval based on the current query state and informs the **Question** component to guide subsequent interactions.

Query Refinement The refined query $Q^{(n)}$ is constructed by combining the initial description $Q^{(0)}$ with accumulated modifiers:

$$Q^{(n)} = Q^{(0)} + \sum_{i=1}^{n} \delta^{(i)}. \quad (2)$$

Modifiers $\delta^{(n)}$ are extracted from the user's responses to enhance the specificity of the query.

Image Retrieval Using the refined query $Q^{(n)}$, the **Search** component retrieves a ranked list of images $\mathcal{R}^{(n)} = \{I_1^{(n)}, I_2^{(n)}, \ldots, I_K^{(n)}\}$ from the image database $\mathcal{D}$. The rank of the target image I^* within $\mathcal{R}^{(n)}$ is recorded as Rank_n. The retrieval results are then passed to the **Question** component for further analysis.

3.2.3 Question Component

The **Question** component analyzes the retrieval results to generate contextually relevant questions aimed at reducing the search space.

Sampling Strategy A subset $\mathcal{S}^{(n)}$ is sampled from the retrieval results $\mathcal{R}^{(n)}$ to focus the analysis on the most informative images. The sampling criteria include selecting images satisfying $\text{sim}(I, Q^{(n)}) > \tau$ and ensuring that $\mathcal{S}^{(n)}$ contains images with diverse features to maximize informational gain. Notably, the target image I^* does not need to be included in $\mathcal{S}^{(n)}$, as the goal is to identify discriminative features among the top candidates.

VLLM Analysis and Question Generation The Vision Large Language Model (VLLM) analyzes $\mathcal{S}^{(n)}$ to identify discriminative features $f^{(n+1)}$ that differentiate images within the subset. It is prompted with instructions to identify the most significant differences among images in $\mathcal{S}^{(n)}$ and to generate a question $q^{(n+1)}$ based on $f^{(n+1)}$ to elicit specific information from the user. The VLLM incorporates the dialogue history $\{(q^{(i)}, a^{(i)})\}_{i=1}^{n}$ to avoid redundant or irrelevant questions.

3.2.4 Multi-round Iteration

The process repeats for multiple rounds, refining the search with each interaction. The dialogue continues until a termination condition is met—either when the target image I^* achieves a satisfactory rank Rank_n within the retrieval results or when a predefined maximum number of interaction rounds N is reached. Upon termination, the system presents the final retrieval results to the user.

Table 1 summarizes the Recall@K performance of our proposed QuesIR method compared to the baseline models across the VisDial, Flickr30K, and MS COCO datasets. Our method consistently outperforms the baselines at all recall levels and datasets, demonstrating the effectiveness of incorporating iterative questioning into the retrieval process.

4 Experiment

In this section, we evaluate the effectiveness of our proposed QuesIR method on three widely-used image retrieval datasets: **VisDial**, **Flickr30K**, **MS COCO**, and **ECO50K**. We aim to demonstrate that our method not only improves retrieval performance through iterative questioning but also generalizes well across different datasets.

4.1 Datasets

VisDial The Visual Dialog (VisDial) dataset (Das et al., 2017a) includes 120,000 COCO images, each paired with 10 rounds of question-answer dialogues. It is used to simulate interactive image retrieval, where iterative questioning improves search outcomes.

Flickr30K The Flickr30K dataset (Young et al., 2014) consists of 31,000 images with five captions each. Known for its linguistic diversity, it serves as a challenging benchmark for image captioning and retrieval tasks.

MS COCO The MS COCO dataset (Lin et al., 2014) contains over 120,000 images with rich annotations, including captions and segmentations for 80 object categories. It is widely used for evaluating retrieval methods across diverse, real-world scenes.

ECO50K We constructed an e-commerce dataset containing 50,000 product images to evaluate the applicability of our method in a shopping assistance scenario, assessing its performance in practical product search and recommendation tasks.

These datasets ensure a comprehensive evaluation of our method across various image types and retrieval challenges.

4.2 Baselines

We evaluate our method by comparing it with the following baselines:

Table 1: Performance (Recall@10) comparison across different datasets and methods. The improvement achieved by QuesIR is significant (p-value $\ll 0.05$).

	VisDial				Flickr30k				MsCOCO			
	R_1	R_3	R_5	R_{10}	R_1	R_3	R_5	R_{10}	R_1	R_3	R_5	R_{10}
ChatIR - Clip	-	-	-	-	70.0	73.8	75.2	75.2	73.1	77.1	79.1	79.3
NavigationIR - Clip	-	-	-	-	71.1	76.6	81.1	-	75.1	82.6	85.6	-
ChatIR - Blip	-	-	-	-	80.1	82.7	83.7	84.1	87.1	90.0	90.8	91.2
NavigationIR - Blip	-	-	-	-	81.3	85.4	89.2	-	88.2	92.7	94.2	-
ChatIR - Customized	69.8	75.7	78.3	81.3	81.0	84.2	86.0	88.0	91.1	93.3	94.7	95.1
NavigationIR - Customized	71.5	79.4	**83.8**	-	83.2	88.5	**91.9**	-	92.3	95.7	**97.1**	-

- **CLIP**: Based on the CLIP model (Radford et al., 2021), this baseline performs image retrieval by aligning image and text embeddings without task-specific fine-tuning. It represents a strong zero-shot retrieval approach.

- **BLIP**: Utilizing the BLIP model (Li et al., 2022), this baseline combines vision and language understanding for zero-shot image retrieval. It benchmarks performance in tasks that do not rely on iterative user interaction.

- **ChatIR**: This system approaches the target image through multiple rounds of questioning, but the questions are generated solely based on the dialogue context and do not change according to the retrieval results. It emphasizes interaction through contextual questioning rather than dynamically adjusting based on retrieved images.

These baselines were selected to emphasize the benefits of our iterative questioning approach over established retrieval techniques.

4.3 Implementation Details

To evaluate the effectiveness of our proposed NavigationIR method, we conducted experiments on the VisDial, Flickr30K, and MS COCO datasets using four different models: CLIP, BLIP, ChatIR, and our NavigationIR. For each dataset, we measured the Recall@10 metric from round 0 to round 5, allowing us to assess how quickly each method could retrieve the target images through iterative interactions.

For the VisDial dataset, which inherently supports interactive dialogue, we simulated user interactions by allowing the models to ask up to five questions per retrieval session. In contrast, for Flickr30K and MS COCO, which do not natively support multi-round dialogues, we adapted the datasets by creating synthetic dialogue contexts. Specifically, we generated clarifying questions based on image captions and annotations to mimic an interactive retrieval scenario.

In terms of sampling, we applied a top-50 random sampling strategy to select retrieval candidates during each iteration. The question generation model was powered by GPT-4o, which formulated context-aware queries for image clarification. The answering model also utilized GPT-4o, designed to provide relevant and consistent responses based on the context. This combination of models ensured that the simulated dialogues remained coherent and closely mirrored real user interactions, enhancing the robustness of our experimental setup.

4.4 Evaluation Metrics

We employed standard retrieval metrics to evaluate the performance of each method. For the VisDial, Flickr30K, and MS COCO datasets, we used Recall@10 at each interaction round. Hits@10 measures the proportion of times the correct image is among the top 10 retrieved results, providing insight into the models' ability to surface relevant images quickly.

4.5 Main Results

Our experiments reveal that NavigationIR significantly outperforms the baseline ChatIR in image retrieval tasks on VisDial, Flickr30K, and MS COCO datasets. NavigationIR achieves higher Hits@10 scores with fewer dialog rounds, demonstrating its efficiency in refining retrievals through iterative interaction. The consistent performance gains across different models and datasets underscore the method's robustness and scalability. Detailed re-

sults for each dataset are provided in the following subsections.

4.5.1 Performance on VisDial

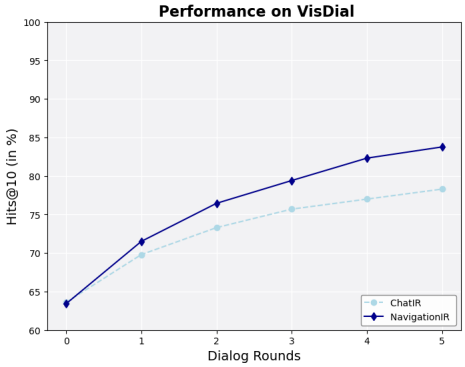


Figure 2: Performance on VisDial: Hits@10 performance over 5 dialog rounds.

As demonstrated in Table 1 and Figure 2, NavigationIR exhibits superior performance compared to ChatIR across all evaluation rounds. The performance gain becomes increasingly evident as the number of dialogue rounds increases, highlighting the method's ability to refine retrieval results through iterative interaction. Specifically, by round 5, NavigationIR achieves a Hits@10 score of 83.8%, representing a 5.5% improvement over ChatIR's 78.3%. This trend continues even when comparing NavigationIR's round 5 performance to ChatIR's round 10 performance, where NavigationIR outperforms ChatIR by 2.5 percentage points (81.3% vs. 83.8%). These results underscore the efficiency of NavigationIR in achieving higher retrieval accuracy within fewer rounds of interaction, demonstrating its robustness and effectiveness in the zero-shot setting on VisDial.

4.5.2 Performance on Flickr30K

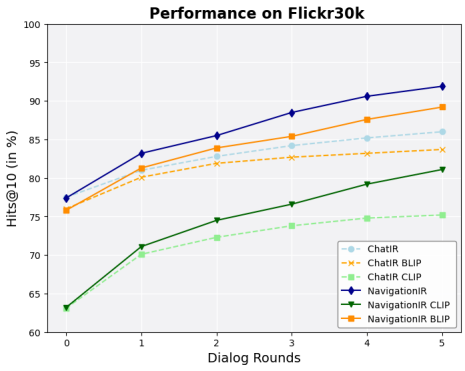


Figure 3: Performance on Flickr30k: Hits@10 performance across 5 dialog rounds for different models.

As presented in Table 1 and Figure 3, NavigationIR demonstrates a clear performance advantage over ChatIR across all retrieval models tested (CLIP, BLIP, BLIP-FT). The improvement becomes more pronounced with each additional round, showcasing the strength of NavigationIR in iterative retrieval tasks. By round 5, NavigationIR consistently achieves over a 5% performance gain across all models compared to ChatIR. In particular, with the BLIP-FT model, NavigationIR reaches a Hits@10 of 91.9%, which represents a significant 5.9% increase over ChatIR's 86% at the same round. Furthermore, even when compared to ChatIR's round 10 performance (88%), NavigationIR still maintains a 3.9% improvement. These results emphasize NavigationIR's effectiveness in improving retrieval precision through fewer interactions, consistently surpassing ChatIR across different retrieval models on the Flickr30K dataset.

4.5.3 Performance on MS COCO

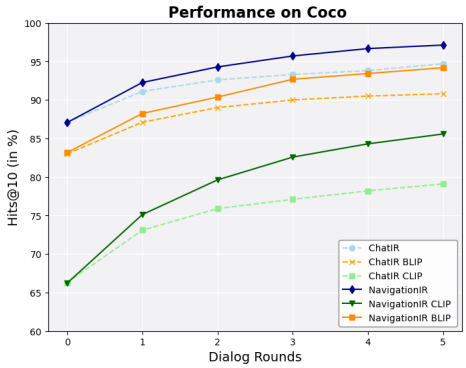


Figure 4: Performance on MS COCO: Hits@10 performance across 5 dialog rounds.

Table 1 and Figure 4 illustrate the superior performance of NavigationIR compared to ChatIR across different retrieval models (CLIP, BLIP, BLIP-FT) on the MS COCO dataset. NavigationIR consistently outperforms ChatIR at every round, with the performance gap widening as the rounds progress. By round 5, NavigationIR achieves a notable increase of more than 3% across all retrieval models. In the BLIP-FT model, NavigationIR reaches a Hits@10 of 97.1%, representing a 2.4% improvement over ChatIR's 94.7% in the same round. Even when comparing NavigationIR's performance at round 5 to ChatIR's round 10 result of 95.1%, NavigationIR maintains a 2% lead. These findings underscore the effectiveness of NavigationIR in enhancing retrieval performance, even in scenarios where the baseline accuracy is already high,

Table 2: Performance (Recall@10) comparison on NavigationIR across different datasets and backbone large language models.

	VisDial			Flickr30k			COCO		
	GPT-4o	qwen-VL	BLIP2	GPT-4o	qwen-VL	BLIP2	GPT-4o	qwen-VL	BLIP2
Round 0	0.634	0.634	0.634	0.774	0.774	0.774	0.871	0.871	0.871
Round 1	0.715	**0.716**	0.712	**0.832**	0.807	0.821	0.923	**0.928**	0.922
Round 2	**0.765**	0.762	0.758	**0.855**	0.832	0.846	0.943	0.945	**0.946**
Round 3	**0.794**	0.789	0.786	**0.886**	0.846	0.869	**0.958**	0.954	0.952
Round 4	**0.823**	0.816	0.814	**0.902**	0.863	0.877	**0.967**	0.962	0.956
Round 5	**0.838**	0.834	0.821	**0.907**	0.879	0.892	**0.971**	0.968	0.962

demonstrating its robustness and scalability across different models and settings.

4.6 Ablation Study

To systematically evaluate the performance of the question-generation model within the NavigationIR framework, we conduct ablation studies across three critical dimensions: (1) the effect of different Answer Model, (2) the influence of Summary Component (3) the influence of Different Sampling Strategies. Each of these components is explored in the following subsections.

4.6.1 Different Answer Models

To evaluate the impact of the answer model component within the NAVIGATIONIR framework, we tested three VLM models: GPT-4o, Qwen-VL, and BLIP2. As shown in Table 2, while the performance variation across these models is within a 1% margin, the choice of answer model can still help in better identifying the target. However, the core improvement comes from the question-generation model. This component drives the most significant gains in retrieval performance by providing more precise and feature-rich queries. Notably, even with BLIP2—the least effective model—our approach still yields an over 3% improvement compared to the baseline CHATIR method. These results highlight that, although different answer models can contribute to fine-tuning, the primary enhancement in retrieval effectiveness stems from the question-generation process, demonstrating the strength and adaptability of our framework.

4.6.2 Summary VS Without Summary

To evaluate the effect of the summary module on system performance, we conducted ablation studies on the Flickr30K dataset. As shown in Figure 5, The experiments showed that removing the

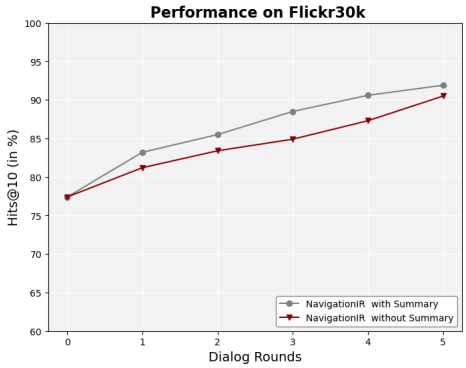


Figure 5: Performance on Flickr30k: With vs Without Summary

summary module resulted in a decrease of approximately 1% to 1.5% in Recall@10 across each interaction round, yet the performance remained superior to that of ChatIR. Specifically, at round 5, the model without the summary module achieved a Recall@10 of 90.5%, which is a 1.4% reduction compared to the full model with the summary module (91.9%). Despite this decrease, it still outperformed ChatIR by 2.5% (90.5% vs. 88%). These findings suggest that generating questions based on retrieval results significantly enhances the overall system performance. Additionally, the inclusion of the summary module contributes further improvements, underscoring its value in refining the effectiveness of our NavigationIR method.

4.6.3 Different Sampling Strategies

In this section, we evaluate the impact of three different sampling strategies: top-50 random sampling, stratified sampling, and clustered sampling. The results indicate that the overall performance remains largely unaffected across these strategies.

To further validate the robustness of the sampling process, we conducted an additional experi-

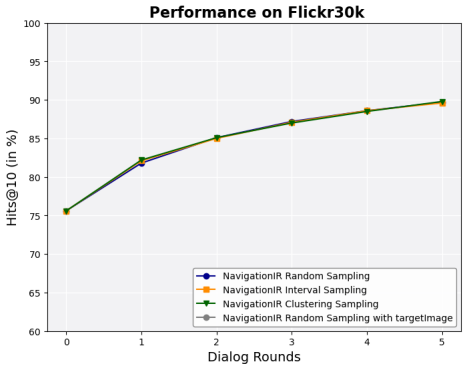


Figure 6: Performance on Flickr30k with Different Sampling Strategies

ment by comparing the inclusion and exclusion of the target image in the feature analysis set (comprising five images). The results showed minimal variation between the two conditions, reinforcing the rationale behind the use of random sampling in our methodology.

4.7 Application Scenario Evaluation on the ECO50K Dataset

We conducted an evaluation on the ECO50K dataset to test the effectiveness of our method in real-world applications. A total of 100 participants searched for products on the NavigationIR website (https://www.navigationllava.com/), each submitting 10 product queries. We recorded the number of conversational turns required to find the desired products.

The results show that chatIR required an average of 11 turns, while NavigationIR reduced this to 4, as shown in Table 3. This demonstrates that NavigationIR minimizes redundant interactions and performs efficiently across different shopping scenarios.

For details on the testing procedure, refer to the appendix or visit the website to experience it firsthand.

Table 3: Comparison of Average Conversational Turns between chatIR and NavigationIR Methods

Method	Average Conversational Turns
ChatIR	10.2
NavigationIR	4.9

5 Conclusion

This paper introduced NavigationIR, an interactive image retrieval system that enhances retrieval through Vision-Language Models and guided query navigation. By structuring the retrieval process as a dialogue, NavigationIR iteratively refines search results based on user feedback.

We conducted experiments on benchmark datasets, including VisDial, MS COCO, Flickr30K, and our newly created e-commerce dataset ECO50K, which contains 50,000 product images and descriptions. Results show that NavigationIR significantly improves both accuracy and efficiency compared to existing methods, demonstrating its robustness in real-world applications.

Future work will focus on integrating advanced open-source Vision-Language Models and optimizing models using Reinforcement Learning from Human Feedback (RLHF) to better extract distinctive features, thereby improving overall performance.

6 Limitation

While NavigationIR offers considerable improvements in interactive image retrieval, like any advanced system, it inevitably introduces certain challenges. The effectiveness of NavigationIR is contingent upon the availability of high-quality data; performance may be limited in environments where annotations are sparse or of substandard quality. Furthermore, as NavigationIR incorporates large language models, it is subject to the biases inherent in its training data, which could subtly influence the questions generated and the interpretation of user responses. These challenges are integral to the current scope of NavigationIR and provide directions for continuous improvement in future versions of the system.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015a. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015b. VQA: Visual Question Answering. In *ICCV*.

Björn Barz and Joachim Denzler. 2017. Automatic query image disambiguation for content-based image retrieval. *arXiv preprint arXiv:1711.00953*.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael

Tschannen, Emanuele Bugliarello, et al. 2024. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*.

Min Cao, Shiping Li, Juntao Li, Liqiang Nie, and Min Zhang. 2022. Image-text retrieval: A survey on recent research and development. *arXiv preprint arXiv:2203.14713*.

Abon Chaudhuri, Paolo Messina, Samrat Kokkula, Aditya Subramanian, Abhinandan Krishnan, Shreyansh Gandhi, Alessandro Magnani, and Venkatesh Kandaswamy. 2018. A smart system for selection of optimal product images in e-commerce. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 1728–1736. IEEE.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017a. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 326–335.

Abhishek Das, Satwik Kottur, José MF Moura, Stefan Lee, and Dhruv Batra. 2017b. Learning cooperative visual dialog agents with deep reinforcement learning. In *ICCV*, pages 2951–2960.

Yicheng Fu, Raviteja Anantha, Prabal Vashisht, Jianpeng Cheng, and Etai Littwin. 2024. Ui-jepa: Towards active perception of user intent through onscreen user activity. *arXiv preprint arXiv:2409.04081*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *CVPR*.

Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, Gerald Tesauro, and Rogerio Feris. 2018. Dialog-based interactive image retrieval. *Advances in neural information processing systems*, 31.

Fuxiang Huang and Lei Zhang. 2023. Language guided local infiltration for interactive image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6104–6113.

Farnaz Khun Jush, Tuan Truong, Steffen Vogler, and Matthias Lenga. 2024. Medical image retrieval using pretrained embeddings. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE.

Saehyung Lee, Sangwon Yu, Junsung Park, Jihun Yi, and Sungroh Yoon. 2024a. Interactive text-to-image retrieval with large language models: A plug-and-play approach. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 791–809, Bangkok, Thailand. Association for Computational Linguistics.

Saehyung Lee, Sangwon Yu, Junsung Park, Jihun Yi, and Sungroh Yoon. 2024b. Interactive text-to-image retrieval with large language models: A plug-and-play approach. *arXiv preprint arXiv:2406.03411*.

Matan Levy, Rami Ben-Ari, Nir Darshan, and Dani Lischinski. 2023. Chatting makes perfect–chat-based image retrieval. *arXiv preprint arXiv:2305.20062*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *ECCV*, volume 8693, pages 740–755.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *NeurIPS*, pages 13–23.

Dimitry Mindlin, Amelie Sophie Robrecht, Michael Morasch, and Philipp Cimiano. 2024. Measuring user understanding in dialogue-based xai systems. *arXiv preprint arXiv:2408.06960*.

Vishvak Murahari, Dhruv Batra, Devi Parikh, and Abhishek Das. 2020. Large-Scale Pretraining for Visual Dialog: A Simple State-of-the-Art Baseline. In *ECCV*, volume 12363 of *Lecture Notes in Computer Science*, pages 336–352. Springer.

Vishvak Murahari, Prithvijit Chattopadhyay, Dhruv Batra, Devi Parikh, and Abhishek Das. 2019. Improving generative visual dialog by answering diverse questions. *EMNLP*.

Kengo Nakata, Daisuke Miyashita, Youyang Ng, Yasuto Hoshi, and Jun Deguchi. 2024. Rethinking sparse lexical representations for image retrieval in the age of rising multi-modal large language models. *arXiv preprint arXiv:2408.16296*.

Hamed Qazanfari, Mohammad M AlyanNezhadi, and Zohreh Nozari Khoshdaregi. 2023. Advancements in content-based image retrieval: A comprehensive survey of relevance feedback techniques. *arXiv preprint arXiv:2312.10089*.

Filip Radenović, Giorgos Tolias, and Ondřej Chum. 2016. Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard examples. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 3–20. Springer.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework. In *ICML*, volume 162, pages 23318–23340.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023. Visual ChatGPT: Talking, drawing and editing with visual foundation models. *arXiv:2303.04671*.

Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2017. Visual question answering: A survey of methods and datasets. *Computer Vision and Image Understanding*, 163:21–40.

Zipeng Xu, Fandong Meng, Xiaojie Wang, Duo Zheng, Chenxu Lv, and Jie Zhou. 2021. Modeling Explicit Concerning States for Reinforcement Learning in Visual Dialogue. In *BMVC*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Yan Zeng, Xinsong Zhang, and Hang Li. 2022. Multi-Grained Vision Language Pre-Training: Aligning Texts with Visual Concepts. In *ICML*, volume 162, pages 25994–26009.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986.

Kai Zhang, Yi Luan, Hexiang Hu, Kenton Lee, Siyuan Qiao, Wenhu Chen, Yu Su, and Ming-Wei Chang. 2024a. Magiclens: Self-supervised image retrieval with open-ended instructions. *arXiv preprint arXiv:2403.19651*.

Xu Zhang, Zhedong Zheng, Linchao Zhu, and Yi Yang. 2024b. Collaborative group: Composed image retrieval via consensus learning from noisy annotations. *Knowledge-Based Systems*, page 112135.

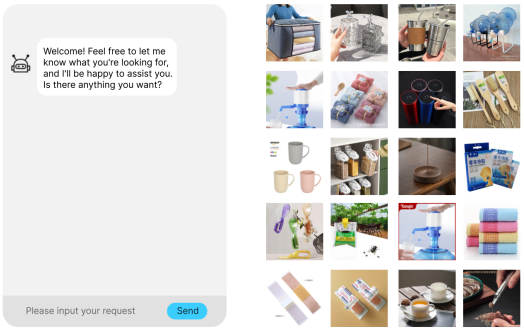
Duo Zheng, Zipeng Xu, Fandong Meng, Xiaojie Wang, Jiaan Wang, and Jie Zhou. 2021. Enhancing Visual Dialog Questioner with Entity-based Strategy Learning and Augmented Guesser. In *EMNLP*.

Hongyi Zhu, Jia-Hong Huang, Stevan Rudinac, and Evangelos Kanoulas. 2024. Enhancing interactive image retrieval with query rewriting using large language models and vision language models. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pages 978–987.

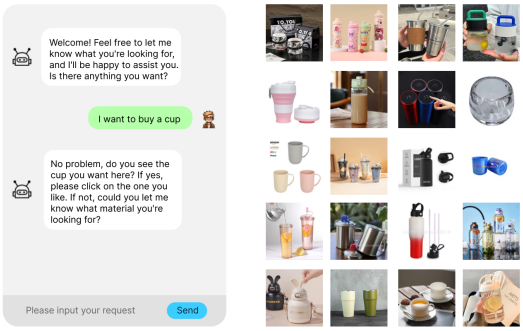
Appendix

Iterative Search Process in NavigationIR

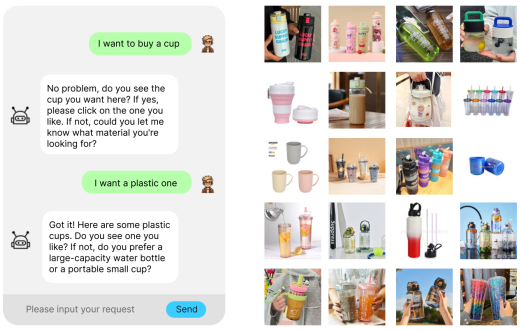
Initial State: The NavigationIR system informs the user that it can assist in locating any desired product.



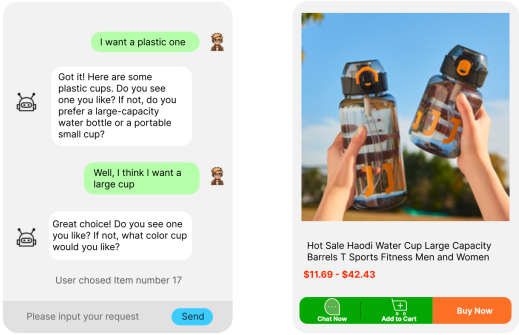
Round 0: The user submits an initial request, and NavigationIR retrieves and presents the top 20 search results based on the initial query. It then analyzes the feature differences among the search results and formulates follow-up questions based on these differences.



Round 1: The user responds to the proposed question, and the entire dialogue context is fed into the Navi-Engine. The system performs a new search and presents another set of top 20 results. It asks the user whether the desired product is among the presented items. If not, the system analyzes the feature differences in the new results and generates a subsequent question based on these differences.



This iterative process continues until the target product is identified within the top 20 results.



NavigationIR can be accessed for an interactive experience via the following link: https://www.navigationllava.com/.