Conversational search for image and video with augmented labelling

Anastasia Potyagalova

Specialist in Mathematics and System Programming

Supervised by Prof.Gareth J.F. Jones (DCU) and

Prof.Benjamin R. Cowan (UCD)

A thesis presented for the degree of Doctor of Philosophy

School of Computing

Dublin City University

June 2025

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Ph.D. is entirely my own work, that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed: Anastasia Potyagalova

(Candidate) ID No.: 20214756

Date: 24th June, 2025

Dedication

This dissertation is dedicated to my mother, Alena, and my grandad, Stanislav. They are in a better world, but it was they who taught me how to stay curious and never give up.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Gareth J. F. Jones, for guiding me through the entire PhD journey and helping me to grow as a researcher and supporting and inspiring me all the way. Additionally, I am grateful to my secondary supervisor, Benjamin R. Cowan, for his support, and to my colleagues from the School of Computing and the Adapt Centre, who provided an excellent opportunity to pursue my PhD I am grateful to all my participants; this research would be impossible without them. I am very grateful to my husband, Vasily, for his invaluable help and encouragement during my PhD journey, and I would express my gratitude to my aunt Anna and my closest friend, Julia, for their support and kindness. I am thankful to the Science Foundation Ireland Centre for Research Training in Digitally-Enhanced Reality (d-real) under Grant No. 18/CRT/6224, and partially as part of the ADAPT Centre at DCU (Grant No. 13/RC/2106_P2) (`www.adaptcentre.ie`) to support this research.

Contents

List of Figures

List of Tables

Conversational search for image and video with augmented labelling

Abstract

The rapid growth of media archives—including text, speech, video, and audio—has driven strong interest in developing advanced search methods for multimedia content. In particular, conversational search has emerged as a promising approach, where users engage in a dialogue with an AI agent to support and enhance their search activities. While most existing systems focus on text-based archives, this research extends conversational search methods to image and video retrieval.

Our approach involves developing an experimental framework to explore how conversational engagement can improve multimedia search. We introduce a prototype system that combines dialogue-based interaction with state-of-the-art visual indexing techniques. While multimedia information retrieval (MIR) has long been studied through conventional user-driven interfaces, the integration of conversational agents introduces a new layer of interactivity. The agent aims to assist users by suggesting relevant content and helping to filter out irrelevant results.

Effective dialogue in this context requires the agent to demonstrate an understanding of the content and its relevance to the user's needs. Although MIR techniques have advanced significantly, little attention has been paid to how retrieved content is represented and communicated during the search process. Our system addresses this gap by incorporating object detection to highlight key visual features, enhancing both the accuracy and contextual relevance of search results.

To evaluate the framework, we conducted three user studies focused on the effectiveness of conversational engagement in multimedia search. These studies examined how AI-driven dialogue affects users' ability to retrieve relevant image and video content and improves the overall search experience. Results indicate that conversational interaction not only refines retrieval accuracy but also increases user satisfaction by creating a more intuitive and responsive search environment.

Chapter 1

Introduction

The growth of multimedia archives has led to significant interest in developing search methods to enable the location of content of interest within these archives. An ongoing challenge of multimedia search is the specification of search queries and interaction with the retrieved content. In parallel with work in multimedia search, recent years have seen increasing interest in conversational search in which the user engages in a dialogue with an AI agent that supports their search activities [57] [17] [25] [48]. Conversational search seeks to enable users to find useful content more easily and reliably than traditional user-driven search interaction frameworks.

The term "conversational" in this research refers to systems that enable users to interact through natural language dialogue, mimicking human-like conversation to support task-oriented goals. In the context of search, a conversational interface allows users to iteratively refine their queries, receive system feedback, and explore results through a back-and-forth exchange. Unlike traditional search interfaces that rely on static keyword input, conversational systems are dynamic and context-aware, adapting to the user's intent as it evolves throughout the interaction.

Within this broad field, conversational search (CS) refers specifically to the use of dialogue-based interactions to support information retrieval. Traditional CS applications have focused primarily on text-based search assistance, helping users refine or reformulate their queries through conversation. However, recent advances have extended this paradigm to multimodal search, where the dialogue guides users not

only through text, but also through visual content such as images and videos.

This research is situated within this expanding space, investigating how AI-driven conversational agents can assist users in searching image and video archives. While conversational search for text has seen substantial development, there remains a notable gap in conversational frameworks designed for video search. This thesis addresses that gap by proposing a novel, multimodal conversational framework that supports interactive search across visual media — with particular emphasis on object detection, query reformulation, and dialogue-driven relevance feedback.

The main research focus for this PhD project is the investigation of how user experience in multimedia search might be enhanced by using a dialogue-based search framework. This will include consideration of how a conversational search agent might be integrated into a multimedia search system, including, for example, the creation of clarification questions and query rewriting.

Another key research focus is the investigation of the potential benefits of augmented reality (AR) features within the conversational search interface — specifically, the integration of augmented labels generated through object detection and contextual textual annotations. These visual and semantic overlays are embedded directly into the search interface, enriching the dialogue with information derived from the visual content itself.

By identifying and labelling objects within images, the system introduces an additional layer of interactivity and meaning to the search process. Users are not limited to text queries alone but can now engage with visual elements, selecting or referencing detected objects as part of the conversational flow. This integration enables a more intuitive search experience, where image content is no longer passive but actively shapes the direction and refinement of the dialogue.

Through this approach, the research aims to explore how the fusion of visual understanding and conversational interaction — supported by AR-style augmentation — can improve search clarity and user engagement in both image retrieval contexts.

1.1 Research Questions

This section focuses on investigating conversational engagement in multimedia search through a series of research questions.

1.1.1 RQ1: User experience in conversational MIR: How does user experience in MIR compare between a standard MIR system and an equivalent one integrating a conversational search agent?

This RQ focuses on exploring different aspects of the user experience using effective MIR systems. This research question contains several subquestions:

1. How can the multimodal conversational search system could be compared with a conventional search system?
2. What aspects of using multimodal features in a search dialogue can be used effectively in multimedia information retrieval?

1.1.2 RQ2: Can clarifying questions be used effectively to resolve ambiguity and improve search effectiveness in conversational MIR?

Clarification question is a query, textual or multimodal, issued by the search system to resolve ambiguity, refine the user's intent, or gather additional information to improve the accuracy and relevance of the search results [106]. RQ2 includes the following subquestions:

1. What are the opportunities and challenges for embedding clarifying questions into the conversational MIR framework?
2. Can multimodal clarification features advantage affect the user's search result preference and the user's perceived workload?

1.1.3 RQ3: Can augmenting media views with text object labels be used to improve the conversational search process in MIR?

RQ3 consists of the following subquestions:

1. Can augmented reality highlighted objects or textual labels in the search results make the user experience more convenient and efficient?

2. Which multimedia representation factors are important for a better user experience?

The next chapter introduces the state-of-the-art methods and findings of conversational search and multimedia information retrieval.

Chapter 2

Literature review

This chapter provides an overview of current multimedia information retrieval (MIR) methods, including techniques for image and video retrieval. It also introduces conversational search (CS) and explores various search interfaces, focusing on aspects such as user engagement, learning, and knowledge acquisition. Additionally, the chapter discusses the challenges of measuring learning and knowledge, along with methodologies to address these challenges. Finally, it provides an overview of query construction and refinement methods.

2.1 *Multimedia information retrieval*

Research in multimedia search encompasses a broad range of activities aimed at improving the retrieval of both static images and dynamic video content. The primary focus has been on advancing semantic analysis techniques to better understand and interpret the content. This includes the development of robust algorithms for object recognition, enabling systems to identify and classify objects within images with higher accuracy and contextual relevance. These advancements not only enhance the precision of retrieval systems but also play a critical role in bridging the semantic gap—the disparity between the visual data in multimedia content and its human-understood meaning. Together, these efforts form a foundational component of multimedia search, supporting more intuitive and effective ways for users to access

and interact with visual information.

2.1.1 *Image search*

The key challenge in image retrieval is the development of feature extraction methods. These seek to provide information regarding the objects contained in images. After feature extraction, these can be used for different purposes: to compare feature sets from different images or to create a text label set for the image, which can be used in text search processes [72]. In this research, we plan to use both. Image-to-text transformation enables the use of textual queries for retrieving images and videos. Additionally, comparing feature extraction methods is crucial when using images as queries in multimedia search. Staying up to date with the latest state-of-the-art techniques in multimedia information extraction and text generation is essential, as these advancements allow for more accurate and meaningful representations of visual content, thereby improving the effectiveness of image-based search queries. Content analysis is an important part of the image search process because effective and accurate object detection provides information that can be used during the search process [70]. For effective preprocessing of the image archive, which is used as a search database for the conversational application, it is important to perform feature extraction with high accuracy, since the relevance of the search results depends on it. The following studies report findings and ideas concerning content analysis, some of which are to be applied in the current project.

Saritha and Paul [81] proposed a content-based image retrieval framework based on the use of the deep belief network methods of deep learning, which are used to extract the features and classification. This algorithm provides high accuracy and good performance on a large volume of data. Meenakshi and Shaveta [64] proposed efficient content-based search using two approaches: text-based and feature vector-based ability. The research presented by Gao and Jin [26] explores text-image matching with the adaptive loss for cross-modal retrieval. The model splits images into patches combined with text tokens. This approach uses an adaptive loss

algorithm which automatically determines the loss weights. The paper presented by Ma and Gu [53] describes a large-scale image retrieval algorithm based on the sparse binary projection matrix through unsupervised training. The results showed an improvement in the performance for various pattern recognition tasks.

The work of Grycuk [29] introduces a novel framework for retrieving images. Their application is based on content-based information retrieval and was designed to retrieve similar images from a large set of indexed images to a query image. The first step relies on automatically detecting objects, finding salient features in the images, and indexing them with database mechanisms. The study conducted by Pawaskar and Chaudhari [70] proposed a web image re-ranking application that learns the semantic meaning of images with numerous query keywords. Portas and Nivaggioli [72] describe an open-source image similarity search engine. This solution allows users to make queries both textually or by using images, relying on similarity search.

Tian and Yang [87] sketch-based image retrieval is a task that learns semantic knowledge and embedding extraction to retrieve similar images using a sketch without any training examples of unseen classes.

Pegia and Jonsson [71] proposed a novel method for supporting multiple modalities in image retrieval. The method takes into consideration the semantic information of the training data through the use of Bayesian regression to estimate the semantic probabilities and statistical properties in the retrieval process.

Mu and Bai [63] presented the research, describing a novel multi-exposure image fusion method via boosting the hierarchical features.

The obtained results in the text image captioning area, presented by Tang and Hu [86], focused on purifying the OCR-oriented scene graph with the master object. The master object is the object to which the OCR is attached, the semantic relationship bridge between the OCR token and the image.

Findings for generating photo-realistic images from given text descriptions, explored by Dong and Wu [19], focused on Generative Adversarial Networks (GAN) for

text-to-image synthesis and provided more explicit category information and richer instance-level details.

Zhang and Xu [109] make the first attempt to realize data forgetting on deep models for image retrieval. The proposed solution provides many opportunities to use artificially generated data for better training for deep models.

The proposed framework uses it during the preprocessing phase to extract information from the image search archive and find relevant information during the multimodal conversational search process [72], [70], [81], [29], [72].

2.1.2 *Video search*

The increase in videos produced by various sources has resulted in online distributed video being provided on various video streaming services. As a result, the problem of effective search engines for videos has become more popular. There are a variety of possible approaches and methods for video information extraction, ranging from simple schemes using text queries only to sophisticated video concept detection methods and textual query analysis. Within this research project, it seems reasonable to implement a video-to-text preprocessing scheme for the video search dataset and use the text descriptions during the search process. The following papers include findings and approaches for video search, which are helpful for this research project.

The study of Choudhari and Bhalla [13] suggested a combined search approach which processes text search queries as feature vectors. The search feature vector searches for information in all potential text-only sources, such as titles, descriptions, comments or annotations.

Investigations performed by Rossetto, Giangreco and Tanase [79] amend a flexible retrieval model supporting multiple query modes for searching in multimedia collections. A framework capable of performing a wide range of search operations, such as using a user's sketches as an image search query, re-using a result as a query, or importing an image as a single video frame into the drawing area.

Markatopoulou and Galanopoulos [59] present a fully-automatic method that combines video concept detection and textual query analysis to solve the problem of ad-hoc video search. This method transforms concept-based keyframes and query representations into a common semantic embedding space. In the research implemented by Garcia [28], static pictures were used to find a specific timestamp or frame within a collection of videos. This approach processes the extracted visual features and performs the search among the video archive.

The framework presented in Zhao and Song [110] focused on the generation of brief text captures for video segments. This approach enhances the use of correspondence between visual and text content.

Wu and Ngyen [96] ad-hoc video search area studies describe a concept-based search solution. This approach relies on concept detectors to detect several concepts, such as a person, object, action, and place in the videos. Then, the detected concepts are indexes of videos to retrieve.

Yang and Lu [102] proposed a novel video interaction framework to automatically generate video montages by retrieving and assembling shots with arbitrary text scripts. The proposed model can generate video montages based on text-to-sequence retrieval and make them more consistent with the input text scripts. So, these findings will be helpful for the creation of various video search archives.

The solution suggested by Pan [69] included the scene-aware network to reduce semantic confusion in remote sensing cross-modal retrieval and enhance the visual representation.

Zacharian and Rao [104] conducted the research focused on video retrieval for everyday scenes with common objects. The system exploits the predictions made by deep neural networks for image-understanding tasks using natural language processing (NLP).

The solution proposed by the Bailer and Arnold [5] enhanced the quality of evaluating text-based queries in benchmarks for video retrieval systems. Also, they proposed a process for reviewing and revising the queries and preparing the assessors

and their findings for the proposed method to improve the clarity of queries and the consistency of judgements.

The multimedia retrieval framework developed by Chen [12] implements the practical text-based video retrieval paradigm. It aims at synchronously retrieving videos and specific video content from a large video collection with given text queries.

Lin and Lu [49] introduced a new video-based multimodal dialogue dataset for intelligent and human-like chatbots with multi-modal context. Also, the video-based multi-modal chitchat task was conducted, and several dialogue baselines were evaluated.

Research results obtained by Jiang and Zhou [34] addressed that video moment retrieval aims at retrieving the most relevant events from an untrimmed video with natural language queries. Their proposed method learns from point-level supervision where each annotation is a single frame randomly located within the target moment.

Li and Hsiao [47] presented the effective dual-encoder model to address the challenging video-text retrieval problem, which uses a highly efficient cross-attention module to facilitate the information exchange between multiple modalities (i.e., video and text). The proposed VideoCLIP was evaluated on two benchmark video-text datasets, MSRVTT and DiDeMo, and it demonstrated relatively high results.

Zhuo and Li [111] introduced an improved CLIP-based model, enhanced by two key innovations: a dynamic weighting strategy and a specially designed min-max hashing layer. These components were identified as the primary contributors to the model's performance improvements. When evaluated on three standard video-text benchmark datasets, their approach significantly outperformed existing state-of-the-art hashing algorithms.

Falcon and Lanz [23] proposed a framework which can organize the cross-similarity of video and text in a joint embedding space and put similar items close and dissimilar items far. It is necessary to note that work addresses text-video retrieval but can be easily extended to other domains where similar ranking losses are used, e.g. in image retrieval.

The video retrieval approach proposed by Dong and Chen [18] focused on the partially relevant video retrieval. An untrimmed video is considered to be partially relevant w.r.t. a given textual query if it contains a moment relevant to the query.

Ma and Ngo [54] prepared the novel interactive video corpus. Known-item video search is effective with human-in-the-loop to investigate the search result and refine the initial query interactively. Also, they have conducted user experiments for video corpus moment retrieval to localize moments from a large video corpus.

These findings and approaches [59], [102], [12], [34], [54] are used in the proposed conversational search framework to preprocess the video search archive and extract more information more effectively during the multimodal conversational search process .

2.2 *Recent works in Conversational Search*

Conversational search has become a topic of significant interest in the IR research community recently. The vast majority of the work has focused on text search, while a small number of studies have discussed its application in multimedia settings. In this section, we describe the concept of conversational search.

2.2.1 *Conversational search for text*

Conversational search is the process of interacting with a conversational system through natural conversations to search for information [106]. Conversational search presents opportunities to support users in their search activities to improve the effectiveness and efficiency of the information retrieval process. While most research in conversational search has primarily focused on text archives, implementing effective conversational search for multimedia content requires a solid understanding of the latest advancements, directions, and solutions in the field.

One important development is the use of mixed-initiative systems, where both the user and the conversational agent actively engage in a dialogue, exchanging

information throughout the search process [1]. This approach aligns with the theoretical framework proposed by Radlinski and Craswell [76], which emphasizes the importance of integrating mixed-initiative features into conversational systems.

Further, research by Aliannejadi and Zamani [2] highlights the critical role of accurately constructed clarification questions within information-seeking conversations, demonstrating how these questions can enhance the performance and effectiveness of conversational search systems.

Zamani [107] addressed the problem of auto-generating questions for a more effective search dialogue by using an encoding model with a bidirectional long short-term memory network. Zamani and Lueck [108] analyzed the user interactions area, highlighting the necessity for large-scale data collection for search clarification needs. They explored user interactions and manual annotations in the proposed datasets and shed light on different aspects of search clarification.

Aliannejadi [105] explored how different conversational search strategies and mixed-initiative approaches can be combined in simulated conversational search sessions in the context of text-based conversational search agents. To do so, they built upon existing interactive information retrieval models to develop a conversational search process model, which explicitly includes the core conversational concept of a mixed-initiative system and explored the impact of clarification properties on user engagement.

In Wadhwa and Zamani [90], the main focus was on applications and ways to model active engagement in conversational information seeking. They define a taxonomy upon which a framework for active engagement could be built. This is divided into three broad dimensions of an active engagement framework: initiation moment (when to initiate a conversation), initiation purpose (why to initiate a conversation), and interaction means (how to initiate a conversation).

The query expansion algorithm presented by Wang, Yang and Wei [92] generates pseudo-documents by few-shot prompting large language models (LLMs), and then expands the query with generated pseudodocuments. This approach is rather simple

yet effective and capable of improving the sparse and dense retrieval systems.

Jagerman [33] conducted the research focused on exploring query expansion by prompting LLMs. Conducted experiments demonstrated that query expansions generated by LLMs can be more powerful than traditional query expansion methods.

The paper by Mackie [56] provides the estimation algorithm focused on the more accurate weighting of expansion terms and making the query expansion with LLMs more precise. The obtained results show improvement in accuracy on several text datasets.

Chuang and Glass [14] described a query expansion and reranking approach for improving passage retrieval, with the application to open-domain question answering. This research first applies a query expansion model to generate diverse queries. Then, it uses a query reranker to select the ones that could lead to better retrieval results.

The relevance feedback approach was explored in Mackie [55] research and revealed that combining generative and pseudo-relevance feedback ranking to achieve the benefits of both feedback classes will significantly increase recall on the conducted experiments.

Vakulenko [88] research paper showed in an end-to-end evaluation that question rewriting is effective in extending standard question-answering approaches to a conversational environment. Obtained results set the new state-of-the-art on the TREC CAsT 2019 dataset. Based on the results of the implemented analysis, question rewriting is a challenging but promising task that can be effectively implemented into conversational question-answering approaches.

The aforementioned findings, concepts, and approaches have been incorporated into the design of our conversational search agent to enhance both the fluency of the dialogue and the overall effectiveness of the search process [1], [2], [107], [105], [33]. In particular, integrating mixed-initiative strategies and clarification techniques contributes to a more natural and responsive user interaction.

Furthermore, insights from research on query expansion have been instrumental

in shaping the query refinement functionality within our conversational dialogue system. By leveraging these techniques, the system can dynamically reformulate and extend user queries based on context. This combination of dialogue intelligence and adaptive query handling ensures a more user-friendly search experience.

2.2.2 *Conversational search for multimedia*

Conversational image search can interactively induce the user's responsibility to clarify their dialogue intent. Several efforts have been dedicated to the conversation part, namely automatically asking the right question at the right time for user preference elicitation. At the same time, few studies have focused on the image search part, given a multimodal conversational query [17], [25].

It is natural to use images as part of a query, in addition to the traditional text. Along with the rapid advancements in multimedia, natural language processing, information retrieval, and conversation technologies mean that it is time for us to explore multimodal conversation and its potential roles in search and recommendation.

Clarifying questions are one of the most studied forms of system initiative in conversational search, which aim to elucidate the user's information need [7]. Recent studies have highlighted the importance of clarifying questions in conversational search; generating them for open-domain search tasks still needs to be studied [2], [105]

Multimodal conversation can help us to uncover and digest a huge amount of multimedia information. Multimodal dialogue also enables natural 2-way interactions between humans and machines, with mutual benefits in enriching their respective knowledge as described in Magalhaes and Chua [57].

A multimodal conversational assistant that utilizes images as part of the search query was introduced by Kim and Yoon [42]. Their work focused on supporting various image editing tasks through a mixed-initiative conversational framework with natural language-formulated commands. The proposed system offered an intuitive

and interactive environment that streamlined the image editing process, making it both faster and more user-friendly for end users.

Nie and Jiao [65] presented a contextual image search approach. They suggested a conversational search interface using an image as a search query and implemented a fashion recommendation dialogue-based framework. Wu, Macdonald and Ounis [97] conducted research on a similar problem and explored user experience results for the multimodal recommendation system. Kaushik [36] introduced a basic multiview conversational image search system. This involved a multimedia search assistant that proactively puts out a fixed number of relevant questions to clarify the intention of the user using a reinforcement learning algorithm.

Yuan [103] introduced the research by exploring the various aspects of asking multimodal clarifying questions containing images in a dialogue-based conversational search. The results obtained during the user experiments proved the methods' high effectiveness.

An interactive video retrieval framework, based on conversational search methods presented by Khan [41] and demonstrated the user relevance feedback of the system, is used to refine a model to support a user's information need through content-based feedback. Moreover, conversational search is used to interactively refine or build upon queries to either directly solve an information need or to provide information to enhance the relevance feedback process.

The iterative sequence refinement for the conversational search presented by Wei[95] In real situations, users only provide ambiguous text queries, making it difficult to retrieve the desired images. To address this issue, the novel conversational composed retrieval method was presented. The provided models process complex user intent through iterative interaction. This paradigm enhances the model's capacity to learn various correspondences.

Wang [94] explored question generation for the conversational search. This research focused on the exploration of generating clarifying questions in a zero-shot setting to overcome the cold start problem. For this purpose, a clarifying ques-

tion generation system uses both question templates and query facets to guide the effective and precise question generation. The experiment results show that the suggested method outperforms existing state-of-the-art zero-shot baselines by a large margin.

Ferreira [24] presented the developed multimodal mixed-interactive framework. The proposed framework is capable of guiding users towards the successful completion of complex manual tasks using humanly shaped conversations and multimodal stimuli, including voice, images, and videos.

Interactive video retrieval system presented by Lyu [52], uses incorporating structured conversational information. Experiments conducted on the Audio Visual Scene-Aware Dialog dataset show that the proposed approach using plain-text queries improves over the previous counterpart mode.

Owoicho [68] explored the abilities of mixed-initiative systems to the users' feedback. This work focused on the exploration of the effectiveness of mixed-initiative conversational search models in combination with simulated user feedback. Proven models were enhanced by including user's answers to clarifying questions and explicit feedback on the system's responses.

Bao [6] designed the novel multimodal interactive framework with the multimodal prompts functionality. This approach uses a transformed text-only query into a multimodal prompt containing image tokens and text tokens. The contrastive learning with two types of losses is designed to learn a more consistent representation of two modalities (image and text) and reduce noise.

The mentioned findings, ideas, and approaches are used in the proposed conversational search in multimedia to make the dialogue flow more fluent and the search process more effective [57], [36], [103], [41], [52], [68]. It is also important to note the existing gap in the literature regarding conversational search for video systems. The current PhD research addresses this gap by exploring interactive video frameworks and extending them into a conversational search setting.

2.3 *Search interfaces*

Any conversational search framework requires dialogue between the user and the conversational search system. Such an interface should correspond to usability principles and standards [60]. In other words, the interface must be clear and convenient. Here we outline examples of different interactive conversational interfaces, visualization data interfaces and multimedia representation. These findings will be helpful and provide guidance for the design and building of a simple-to-use and convenient user interface for our conversational search framework.

The study conducted by McTear [61] explored how a conversational search interface is relevant today and identified some takeaways: the usage of a conversational search interface in messaging apps, personalised chat experience, and search interfaces which learn from previous experiences.

Moreover, Hearst [30] conducted a study in which the participants reported their preference for viewing visualizations in a chat-style interface when answering questions about comparisons and trends. This study's major insights revealed that most participants opted for additional visualizations and charts in addition to the regular textual replies in the chat interface. The results obtained demonstrate the impact of the graphical user interface elements incorporated into the conversational framework on the user experience.

Another research study was the conversational chat interface for stock analysis proposed by Lauren and Watta [46]. They used Slack as the platform for interaction and the RASA[1] framework for Natural Language Understanding and dialogue management. Also, they explored the conversational search interface abilities in applying for real-time stock analysis.

The study conducted by Kaushik [36] introduced a prototype multi-view search interface to a search engine API. The interface combines a conversational search assistant with an extended standard graphical search interface. The user interaction experience results demonstrate that the users found the conversational search

[1] `https://rasa.com/`

framework simple and convenient.

In their research, Doyle [20] explored various aspects of agent-user interactions and concluded that concepts of humanness are core to the design of speech interfaces. Yet, the specific dimensions of humanness used to define these interactions are not fully understood. The study conducted has clearly outlined key themes related to how users view humanness in dialogue interaction and how this varies in speech-based dialogue.

Laban [45] conducted broad research focused on exploring different aspects of intelligent assistants. The study highlights the universality of intelligence as a feature that is not limited to humans nor to non-human objects that appear human-like. Information systems that might not appear humane at all are still valued to be as intelligent based on their competence.

The research presented by Oh and Ju [66] explored people's impressions of different conversational search agent features. Based on these empirical findings, it is suggested that conversational agents should be designed with consideration of the different usage patterns and perceptions across age groups.

In the proposed paper, Doyle [21] developed a novel questionnaire for assessing user experience for conversational frameworks. This questionnaire evaluates the different aspects of perception, such as communicative competence and dependability, human likeness in communication, and communicative flexibility.

The modern search interfaces also include AR and VR interaction features. Spiess [85] conducted a comparative study to distinguish the differences between desktop and VR video browsing interfaces. The study's results demonstrated that VR interfaces can be competitive in browsing performance and indicate that there can even be an advantage when browsing larger result sets in VR.

The paper proposed by Xiao [98] a novel multimodal recommender framework to weaken the redundancy between heterogeneous modalities. Moreover, they designed the gating mechanism to set unequal weights to different modalities. To increase confidence, many experiments were conducted, and the results show that the pro-

posed model has achieved better performance than the state-of-the-art methods on both public and collected industrial datasets.

Rapp [78] offered a detailed recount of how people collaborate with a task-focused chatbot. They identified two main aspects of collaboration, behavioural and conversational, and for each aspect, highlighted the different strategies that users utilise to "work together" with the agent. The strategies identified span from user commitment to acceptance of the chatbot's proposals and their willingness to behave favourably towards the chatbot.

The research presented by Pucci [75] has discussed a new paradigm for conversational Web browsing, as emerged from a human-centred process conducted with a sample of different groups of users. The illustrated results aim to fill the current gap in the literature with concrete guidance on how to design conversational agents for the Web.

Rajaram [77] demonstrated the novel approach extending user-driven elicitation to design techniques for sharing AR content. Also, they explored how to adapt a similar elicitation method to teach designers about considerations for AR in future work.

The study conducted by Xiao [99] investigated the impact of an AI-powered chatbot on enhancing the process of obtaining informed consent online. The results of the user study demonstrated that the chatbot not only improved participants' engagement with consent forms but also fostered a more balanced power dynamic between participants and researchers. Furthermore, it led to higher-quality responses within the study. As dialogue-based applications continue to grow in popularity, these findings offer valuable design insights for developing more effective and user-centred chatbot systems.

The work proposed by Cao [11] explores the potential and challenges of virtual exhibitions through a series of interviews and user surveys. Insights gathered from these user studies were instrumental in identifying both the strengths and limitations of current virtual exhibition practices. Based on this feedback, the study offers a set

of practical guidelines aimed at improving the design of future virtual exhibitions and enhancing the usability of VR interfaces. These findings contribute to the growing body of knowledge on user experience in immersive digital environments.

User models play an important role in interaction design, supporting the automation of interaction design choices. In order to do so, Kerulainen et al. [40] presented a novel approach, which is reducing the computational cost of designing experiments by training a policy for choosing experimental designs with simulated participants. The designed solution learns which experiments provide the most useful data for parameter estimation by interacting with various types of users.

There have been significant advances in simulation models predicting human behaviour across various interactive tasks. One issue explored by Moon [62] demonstrates that an amortised inference approach permits analysing large-scale datasets by means of simulation models. It also addresses emerging opportunities and challenges in applying amortised inference in human-computer interaction (HCI).

Kim and Son [84] presented the framework using an interaction method for predicting a user's intended target based on a user's input. Furthermore, user study insights confirm that the computational cost is significantly reduced compared with the existing framework, and it plausibly detects the point where the user changes goals.

Kim and Lee [43] proposed a technique to quantify reactiveness and proactiveness to determine the degree and characteristics of each input strategy. The technique explored in two empirical studies highlighited how to use the technique to answer questions proactively or reactively.

Saquib [80] designed the solution at the intersection of computer vision and graph analytics by utilizing visual variables extracted from images/videos and some direct manipulation and pen interaction techniques. The design framework is implemented as a sketch-based notebook interface to demonstrate the design possibilities. User studies with scientists from various fields reveal innovative use cases for such an embodied interaction paradigm for graph analytics.

The findings, ideas, and approaches mentioned are used in the designed conversational search web interface, including the search dialogue and multimedia gallery for representing search results. Also, the experiment design and search tasks, which will be discussed in Chapters 5 and 6, are based on the described sources [46], [36], [66], [98], [78], [77].

The next chapter describes the research project hypotheses and discusses the research questions in detail.

Chapter 3

Methodology

3.1 Overview

This chapter presents the experimental design practices employed in our conversational information retrieval investigations for both image and video search. It provides a detailed account of the experimental setup, outlining the methodologies and tools used to conduct the study.

Evaluation plays a pivotal role in multimedia information retrieval (MIR) research, particularly in assessing how effectively a system meets users' information needs. In the context of conversational search, evaluation focuses on two key aspects: the user experience and the effectiveness of the queries generated during the conversational search process. These assessments provide valuable insights into the system's ability to deliver relevant results and facilitate engaging, efficient interactions.

3.1.1 Experimental Methodology

Our methodology emphasizes practical interactions within information retrieval (IR) evaluation methods, aiming to provide a comprehensive understanding of how users engage with the system during the search process. It incorporates real-world user scenarios to assess not only the effectiveness of retrieval outcomes, but also the

dynamics of human-computer interactions. This approach evaluates the user's behaviour, decision-making processes, and the overall usability of the system, offering valuable insights into the interplay between the user and the technology.

Designing experiments to investigate user search behaviour is a complex and multifaceted process. As emphasized by Kelly [37], user search behaviour is influenced by numerous factors, including mood, pre-existing knowledge, and interest in the search topic. Studying the impact of individual factors within an interactive search process poses significant challenges, since these elements often interact dynamically, making it difficult to isolate their specific effects [51]. This makes it very challenging to design an experimental setup which allows multiple users to have the same feel or experience while using it. Another complex task is to understand the relationship between these factors. However, due to the increased interest in human-computer interaction and interactive information retrieval (IIR) communities, multiple studies have appeared [16], [37], [38], [39], [44] that focus on developing standard practice for the design and evaluation of IIR systems. During this PhD research, these studies are used as a source of methods for the design of experiments for our investigations. In the next section, we review some of these studies, highlighting features important to the design of experiments and investigations. In this section, we discuss the topic of remote studies vs lab-based studies, the effect of limited time duration in experimental setups, and the adoption of task sequencing strategies for allocating tasks to the user to avoid biasing effects.

Remote studies

A study conducted by Kelly and Gyllstrom [38] compared lab-based vs remote-based IIR experiments. The investigation was conducted with two groups of people: the first group participated in the experiment remotely and the second group participated in the laboratory. Both groups were studied on the basis of user behaviour, search behaviour and evaluation behaviour. For most of the measures, there were no significant differences between the settings. This demonstrated that user behaviour

does not change significantly based on the experimental search setting. Following these findings, the experiments conducted in this PhD research were performed remotely to make attendance easier for non-DCU and non-Dublin-based participants. The experimental setup and design of each investigation are described in detail in the relevant chapter.

Sequencing effects

It is crucial to eliminate sequence or order effects in an experiment to ensure that results are not influenced by the order in which tasks are presented. Sequence effects can increase the likelihood that outcomes are attributed to the experimental conditions rather than genuine differences in user behavior across tasks. To minimize bias in the experimental setup, the search tasks are systematically rotated and counterbalanced. Studies such as [37] and [27] describe using the Latin square method to arrange the search tasks to avoid order effects. Based on these methods, the experiments conducted in this PhD research all arrange the search tasks using Latin square sequencing methods. In Chapter 4, Chapter 5 and Chapter 6, each user had to perform search tasks using conversational and conventional frameworks.

3.1.2 Experimental Procedure

In this section we describe the details of the experimental setups for our user studies. The studies aim to enable us to observe and better understand the behaviour of non-specialist searchers whose techniques for use of search engines are generally learned from personal experience. To address the raised research questions, introduced in Chapter 1, the main goal of the experiments is to compare the user experience for conversational and conventional search frameworks during the multimedia search process. We seek to gain insights into how conversational engagement might be helpful to make the multimedia information retrieval process easier using conversational assistance and seeking opportunities to enhance the user's search experience.

The studies aim not just to observe user behaviour in completing their search

tasks, but also to gain their feedback about the approach. In doing this, we hope to gain insights into the relationship between user actions and information seeking in order to be able to make use of this in the design of future conversational search agents.

Participants in our studies are required to complete a search session consisting of two search tasks per framework. As part of their search session they complete a questionnaire before and after undertaking each task. In this section, we first give details of the standard structure of experiment, and then describe our experimental setup and questionnaires, and follow this with the procedures used for our studies (ethical permission, recruitment, pilot studies and thematic studies). The details of each step mentioned above are discussed in the individual next subsections.

Search Tasks

For our search tasks, we aim to provide participants with realistic information needs that could be addressed using a standard image or video search engine. The process of creating these search tasks, tailored to align with the objectives of the experiments, is detailed in the Chapters 4-6, outlining the design of each experiment.

Questionnaire

As part of completing each task the searcher is required to complete an online questionnaire. The detailed questionnaires can be found in Appendix 1.

Experimental Setup

Participants engage in a pre-search survey, then conduct the assigned search tasks using a conventional search application and a conversational agent framework, and finally, complete a post-search survey to provide feedback or compare their experiences with the two frameworks.

The entire study was conducted online, with the search frameworks deployed on Amazon Cloud infrastructure. Separate versions were set up to support both the

conventional and conversational search approaches.

Recruitment

Participants for the experimental studies were recruited following the approval of ethical clearance. Recruitment was carried out through multiple channels, including targeted emails to university groups, posts on social media platforms, and promotional flyers and posters distributed at conferences.

Participation in all experiment was voluntary. Participants were informed of the purpose of the experiment and their role in a Participant Information sheet prior to beginning work, and it was made clear to them that they could withdraw from the study at any time if they were willing to. Details of the participants for each study are given in the relevant part of each Chapter.

The following sections describe the design of the experimental studies and the prototypes of experimental search frameworks implemented for these studies.

Pilot Studies

A pilot study involving a small group of PhD candidates was conducted prior to the main study for each experiment. This pilot study utilized the provided search instructions to evaluate the time required to complete different sections, gain insights into the anticipated behavior of participants, and identify and resolve any issues in the experimental setup.

Participants were given details of the instructions for their search sessions and each search task shared in the Google Docs, and an interactive tutorial before the performing their assigned search tasks. Each pilot search task, whether for image or video search, took approximately 30 minutes to complete. The feedback collected during the pilot study was instrumental in refining the interface and incorporating new recommendation features to enhance the user experience.

Ethical Permission

Approval was obtained from the DCU Research Ethics Committee prior to beginning the user studies conducted during this PhD.

3.2 Experimental Design in Conversational Search

In this section, we provide a comprehensive overview of the design and features of the two experimental frameworks: conventional and conversational. Both frameworks are developed to support image and video search, with similarities in functionality within each mode. However, the conversational framework incorporates additional features for multimodal interactions. While the general functionality of the conversational framework remains consistent across image and video search modes, the implementation of multimodal clarification questions differs technically between the two. These technical distinctions are elaborated upon in the subsequent sections.

The conventional framework offers a straightforward and uniform functionality for both image and video search modes, focusing on traditional input-output search mechanisms. In contrast, the conversational framework includes augmented reality (AR) capabilities, but this feature is implemented exclusively for image search. The AR functionality enhances user engagement by overlaying visual labels of detected objects on images, providing an intuitive way to filter and refine search results interactively.

After detailing the design and features of the interfaces, we describe the process of user interaction and engagement with the system. This includes how users navigate the search process, respond to multimodal clarification prompts, and utilize AR functionality. Additionally, the technical details of the system architecture, highlighting the integration of conversational elements, multimodal capabilities, and backend support for efficient search processing were presented in the section below.

3.3 Prototype Systems

3.3.1 Conversational Search Interface

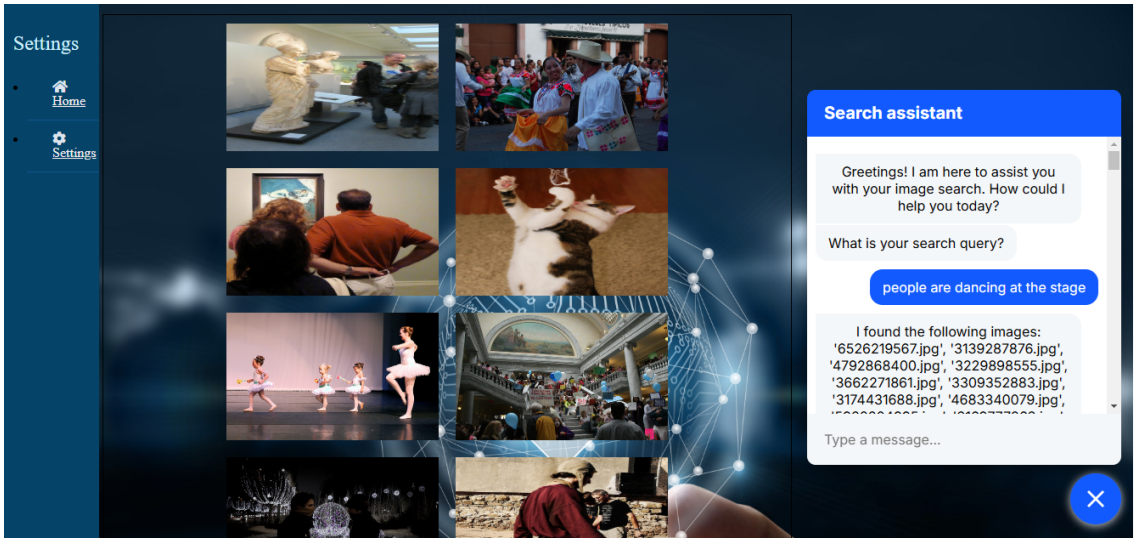

Figure 3.1: The interface of conversational search framework

Figure 3.1 shows the search interface for our prototype conversational multimedia search system.

This interface includes the following components:

1. Agent Display: Provides users with the ability to communicate with the search system using natural language instructions or predefined buttons for quick actions.

2. Image and video gallery: Presents the search results to the user, showcasing images or video frames retrieved based on the search query.

3. Additional question: Enables users to determine the next step in their interaction with the search results. Options include:

    - Restart Search: Clears the gallery and resets the search session, allowing users to begin a new search scenario.
    - Reduce Search Output: Suggests additional images from the Flickr30k dataset to help refine the search output by reducing irrelevant results.

These images are selected based on their proximity to the user's text query.

- Refine Text Query: Allows users to expand or modify their text search query to obtain a broader or more targeted search output.
- Show Visual Labels: Displays detected objects overlaid on the images retrieved, helping users identify relevant results more effectively.

4. Select mode menu: Allows the user to choose the search option:

- Image: Starts the search process for image search archive
- Video: Starts the search process for video search archive

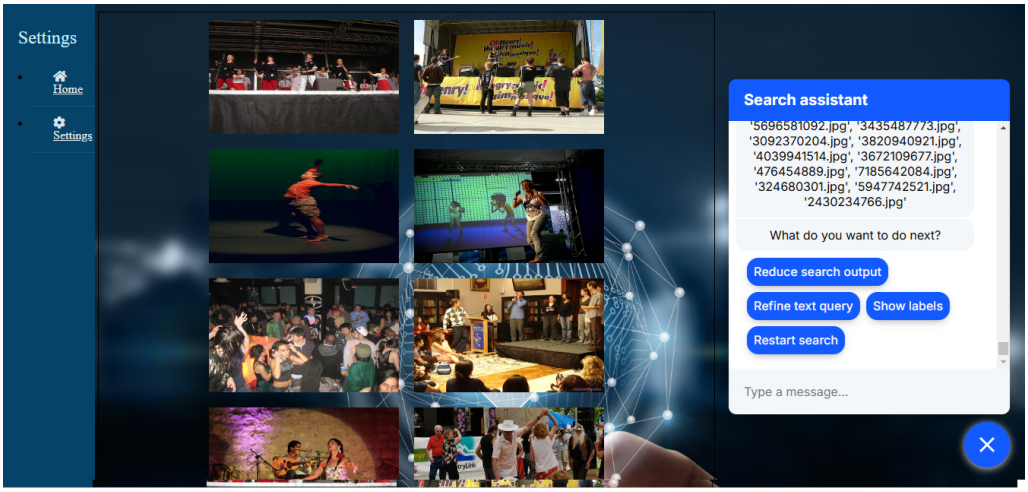

Figure 3.2: Process of clarifying of the image search

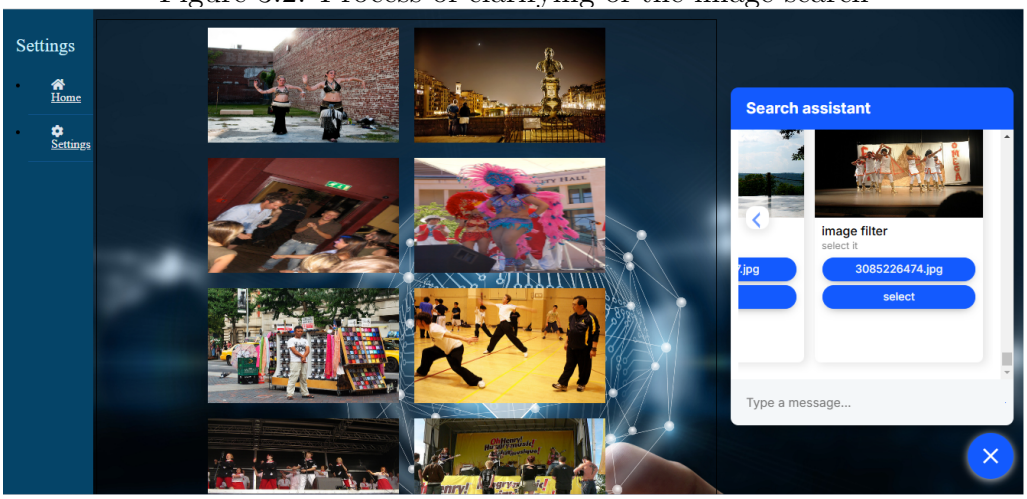

Figure 3.3: The suggested images for the search output refinement

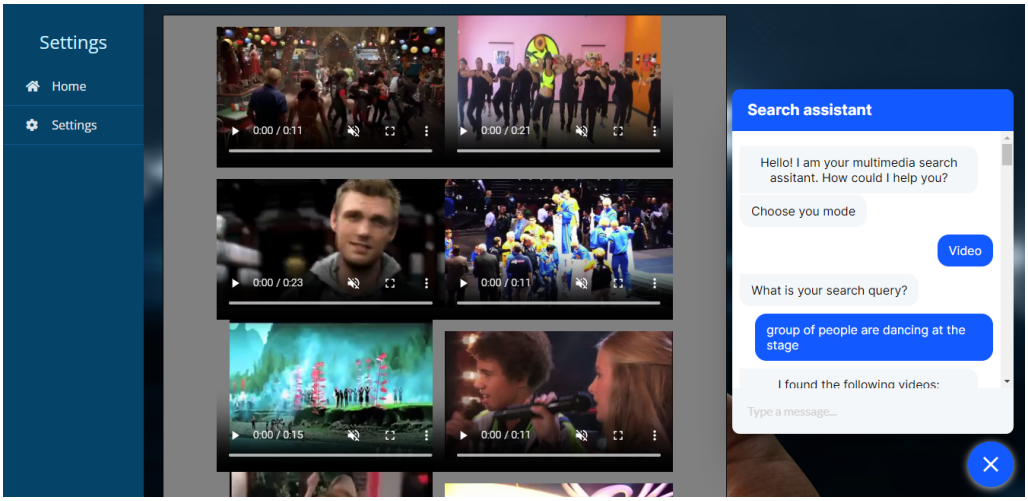

Figure 3.4: Initial video search outcome

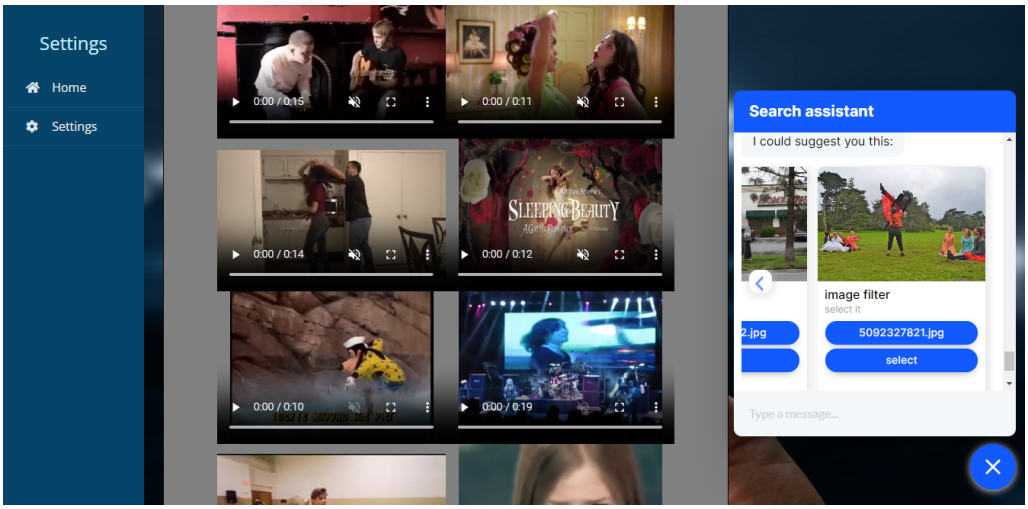

Figure 3.5: Clarification process for the video search

Figure 3.2 illustrates the potential actions of the conversational agent, focusing on the process of refining search outputs. In this example, the user selects the 'Reduce Output' option, prompting the framework to suggest two images as potential filters, as it shown on Figure 3.3. When the user selects one of the suggested images, the framework calculates the L2 distance between the images in the gallery and the selected image. Images with distances below a certain threshold remain in the gallery, effectively narrowing the search results.

For the conversational search in video mode, the system employs a similar approach but tailored for video content. It begins by posing clarifying questions based

on the user's initial query, as shown in Figure 3.4. The user's responses guide the search process and help refine the results.

The video search functionality incorporates a comparable algorithm to that used for image search. However, instead of comparing images in the gallery, the selected image in the search dialogue is compared to video frames. The framework calculates the L2 distance between the selected image and video frames, retaining only those frames that are sufficiently similar. This process, demonstrated in Figure 3.5, reduces the output dynamically, enabling more precise and relevant search results for video content.

3.3.2 Dialogue Flow

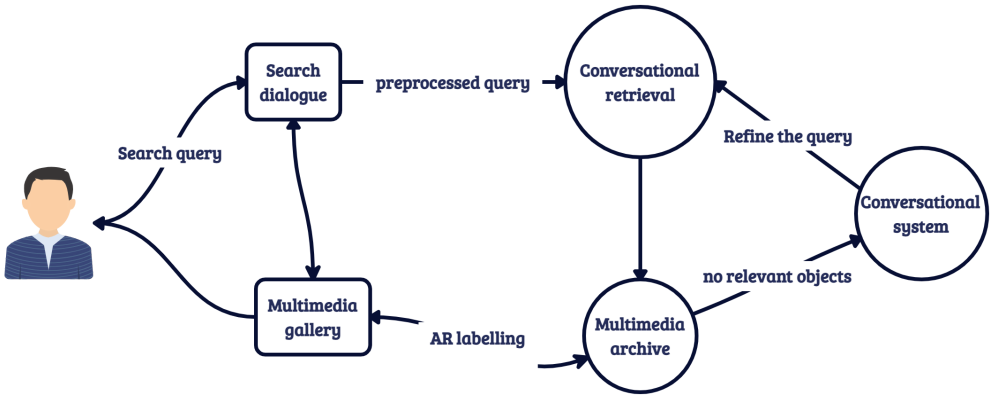

Figure 3.6: Representation of conversational MIR workflow.

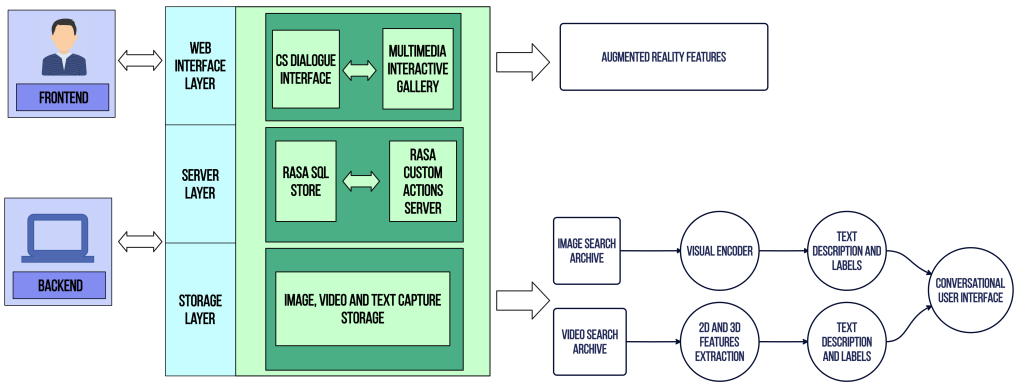

Figure 3.7: Representation of framework structure.

The user engages with the search system through a conversational dialogue interface. Our prototype includes a Natural Language Understanding (NLU) model that supports and facilitates the search interaction. First, the user's initial text query is preprocessed and then compared against text descriptions of the images or videos from the multimedia databases in a retrieval process. The detailed description of archives structure are described below in the "System implementation" section. If the results of this search are not found to be relevant by the user, the displayed result can be revised. Figure 3.6 shows a simple representation of the workflow of our conversational MIR system.

In the dialogue, the conversational agent proactively seeks clarification when the user indicates that the current results are irrelevant. If the initial search output contains some relevant results, the user can refine the results further by reducing the search output through the dialogue, as it shown on the Figures 3.2 and 3.3.

This interactive process progresses within the search dialogue, enabling users to iteratively narrow down their results. The updated search results are displayed in the image or video gallery linked to the search agent, ensuring a seamless and intuitive search experience.

3.3.3 System implementation

Figure 3.7 shows the architecture of our prototype conversational multimedia information retrieval (MIR) system. This is implemented as a web application developed using the Python and the Flask framework. It is deployed on the Amazon virtual machine. The architecture is composed of three layers:

Storage layer: This includes the preprocessed search datasets, detailed in the datasets subsection below. These datasets have been preprocessed to extract visual features and objects suitable for performing content-based MIR. Each image and video is accompanied by text descriptions that capture their content [35]. Additionally, the image archive includes supplementary content with detected objects and corresponding text labels [89]. These labels are utilized to enhance the gallery

display, supporting the augmented labels functionality. This layer is not visible to the user.

Server layer: This is responsible for performing the search process. The search framework includes the dialogue system that is responsible for managing the conversation and search process, introduced in the previous section. The dialogue system uses a RASA[1] API, which is a popular open-source framework for building conversational AI systems. The Whoosh[2] library is used to understand and parse the user text query and to support the search process. This layer is also not visible to the user.

Interface layer: This layer represents the web interface of the framework, which serves as the primary point of interaction for users. The interface comprises two main components:

- Agent Display: This window allows users to input their search queries and interact with the system using natural language dialogue. It facilitates seamless communication between the user and the search framework.

- Multimedia Gallery: Displays the search results, including images or video frames, dynamically updated based on user interactions and search refinements.

The search framework integrates a dialogue system that manages the conversation and oversees the entire search process. Users have access only to this interface layer, making it the central hub for all interactions with the system.

[1] `https://rasa.com/`
[2] `https://whoosh.readthedocs.io/`

3.3.4 Conventional framework

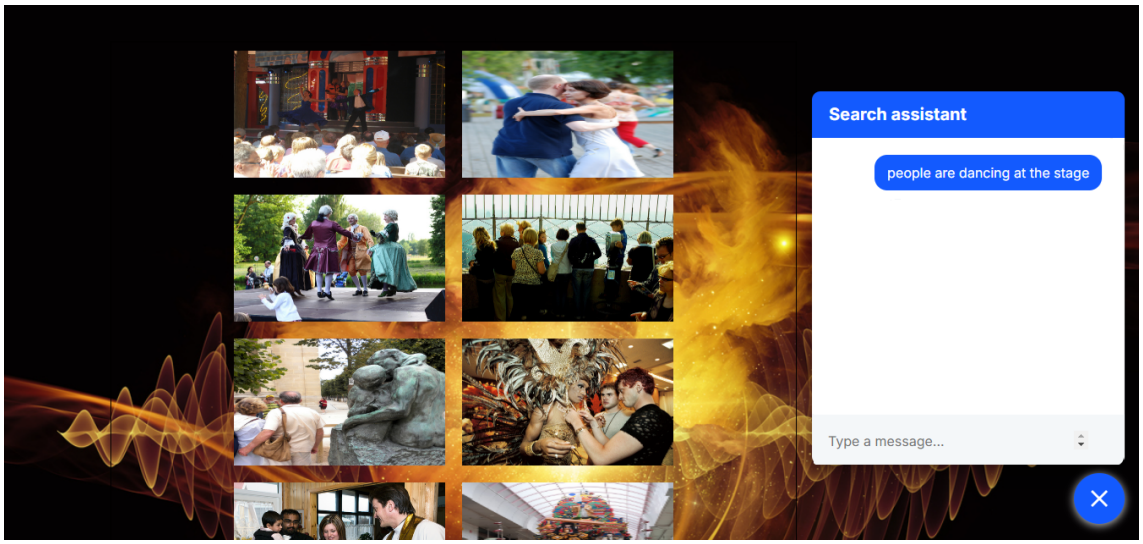

Figure 3.8: The interface of traditional search system

An equivalent conventional image search framework was developed using the same technologies and adopts the same image search archive. Figure 3.8 illustrates the interaction search process for the traditional framework. The search results are shown in the multimedia gallery connected to the search window on the right. The framework supports only text search functionality, so participants must manually scroll the search output or refine the query to obtain the desired image or video. Other functionality, such as clarification questions or augmented labels are not supported in this search framework.

3.3.5 Datasets for Experimental Investigations

The following publicly available datasets are indexed within our prototype system to support the investigation: Flickr30k and MSR-VTT. Both datasets were pre-processed to extract visual features and detected objects suitable for performing content-based MIR [89], [101]. These datasets were chosen for their richness in annotations and their relevance to both image and video retrieval tasks, providing a solid foundation for developing.

Flickr30K The Flickr30k dataset is a comprehensive image dataset that includes

approximately 31,000 images. Additionally, the dataset is enriched with 276,000 manually annotated bounding boxes, associating specific entities in the captions with their corresponding regions in the images [31]. This dataset was selected for our study due to its wide variety of general-domain content, including images of people, objects, and places. The extensive captions and annotations make it particularly well-suited for investigating conversational image search, as they provide rich textual information for query construction and result refinement.

MSR-VTT (Microsoft Research Video to Text) The MSR-VTT dataset complements the image-focused Flickr30k by providing video data suitable for conversational search in a multimedia context. This dataset, which contains 10k short (9-15 seconds) video clips paired with detailed captions [93] [100] [91], enables the evaluation of search and retrieval methods specific to dynamic content. Although further details on MSR-VTT are described elsewhere, it plays a critical role in our system by supporting the exploration of conversational video search scenarios. By indexing these datasets, our system enables the investigation of both static (images) and dynamic (videos) content retrieval within a conversational search framework, leveraging their rich annotations and diverse content to simulate real-world search interactions.

3.3.6 Search scenarios

In this section, we outline typical search scenarios for both image and video retrieval that are expected to align well with the capabilities of our conversational search system. These scenarios demonstrate how the system facilitates effective user interactions and showcases its suitability for addressing diverse multimedia search needs.

Image search scenario

Our prototype MIR system enables users to perform a search for images and videos. A common search scenario for a conversational search for images could include the

following steps:

1. **The user enters a text query:** The user initiates a search by entering a text query to the conversational search framework.

2. **Framework obtains results:** The conversational search framework processes the user's query and retrieves potentially relevant results from the preprocessed archive. Results are shown in the multimedia gallery.

3. **User selects the suggested option:** User may select one of the suggested options in the dialogue, which were described in the previous sections.

4. **User selects a filter image:** The framework presents the user with the initial set of results. The user can then select an image in the search dialogue to refine the search. For instance, they might choose a specific image representing their preferred style or composition.

5. **Framework reduces the output:** Based on the user's selected filter image, the conversational search framework applies a filtering mechanism to narrow down the output. The framework uses visual similarity to reduce the set of images to those closely matching the selected filter image.

6. **Reformulate query with the LLM (GPT-4) [67]:** If the user opts to expand their query, the system can perform this expansion automatically. The user's initial request is reformulated by the GPT-4 model to generate a more detailed and comprehensive search query. The system then executes the search again using the enhanced query, resulting in a broader and potentially more inclusive set of search results.

7. **Show augmented labels:** The user can choose to display images with overlaid detected objects and further refine the search output by typing one of the detected labels displayed on the images. This allows the user to filter results more effectively based on the visual content of the detected objects.

8. **Repeat the process:** If the user is unsatisfied with the refined results or wants to explore further, they can repeat the process by sending a new text query, selecting a different filter image, and obtaining a new set of reduced output. This iterative approach allows the user to narrow their search and gradually find the desired images.

Video search scenario

A common search scenario of a video collection could include the following steps:

1. **The user enters a text query:** The user initiates a search by entering a text query to the conversational search framework.

2. **Framework obtains results:** The conversational search framework processes the user's query and retrieves relevant results from the preprocessed video archive. Results are shown in the multimedia gallery.

3. **User selects the suggested option:** User may select one of the suggested options in the dialogue, which were described in the previous sections.

4. **User selects the filter image:** The user can then select a suggested image from the search dialogue to refine the search. For instance, they might choose a specific image representing their preferred style or composition.

5. **Framework reduces the output:** Based on the user's selected image, the conversational search framework calculates the L2 distance between selected image and video frames to narrow down the output.

6. **Reformulate query with the LLM (GPT-4) [67]:** If the user opts to expand their query, the system can perform this expansion automatically. The user's request will be automatically reformulated by the GPT-4 model, and the search will be performed again with the more detailed search request, which provides a broader search output.

7. **Repeat the process:** If the user is unsatisfied with the refined results or wants to explore further, they can repeat the process by sending a new text query, selecting a different filter image, and obtaining a new set of reduced output. This approach allows the user to narrow their search and find desired videos gradually.

Following this scenario, the conversational search framework enables users to engage in a conversational and interactive search process. They can provide queries, select filter images, and refine the output iteratively until they find the most suitable videos based on their preferences and criteria.

3.4 Evaluation methodology

3.4.1 Design of user studies

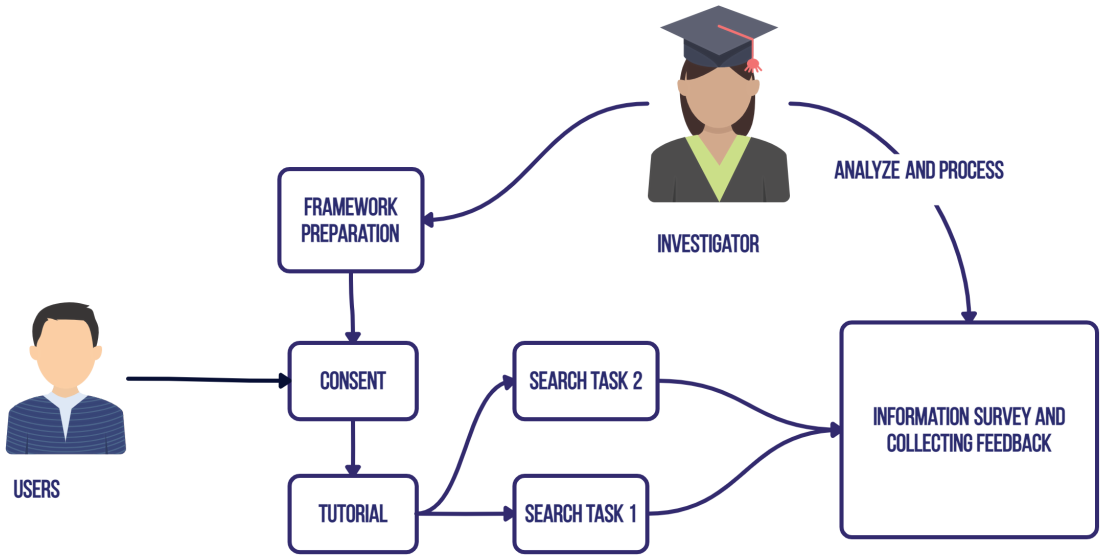

Figure 3.9: Proposed scheme for the design of user experiments

Figure 3.9 shows the proposed procedure adopted for the user studies conducted within this PhD. The experiments aim to enable us to observe and gain a better understanding of the behaviour of users. The objective of these studies is to explore the effect of previously described conversational dialogue features on existing MIR search approaches and the user search experience.

The experimental studies examine the following topics:

- **Comparison of the conversational search interface with a traditional MIR search system.** This evaluation enables the comparison of traditional and conversational search interfaces. The user completes several search tasks, one each using each search setting (traditional and conversational). For each study, the user completes a pre-search questionnaire and a post-search questionnaire. This experiment is focused on providing better insights into the operation of a CS system and contrasting user feedback of each type of interface.

- **Evaluating only a conversational interface based on selected performance metrics.** Pre-search and post-search questionnaire scores could be compared using standard benchmarks [22], [8]. This provides an opportunity to explore the conversational search interface with a standard system benchmark. Moreover, this investigation allows us to understand user expectations and can be useful in understanding how far the current CS interface is from the user's expectations.

- **Comparison of the Conversational Search Interface With and Without Augmented Labeling:** This evaluation facilitates a comparative analysis of the functionalities within the conversational search framework, focusing specifically on the impact of augmented labeling [77]. The structure of the study follows a similar approach to the previously described experiment, ensuring consistency in methodology. This experiment aims to provide deeper insights into the functionality of augmented labeling and its influence on user experience in MIR.

Experiments in CS to date have focused on user feedback and various aspects of interaction with the system, and changes in the user's knowledge [1].

3.4.2 Evaluation tools for the user experiments

There are numerous metrics available for evaluating the usability of interactive systems [8], each designed to measure different aspects of user interaction [82], [4], [15] system performance [44], [16] and overall user satisfaction [62], [40]. These metrics are critical for understanding how effectively a system meets user needs, facilitates task completion, and enhances the user experience.

For our experiments, we propose to utilize the following metrics, which are tailored to assess the usability and performance of our conversational search framework in the context of MIR:

1. **Chatbot Usability Questionnaire**

   The Chatbot Usability Questionnaire (CUQ) is derived from the chatbot UX principles outlined by the ALMA Chatbot Test tool. This tool evaluates key aspects of chatbot design, including personality, onboarding, navigation, comprehension, response quality, error handling, and intelligence [32]. The CUQ consists of 16 items tailored specifically for assessing chatbot usability while maintaining comparability with broader usability metrics. These scores highlight that the conversational framework is user-friendly, convenient, and straightforward to navigate, making it accessible to a wide range of users.

2. **User Experience Questionnaire**

   The User Experience Questionnaire (UEQ)[22] is a fast and reliable questionnaire to measure the user experience of interactive products. By default, the UEQ does not generate a single score for each participant but instead provides six scores, one for each attribute [22]. Attributes score the UI on six qualities: Attractiveness(overall characteristics), Perspicuity, Efficiency, Dependability (pragmatic qualities), Stimulation, and Novelty (hedonic qualities). The scores given by the users are on a scale of 1 to 7. This metric will present the overall user experience of using the conversational search framework. The User Experience Questionnaire contains 6 scales with 26 items:

- Attractiveness: Overall impression of the product. Do users like or dislike the product?
- Perspicuity: Is it easy to get familiar with the product? Is it easy to learn how to use the product?
- Efficiency: Can users solve their tasks without unnecessary effort?
- Dependability: Does the user feel in control of the interaction?
- Stimulation: Is it exciting and motivating to use the product?
- Novelty: Is the product innovative and creative? Does the product catch the interest of users?

Attractiveness is a pure valence dimension. Perspicuity, Efficiency and Dependability are pragmatic quality aspects (goal-directed), while Stimulation and Novelty are hedonic quality aspects (not goal-directed).

Also, we will compare our obtained data with the data from UEQ benchmark [83]. A benchmark for the User Experience Questionnaire (UEQ), a widely used tool for assessing the usability and user experience of interactive systems. The benchmark also serves as a valuable reference for supporting quality assurance processes in individual projects.

The benchmark was established by aggregating data from numerous UEQ evaluation studies contributed by researchers and industry professionals. While some data came from scientific research, the majority originated from industry projects. Currently, the benchmark includes data from 246 product evaluations, representing a diverse range of applications, and contains a total of 9,905 individual responses.

The number of respondents per evaluation varied significantly, ranging from as few as 3 to as many as 1,390 participants, with an average of 40.26 respondents per study. This comprehensive benchmark provides valuable insights and a reference point for comparing and interpreting UEQ results across different contexts and applications.

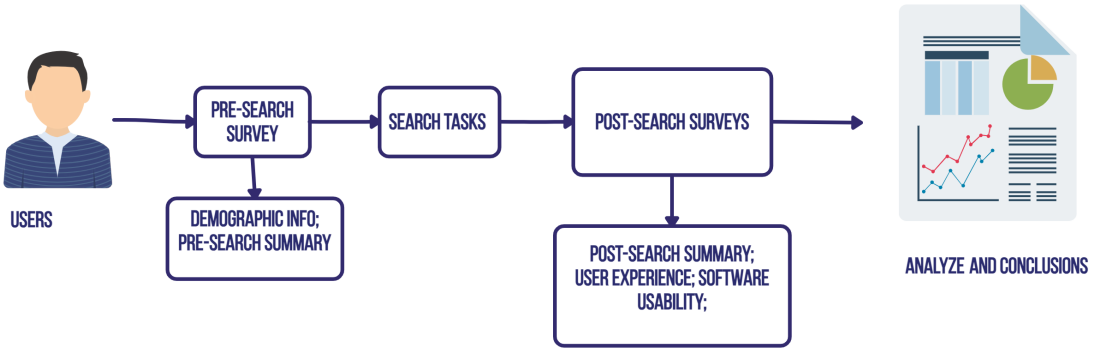

Figure 3.10: Evaluation process including pre-search and post-search surveys

The overall evaluation process is illustrated in Figure 3.10. Participants begin by completing a pre-search online survey to provide baseline information, followed by a post-search survey to capture their experiences and feedback during the search process. After the user experiments are concluded, the collected data will be analyzed to gain insights from the study. This analysis will help evaluate the multimodal conversational search framework against standard usability metrics and address key exploratory research questions (RQs), providing valuable information for further system improvements and research.

3.5 Concluding remarks

Conversational search and its impacts remain an evolving area of research. A critical component of this PhD study is the preparation and development of suitable methodologies and processes for evaluating conversational MIR.

In this chapter, we present an experimental methodology and an evaluation framework designed to assess multiple dimensions of the effectiveness of the interface. This framework provides a structured approach to understanding the usability and performance of conversational search systems.

In the subsequent chapters, we apply the designed interface to explore user experiences in conversational search (CS) for both image and video retrieval tasks, highlighting the system's capabilities and user-centric enhancements.

Chapter 4

A Comparative Study of Conversational and Conventional Search Methods for Image Retrieval

4.1 Overview

The incorporation of conversational engagement in search systems presents opportunities to support users in their search activities, improving the user experience in completing search tasks. In this study we examine the integration of conversational search in MIR for an image retrieval task. Our conversational MIR system seeks to improve user identification of potentially relevant information by using a conversational agent search assist to engage with the user while carrying out their search.

The user can engage directly with the search system while receiving suggestions from the search assistant to help them refine their queries and guide their interactions with the retrieved content. In this investigation, we conduct a user study focused on the comparison of user experience when using our conversational interface and

an identified search system using conventional interface. Our study seeks to address the following research questions:

RQ1: User experience in conversational MIR: How does user experience in MIR for image retrieval compare between a standard MIR system and an equivalent one integrating a conversational search agent? This RQ focuses on exploring different aspects of the user experience using effective MIR systems. This research question contains two subquestions:

1. How can multimodal conversational search system be compared with a conventional search system?

2. What aspects of using multimodal features in a search dialogue can be used effectively in the image search process?

RQ2: How clarifying questions could be used effectively to resolve ambiguity and improve search effectiveness in conversational MIR?

Clarifying questions are one of the most studied forms of system initiative in conversational search, which aim to elucidate the user's information need [7]. Recent studies have highlighted the importance of clarifying questions in conversational search; generating them for open-domain search tasks still needs to be studied [2], [105]. RQ2 includes the following subquestions:

1. What are the opportunities and challenges for embedding clarifying questions into the conversational MIR framework?

2. Can multimodal clarification features advantage affect the user's search result preference and the user's perceived workload?

4.2 Experimental methodology

This study aims to compare conversational and traditional approaches for image search scenarios, and make suggestions for improvements to the conversational search interface. In this section, we describe the details of our user study exploring

our prototype conversational image search application and contrasting this with an equivalent conventional image search system without conversational support. This aims, in particular, to enable us to examine and better understand the search behaviour of non-specialist users whose techniques for use of search engines are learned from personal experience, rather any type of formal training [3]. We first describe the experimental framework developed for the image applications, and then the methodology and procedures used for the experimental study. The study was performed online, and participants using two systems deployed in the Amazon Cloud.

4.2.1 Experimental framework

In this section, we provide details about a prototype conversational image search application and compare it to a conventional image search system that lacks conversational support.

Dialogue Flow

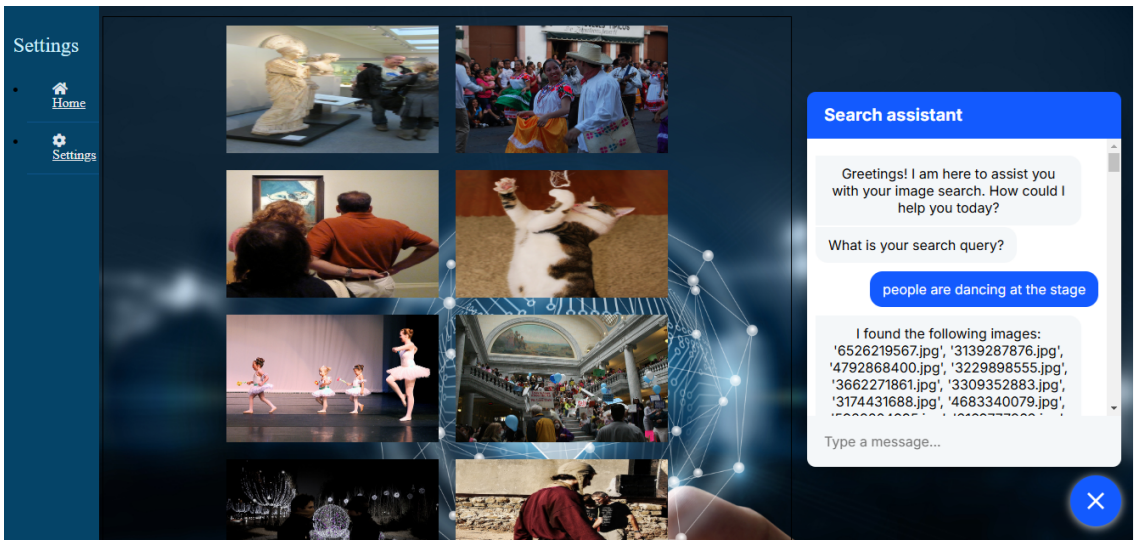

Figure 4.1: The interface of conversational search system

The dialogue module of the conversational application includes a natural language understanding (NLU) model, which supports the search dialogue. The dialogue module of the conversational application incorporates a natural language under-

standing (NLU) Dual Intent and Entity Transformer model [9] to support the search dialogue. In this module, the user's initial text query is preprocessed through techniques such as lemmatization and the removal of articles. The preprocessed query is then compared against the textual descriptions of images stored in the multimedia image database. The retrieval process involves indexing the data (before application deployment) and performing a search within the indexed archive using the Whoosh library API, ensuring efficient and accurate query matching. The retrieved results presented using TF_IDF model and if they are deemed irrelevant by the user, they can be refined using various conversational search assistance features, which will be described below. Figure 4.1 illustrates the initial step of the image search process. During the dialogue, the conversational search agent asks a user when the presented results are unsatisfactory. If user decides that results are not satisfactory, conversational assistant will ask him the multimodal clarification question, as illustrates in Figure 4.1. The user is presented with two images, and invited to say if one of them is similar to the features desired in relevant target images. The distance between a selected image and the images in retrieved gallery is calculated using L2 distance. The gallery is then filtered to remove images with low similarity to the selected image, making navigation of the gallery easier for the user. The images for clarification questions are selected randomly from among those with high text matching scores to the user's query. Since the images in the clarification questions are taken from the target search collection, L2 distance between a selected clarification image and the images in the gallery can be precomputed to improve user interaction. If the user does not find either of the presented clarification images useful, they can reject them both, and be presented with a new pair of images. They can quit the clarification process at any point if they do not find it productive, and enter a revised text search query in an attempt to progress the search process.

The user has the option to terminate the current search dialogue by restarting the process or by submitting a new search query. The search results updated in the image gallery connected to the search agent. A search dialogue concludes either

when the user successfully finds the desired results or decides to terminate the process after an unsuccessful search.

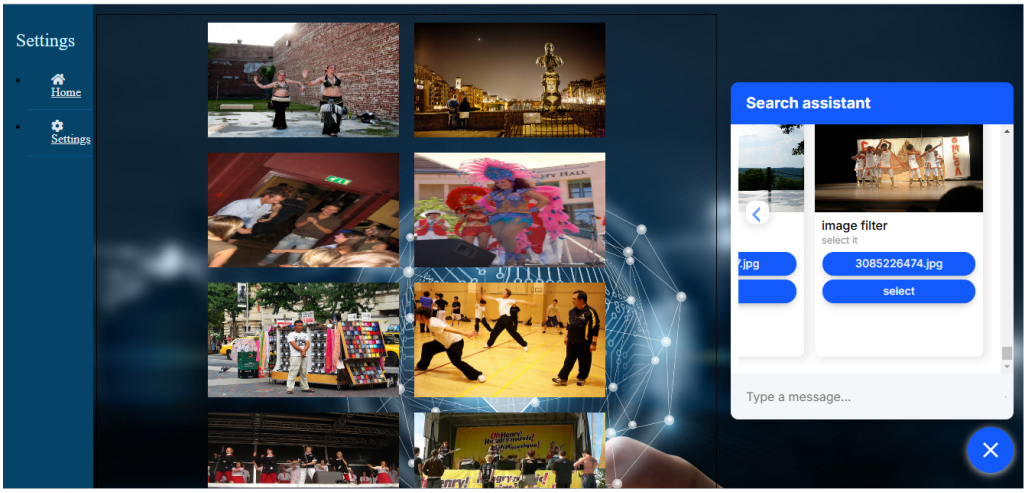

Figure 4.2: Conversational framework asks about the clarification

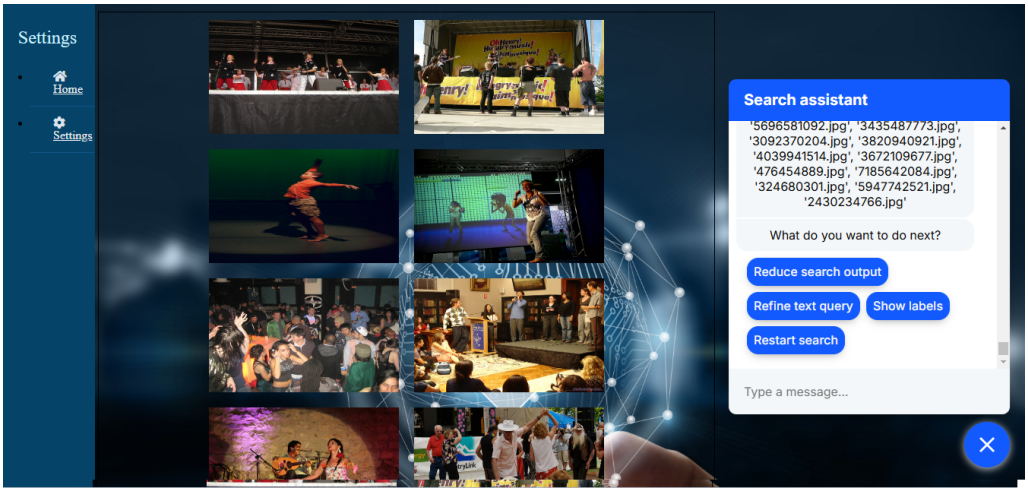

Figure 4.3: Image search output representation

Where the initial search output contains some relevant results, the user can use the dialogue to reduce the amount of provided images with multimodal clarification questions. To do this, the framework selects several images which are close to the user's search query (using text keyword labels) and displays them in the search dialogue to the user as potential filters, as illustrated in Figure 4.2. The search results are updated interactively as the user selects filters while the search dialogue progresses, which means the calculation the L2 distance between selected image and

images in the gallery. The search results are shown in the image gallery connected to the search agent, as demonstrated on Figure 4.3.

Conventional framework

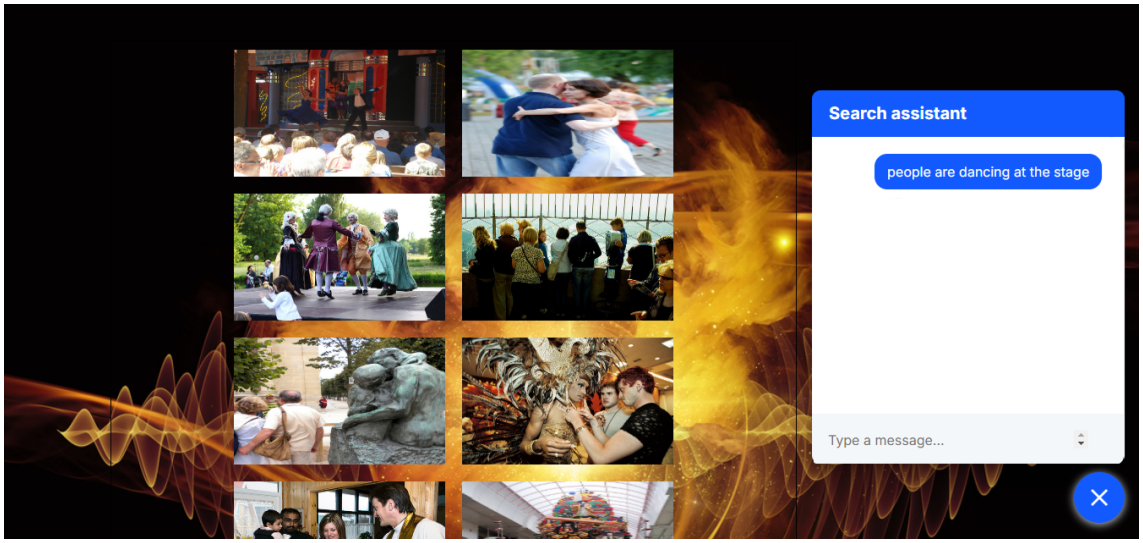

Figure 4.4: The interface of the traditional search system

The equivalent conventional image search framework was developed using the same technologies and adopts the same image search archive, including preprocessing and indexing the archive with the Whoosh Search API, and utilizes the same image search archive.[1]. This ensures consistency in functionality while allowing for a comparison between conversational and conventional search approaches.

Figure 3.8 illustrates the interaction process within the conventional framework: the user submits a text-based search query, and the corresponding image results are retrieved and displayed in the gallery for review. The search results appear in the image gallery adjacent to the search window on the right.

Unlike the conversational framework, the conventional system supports only text-based search functionality. As a result, participants must manually scroll through the search output or refine their query to locate the desired image. Alternatively, users can restart the search process entirely by entering a new search query. We

[1]https://anonymous.4open.science/r/traditional-framework-for-images-AFAD/

designed the conventional framework by referring to established image search platforms, such as COCO Explorer [2] for MS COCO and the more advanced FiftyOne [3], which support text-based search functionality along with some content-based features. However, for our specific objective—examining the impact of conversational search (CS) assistance in image retrieval, we implemented a text-based search system only, ensuring a controlled evaluation of the conversational framework's effectiveness.

4.2.2 Search task design

In the experimental session, participants are required to complete designated multimedia search tasks and corresponding questionnaires. The session begins with a training task to familiarize participants with the search application and task requirements. Following the training, participants perform two search tasks, each focusing on finding a specific image. The selected user scenarios were designed to reflects those commonly encountered by content creators, such as writers, illustrators looking for image references and individuals preparing presentations, and others who seek appropriate images to complement their work. These scenarios reflect typical use cases in the context of image search, making them highly relevant for our study. During these tasks, their search activities will be systematically logged for future analysis.

4.2.3 Experimental procedure

In this section we describe the experimental procedure for our comparison of conversational and conventional approaches for image search. We seek to gain insights into how conversation engagement might be directly incorporated into current user search activities, and to explore opportunities to enhance the user's search experience [10].

The study utilized the previously described search task instructions to assess the

[2] `https://cocodataset.org/`
[3] `https://docs.voxel51.com/`

time required to complete each section of the study, analyze participant behavior, and collect feedback on their experiences with the conversational search approach for the image search process. Participants engaged in a pre-search survey, then conducted search tasks using a conversational image search application and a corresponding conventional one, and finally, completed a post-search survey to provide feedback and compare their experiences with the two systems.

Pilot study

Prior to conducting our main study, a pilot study with three PhD students with STEM background was conducted using image search tasks to see how long it took them to complete the sections of the study, gain insights into the likely behaviour of participants, and to generally debug and refine the experimental setup.

The following feedback was gathered during the pilot study and informal interviews with the participants:

- Enhance the error handler and make the conversational agent more dependable, there are two critical improvements to make.
- Update the restart scenario and provide an option to enter a new query without restarting the entire search story.

Furthermore, during discussion, the following useful recommendations were made:

- It would be helpful to have a cheat sheet or assistance that makes the conversational agent's actions more transparent
- To make the agent more proactive, it was suggested to provide the option to select the reformulated text query.
- Thus, it may be beneficial to develop a more precise mechanism for making image suggestions.

Based on user feedback and identified challenges in formulating effective search queries, we designed and implemented a query expansion feature to assist users in refining and broadening their search inputs. This feature leverages a Large Language

Model (LLM), specifically GPT-4, to generate multiple alternative, reformulated query options based on the user's initial input.

By providing several reformulated options, the framework enables users to select the most relevant phrasing, thereby increasing the likelihood of retrieving desired images. This functionality fosters a more dynamic and interactive search process, allowing users to refine their queries efficiently and achieve more accurate and relevant search outcomes.

In addition to the query expansion feature, other user-recommended improvements were implemented. However, enhancements to the algorithm for image suggestions were deferred, as this would require significant time for further investigation and development.

4.3 User study

In this section, we describe the details of our user study, which aims to enable us to observe and better understand and contrast the behaviour of non-specialist searchers whose techniques for using search engines are generally learned from personal experience.

The main study involved 20 participants, all with STEM backgrounds, including MSc students, PhD candidates, and postdoctoral researchers. The participant group comprised 9 males and 11 females, with ages ranging from 23 to 37. Among them, 15 participants reported using image search engines, predominantly Google Images, for over 10 years. Additionally, 11 participants had experience using conversational assistants for general search tasks, which they described as user-friendly but not highly informative and 4 of them are using CS tools regularly. The average level of interest in the topic of conversational search, rated on a scale from 1 to 7, was 4.8, with 1 being the minimum and 7 the maximum.

A total of 25 instruction sheets were prepared for the study, each containing detailed step-by-step instructions and two randomly selected images sourced from the MS COCO dataset. These instructions provided participants with clear guidelines

on how to complete the search tasks and ensured consistency. The images included in each sheet were chosen to represent a variety of content types, such as objects, scenes, and activities, to evaluate the versatility of the search frameworks.

To ensure fairness and minimize bias, the instruction sheets were distributed randomly among the participants. This randomization aimed to balance the difficulty of search tasks across the participant pool and eliminate any systematic variations that could influence the study results. By using this approach, the study maintained a diverse set of search scenarios, enabling a comprehensive evaluation of both the conversational and conventional search frameworks. As demonstrated in similar user studies, as demonstrated in previous studies this participant sample was expected to be sufficient to make meaningful insights and conclusions [3].

4.3.1 User experience questionnaire results

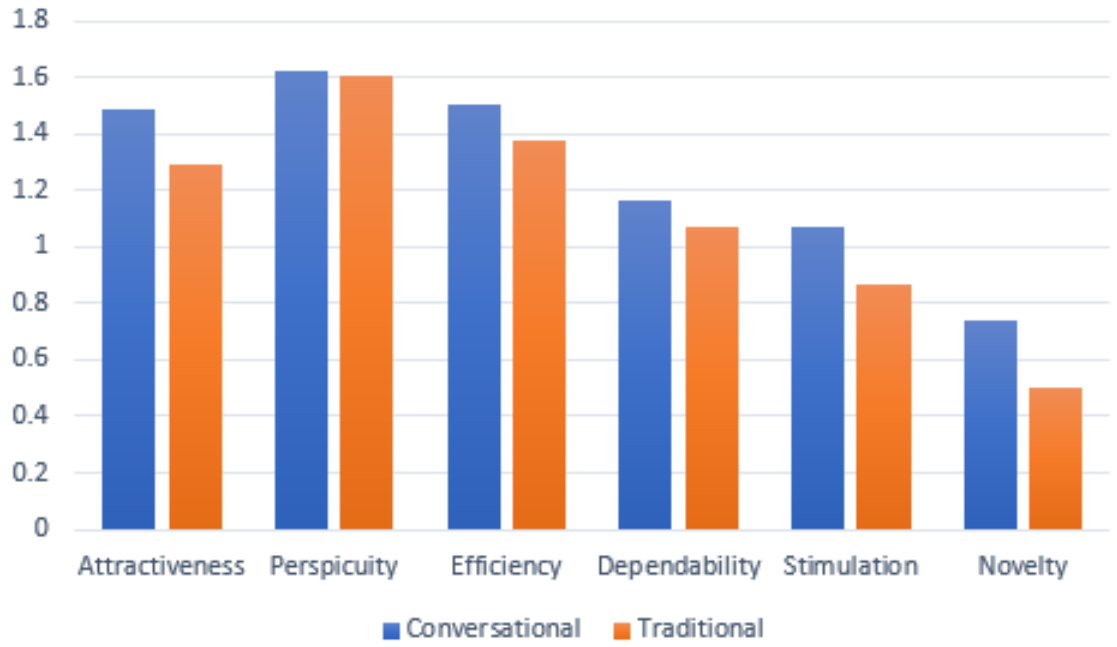

Figure 4.5: The comparative results of UEQ for conversational and traditional frameworks

After completing all search tasks, users were asked to evaluate their experiences using post-search online questionnaires. Initially, participants assessed their experience with the conversational framework, followed by an evaluation of the con-

ventional framework for image search.

Using the UEQ metrics, both frameworks were systematically evaluated. The results, presented in Figure 4.5, reveal that the conversational framework outperformed the conventional framework in key areas. It achieved higher scores in Attractiveness (overall appeal), Efficiency, and Dependability (pragmatic qualities), as well as Stimulation and Novelty (hedonic qualities). Importantly, Perspicuity (another pragmatic quality) was rated relatively high for both frameworks, indicating their ease of understanding and usability.

These findings highlight the conversational framework's ability to provide a more engaging, efficient, and user-friendly search experience compared to the conventional approach.

Scale	Conversational	Traditional
Attractiveness	1.49 (Above Average)	1.29 (Above Average)
Perspicuity	1.62 (Above Average)	1.61 (Above Average)
Efficiency	1.51 (Good)	1.38 (Above Average)
Dependability	1.16 (Above Average)	1.07 (Below Average)
Stimulation	1.07 (Above Average)	0.87 (Below Average)
Novelty	0.74 (Above Average)	0.50 (Below Average)

Table 4.1: Detailed results of comparison conversational and traditional frameworks on the UEQ benchmark

The measured scale means were compared against existing values from a benchmark UEQ dataset [8], [83]. This benchmark dataset comprises responses from 21,175 participants collected across 468 studies evaluating various products. It provides a standardised framework for interpreting UEQ scores by categorising them into predefined quality levels, such as 'Below Average,' 'Above Average,' 'Good,' and 'Excellent.' These categories allow for a meaningful comparison of results and the contextualization of user experience evaluations.

According to the benchmark measurements, as presented in Table 4.1, the UEQ results for the conversational search framework fell within the 'Above Average—Good' range across most dimensions. In contrast, the traditional search framework scored lower, ranging between 'Above Average' and 'Below Average,' particularly in Effi-

ciency and hedonic qualities such as Stimulation and Novelty.

The obtained results reflected that users found the conversational search approach effective and convenient for the image retrieval task.

4.3.2 Statistical analysis

A one-way MANOVA (Multiple Analysis of Variance) was conducted to examine the variation between UEQ scales for both framework types, conversational and traditional. The one-way MANOVA allows for a test on each dependent variable (DV) to understand whether the scale result is changed by the framework type selected as the independent variable (IV).

Independent Variable	Dependent Variable	F-Test
Framework type	Attractiveness	$F = 6.5$, $p = 0.001$
	Perspicuity	$F = 0.52$, $p = 0.71$
	Efficiency	$F = 4.74$, $p = 0.0036$
	Dependability	$F = 5.8$, $p = 0.0011$
	Stimulation	$F = 9.5$, $p << 0.01$
	Novelty	$F = 12.61$, $p << 0.01$

Table 4.2: Summary table of results

The one-way MANOVA revealed a significant effect of the group on the scores of all six scales (4.2). Post hoc analyses showed that the participants rated the conversational search framework significantly better than the traditional one on Attractiveness ($p = 0.001$), Efficiency ($p = 0.0036$), Dependability ($p = 0.0011$), Stimulation ($p << 0.001$), and Novelty ($p << 0.001$), and although this difference was also nearly significant for Perspicuity ($p = 0.71$), which correlated with comparing results for the UEQ benchmark.

4.3.3 Search strategies and behaviour

In this section, we discuss user behaviour during the search session. Users tend to send an equal number of requests to both the conversational and traditional search interfaces. However, it is important to note that conversational search scenarios

generally take longer to complete. This is primarily due to participants taking additional time to explore the various features of the interface, as highlighted in the feedback received from users.

User's activity	Conversational	Traditional
Number of interactions	30.25	8.15
Number of queries	7.45	6.4
Restart	2.8	1.75
Reduce the search output	4.4	N/A
Rewrite the query	1.2	N/A
Ask about another image	4.75	N/A
Return to previous results	0.6	N/A
Time for the search task	24.5 (min)	21.4 (min)

Table 4.3: Summary table of average results

The quantitative results are shown in Table 4.3, which presents the average number of interactions and the time spent per search task per user. For the purposes of this study, one interaction is defined as a single user-initiated action within the search interface. This can include either typing a command into the input field or selecting a predefined option by clicking a button from the list of available commands. These results indicate that participants engaged in more interactions within the conversational search framework compared to the traditional search method. This increased number of interactions can be attributed to the additional functionalities provided by the conversational interface, which encourage users to explore various options and tools.

We aimed to analyze user behaviour while interacting with the search tool, focusing on the types of interactions they performed and the underlying reasons for their behaviour. To facilitate this analysis, we categorized search behaviour into four distinct categories, enabling a structured approach to understanding user actions and motivations:

- **User type 1:** The user enters one text query and marks one image from the search output to fulfill the information need.

- **User type 2:** The user enters one text query and chooses several images from the search output to fulfill the information need.
- **User type 3:** The user progressively refines their text search queries, adding more detail with each iteration to narrow down the search output and achieve more accurate results. Additionally, the user selects specific images from the automatically refined search output to better fulfill their information needs.
- **User type 4:** The user simplifies the text search query to a single word and interacts with the conversational agent by using image filter suggestions to automatically refine the search output, selecting several images as relevant. In contrast, within the conventional framework, the user manually scrolls through the search results to locate the desired content and similarly identifies several images as relevant.
- Additionally there is the possibility of the case where the user issues one or more queries, but does not select any of the retrieved image. This may indicate that either the user retrieves no relevant items or cannot identify retrieved relevant items or is able to satisfy their information need from the provided images.

In conclusion, the analysis highlights a diverse range of user behaviours, reflecting the varying strategies adopted in both conversational and conventional search environments. The conversational framework not only supported a higher level of engagement but also accommodated different user preferences in query formulation and interaction style. These findings underscore the potential of conversational interfaces to enhance user experience through more personalized and flexible search pathways.

4.3.4 Chatbot usability questionnaire results

Table 4.4 shows the odd question numbers of the CUQ have statements that relate to the positive aspects of the conversational agent. On a scale of 1—Strongly

Positive aspects	Scale (1-5)
Q1:Realistic and engaging	3.5
Q3:Welcoming at setup	4.5
Q5:Explained purpose well	3.5
Q7:Easy to navigate	4.5
Q9:Understood me well	3
Q11:Useful informative responses	3.5
Q13:Coped well with mistakes	2.5
Q15:Very easy to use	4

Table 4.4: The average ranking for the positive aspects of the agent's usability

Disagree to 5—Strongly Agree to the positive statements about usability, Question 3, which states 'The chatbot was welcoming during initial setup', had the highest average ranking of 4.5, corresponding to Agree and Question 7, which states 'Easy to navigate'. The lowest average ranking was 2.5 for Question 13, which states, 'Coped well with mistakes'.

In Table 4.5, the average ranking on a scale of 1—Strongly Disagree to 5—Strongly Agree of the CUQ even question numbers with statements related to the negative aspects of the chatbot are shown. Question 2, which states 'Too robotic' has an average ranking of 4. With an average ranking of 1, Question 16, which states 'Very complex', had the lowest ranking.

Negative aspects	Scale (1-5)
Q2:Too robotic	4
Q4:Very unfriendly	2
Q6:No purpose indication	2
Q8:Confusing	2.5
Q10:Failure to recognise inputs	2.5
Q12:Irrelevant responses	1.5
Q14:Unable to handle errors	2.5
Q16:Very complex	1

Table 4.5: The average ranking for the negative aspects of the agent's usability

The CUQ results indicate that participants generally found the conversational search framework to be clear, user-friendly, and intuitive to navigate. Users appreciated the overall structure and ease of interaction, which contributed to a positive experience. However, one recurring concern highlighted by several participants was

the limited flexibility of the predefined or suggested responses offered by the search assistant during the dialogue. This constraint sometimes hindered the fluidity of the conversation and reduced the system's ability to fully adapt to user needs in more complex or nuanced search scenarios.

4.4 Analysis Summary

Our evaluation reveals that the conversational assistant can be successfully used for image retrieval, offering a more interactive and engaging user experience than the equivalent conventional search framework. Through our study, we identified several key components crucial to fostering successful user adoption of this technology:

- Some users spend additional time exploring the various interactive features of the conversational interface. The additional complexity of navigating the interactive options contributes to better precision, but requires more time compared to working with traditional image search interfaces.

- Nearly all participants reported that the conversational search framework was both effective and easy to understand. They found the interface intuitive, and the conversational flow helped guide their search process more naturally than conventional methods.

- The use of multimodal clarification questions within the search dialogue was widely adopted by participants. They noted that this feature significantly improved the efficiency of the search process, as it allowed them to 'click and reduce output' rather than scrolling through long lists of results manually. This interaction style makes the process smoother and more user-friendly.

4.4.1 Limitations

While this study provides valuable insights into the effectiveness of conversational search frameworks for multimedia information retrieval, certain limitations must be

acknowledged. These limitations arise from factors such as the experimental design, participant demographics, dataset constraints, and the data collection process. Addressing these challenges in future research will help to strengthen the findings and further refine the development of conversational search systems. Below, we discuss the key limitations identified in this study:

- Sample profile: the profile of many of the participants may be similar (e.g. age, background, current academic status), which may not accurately represent wider population groups and the results that may have been obtained from a more diverse sample.

- Data collection process: social desirability bias may be present in the questionnaire, as users may under-report bad experiences and over-report good experiences with the assistant in order to please the researcher.

- User interactions: Users have to do more actions in the conversational search interface, especially during the first attempt of usage, because the conversational interface contains more functions than the traditional interface.

4.5 Concluding Remarks

This user study compared conversational and non-conversational search frameworks, examining their impact on user experience during an image search process. Based on our results, users found the conversational search approach to be both helpful and effective, enhancing their overall interaction with the search system.

The core aim of our study was to investigate how multimodal conversational assistance could simplify and clarify the image search process for users. Through our experiment, we demonstrated that our multimodal dialogue-based search assistant makes the overall user experience more intuitive and engaging. The structured, interactive dialogue provided by the assistant enabled users to navigate the search more efficiently, reducing cognitive load and making the process more user-friendly.

Our findings underline the potential of conversational search to streamline multimedia retrieval tasks, offering a more flexible and personalized experience. In the next chapter, we aim to extend the capabilities of our multimodal conversational search assistant to video retrieval. This will involve integrating conversational features to support the growing demand for effective video search, further advancing our approach to multimodal search interfaces and expanding the scope of our framework's application.

Chapter 5

A Comparative Study of Conversational and Conventional Search Methods for Video Retrieval

5.1 Overview

As we saw in the previous chapter, the role of conversational engagement in search systems has the potential to enhance the search experience of users when carrying out image search tasks. Next study move beyond image search to the video retrieval task. Again we examine its impact on the user's search experience and the effectiveness of their search.

Our study uses a version of our conversational search framework extended to support engagement with video content. The user can again engage directly with the search system while receiving suggestions from the search assistant to help them refine their queries and guide their interactions.

5.2 TRECVID workshops results

Participation in the TRECVID (TREC Video Retrieval Evaluation) workshops offered a valuable opportunity to gain hands-on experience with state-of-the-art video processing techniques and evaluation methodologies. TRECVID is a well-established benchmarking initiative, coordinated by the National Institute of Standards and Technology (NIST), which focuses on advancing the research in video information retrieval by organising a series of annual shared tasks.

During my involvement in TRECVID 2021, I participated in both the video summarisation and video question-answering tracks. These experiences significantly deepened my understanding of large-scale video datasets, annotation methods, and video preprocessing pipelines [73]. The summarisation task, in particular, provided insights into how to efficiently extract meaningful segments from long video content — a skill that later informed the development of the multimedia preprocessing module in this research. Additionally, my engagement in the video question-answering track enhanced my capacity to link visual content with text queries, which aligned closely with the goals of this thesis.

In TRECVID 2022, I contributed to the Deep Video Understanding (DVU) challenge, which introduced more advanced and semantically rich methods of video analysis. A key outcome of this experience was the implementation and exploration of techniques and tools, which were used to generate detailed and structured textual representations of video content. These enhanced captions and semantic descriptors supported more accurate video retrieval and were particularly influential in shaping the design of the search framework in this PhD project [74]. The importance of generating rich, full-text descriptions for video segments — as highlighted in the DVU challenge — directly contributed to the design of the video archive used in this work, as well as the integration of automatic captioning tools and indexing strategies [58].

5.3 Experimental methodology

In this section, we describe the details of our user study exploring our prototype conversational video search application and contrasting this with an equivalent conventional video search system without conversational assistance. This study compares conversational and traditional search approaches within video search scenarios, focusing on outcomes that influence user engagement. The findings provide valuable insights for refining and improving the framework. We seek to gain insights into how conversation engagement might be directly incorporated into current user search activities, and to explore opportunities to enhance the user's search experience. The study was performed online, and participants used a setup of two systems deployed in the Amazon Cloud. One system is a traditional video search application, while the other incorporates a conversational agent.

5.3.1 Experimental framework

In this section, we provide details about a prototype conversational video search framework and compare it to a conventional video search system that lacks conversational assistance.

Dialogue Flow

The dialogue module of the conversational application utilizes a similar DIET model to support search dialogue and the Whoosh Search API for query processing, as previously described in Chapter 4. The video database, MSR-VTT [101], was also indexed using the same API to ensure consistency in retrieval methodology.

The retrieved results are presented using the TF_IDF model. If the results are deemed irrelevant by the user, they can be refined through various conversational search assistance features, detailed below. This refinement process enables users to engage in an interactive dialogue to apply additional filters, further narrowing down the search results. This iterative and dynamic approach improves the accuracy of search outcomes while creating a more engaging user experience.

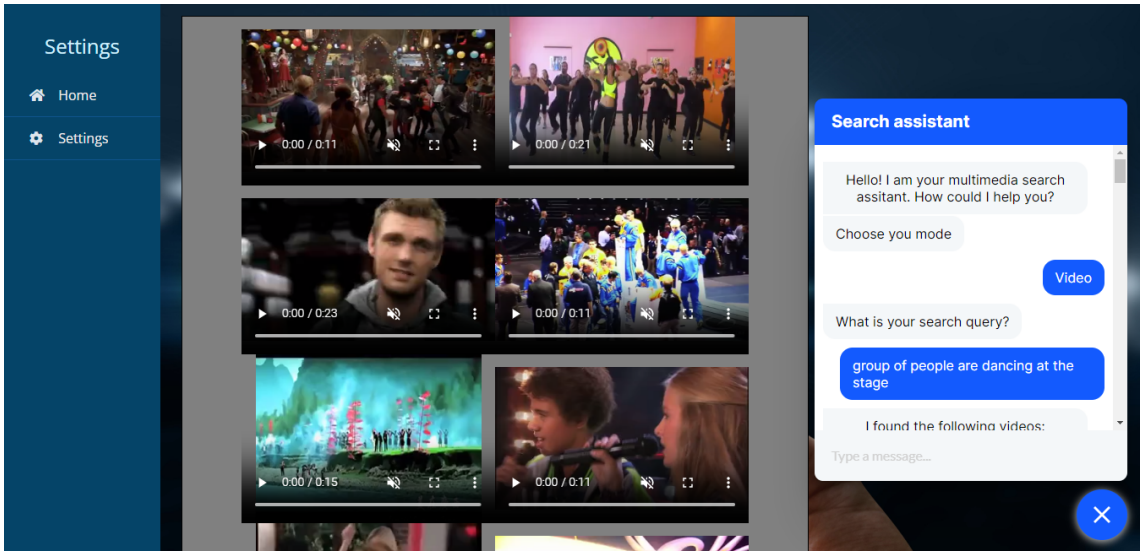

Figure 5.1: The interface of conversational search system

When the initial search output includes some relevant results, as illustrated in Figure 5.1 the user can use multimodal clarification questions to refine the results. Users can play a video within the provided interface—only one video at a time—by clicking the play button. The conversational assistant suggests clarification by selecting images from an image search archive [31] that are closely aligned with the user's text query. The similarity between the user's query and the text captions of preprocessed images is calculated using the Whoosh library API. The clarification images are presented in the search dialogue, as demonstrated on Figure 5.2, enabling the user to filter results interactively. If the user decides that the results remain unsatisfactory, the assistant will suggest additional clarifications or allow the user to restart the search process with a new query. The search dialogue concludes when the user successfully retrieves the desired results or terminates the session after an unsuccessful search attempt.

As part of the multimodal refinement process, when relevant results are present in the initial output, the framework dynamically reduces the set of displayed videos using clarification questions. The system selects several images based on their textual keyword labels, presenting them in the search dialogue as potential filters. As the user selects filters, the framework calculates the L2 distance between the chosen

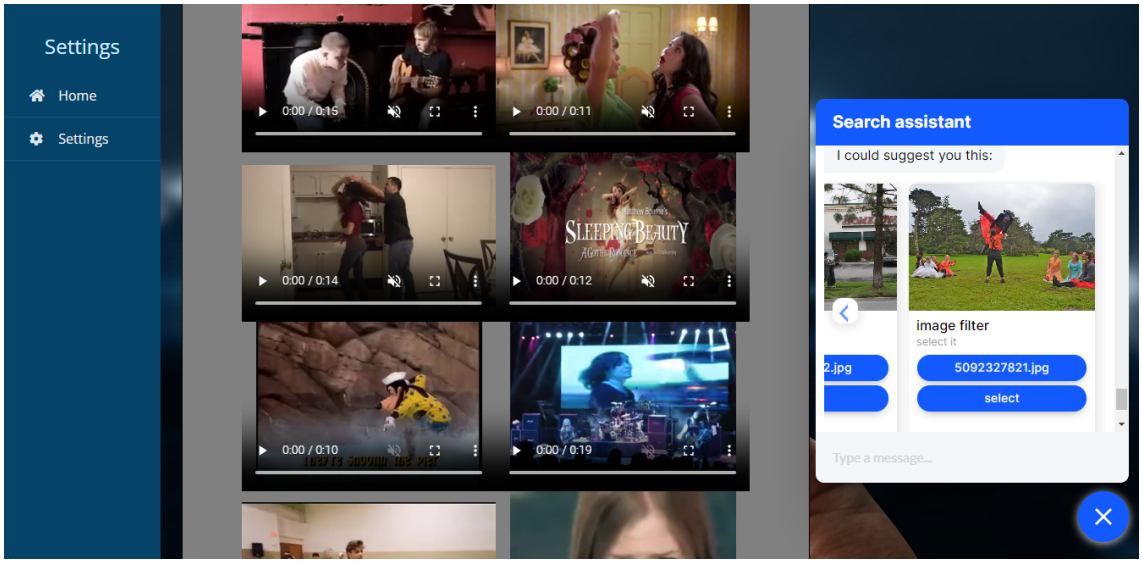

Figure 5.2: Multimodal clarification question for video search

image and video frames. This real-time calculation ensures that only the most relevant frames are retained in the search results, which are then displayed in the video gallery linked to the search agent. This seamless and responsive process enhances the search experience.

Conventional framework

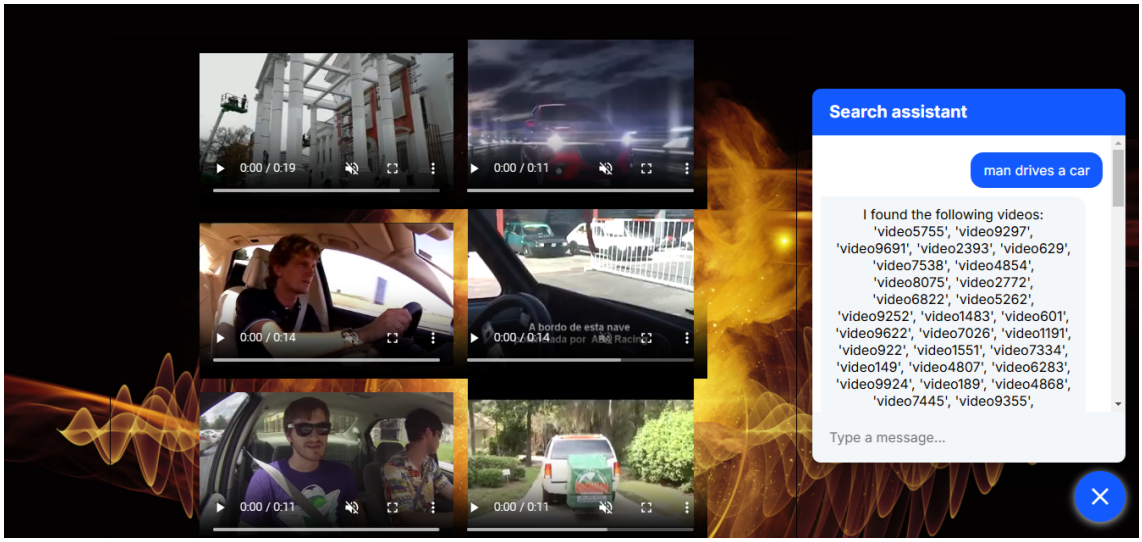

Figure 5.3: The interface of traditional search system

The equivalent conventional video search framework was developed using the

same underlying technologies, including preprocessing and indexing the archive with the Whoosh Search API, and utilizes the video search database MSR-VTT. This ensures consistency in functionality while enabling a direct comparison between conversational and conventional search approaches.

Figure 5.3 illustrates the interaction process within the conventional framework: the user submits a text-based search query, and the corresponding video results are retrieved and displayed in the gallery for review. Unlike the conversational framework, the conventional system supports only text-based search functionality. Consequently, participants must manually scroll through the search output or refine their query to locate the desired video content. If the results remain unsatisfactory, users can restart the search process entirely by entering a new search query. This process, while functional, lacks the interactive and dynamic refinement options offered in the conversational framework.

5.3.2 Search tasks design

The design of the search tasks is largely similar to the previous experiment, with the primary distinction being the focus on video retrieval as the search problem. In this session, participants will complete assigned multimedia search tasks and associated questionnaires. The session begins with a sample training task designed to familiarize participants with the search application and its requirements. After completing the training, participants will engage in two search scenarios, each focused on locating a specific video. Throughout these tasks, their search activities will be systematically recorded for subsequent analysis.

5.3.3 Experimental procedure

In this section, we outline the methodology for our experimental comparison of conversational and traditional approaches to video search. The primary objective of this study was to gain insights into how conversational engagement can be directly integrated into existing user search activities and to explore opportunities for enhancing

the user experience [82].

The study employed the previously described search task instructions to evaluate the time required to complete each section, analyze participant behaviour, and collect user feedback on their experiences with the conversational search approach for video retrieval. Participants began with a pre-search survey to establish baseline information, followed by performing search tasks using both a conversational video search application and a corresponding conventional search system. After completing the tasks, participants were asked to complete a post-search survey to provide detailed feedback and compare their experiences with the two systems.

The study methodology was designed to assess the effectiveness, usability, and overall engagement offered by the conversational framework compared to its conventional counterpart.

Pilot studies

A pilot study with two PhD candidates was conducted to see how long it took them to complete the study sections, gain insights into the likely behaviour of participants, and generally debug the experimental setup.

Participants of the pilot study provided the following feedback on the conversational framework:

- Clarity and Usability: The framework was clear and easy to use, making it accessible for users.
- Helpfulness: Participants found the conversational approach very helpful for conducting searches.
- Loading Speed: The speed of loading the searched videos could be improved to enhance user experience.
- Image Accuracy: The accuracy of suggested images in the search dialogue could be further refined to better align with user queries.

The loading issues for searched videos were resolved prior to the commencement of the main user study by improving the characteristics of virtual machines. However, the accuracy of image suggestions remained unchanged, as addressing this would require significant time and resources for further updates.

5.4 User study

In this section, we describe the details of our user study, which aims to enable us to observe and better understand and contrast the behaviour of non-specialist searchers whose techniques for using search engines are generally learned from personal experience.

The main study was conducted used 17 participants, all of whom had STEM backgrounds, ranging from MSc students to PhD candidates and postdoctoral researchers. The participant group consisted of 8 males and 9 females, with an age range of 23 to 37. Among them, 13 participants reported using video search engines regularly, predominantly Youtube, for over 10 years. Additionally, 6 participants had experience using conversational assistants for general search tasks (mainly ChatGPT), which they described as user-friendly but not highly informative and only 2 participants are using the CS tools regularly. The average level of interest in the topic of conversational search, rated on a scale from 1 to 7, was 5.1, with 1 being the minimum and 7 the maximum.

A total of 25 instruction sheets were prepared for the study, each containing detailed step-by-step instructions and two randomly selected and rewrite video captures sourced from the MSR-VTT dataset. These instructions provided participants with clear guidelines on how to complete the search tasks and ensured consistency.

Each search task sheet contains two distinct textual queries, each describing a desirable search video. For each scenario, participants describe the content of the target video in their own words, generating a natural language text query based on what they observe. They then input this text query into both frameworks, which process the descriptions and return a set of search results. Using the tools provided

by the frameworks, participants navigate the search results to identify and select the videos they find most relevant.

The task is considered complete once the user has successfully found videos that closely match the content of the provided videos for both scenarios. This process helps collect valuable feedback on the frameworks' performance, highlighting how well they handle natural language queries and provide relevant results based on user input.

5.4.1 User experience questionnaire results

After completing all search tasks, users were asked to evaluate their experiences using post-search online questionnaires. Participants completed a post-search survey, which included the Chatbot Usability Questionnaire to evaluate the conversational interface and the User Experience Questionnaire for both frameworks. Additionally, participants indicated whether they found relevant videos (if any) for each search task across both frameworks and participants had an opportunity to leave any additional feedback if they had any. Using the UEQ metrics, both frameworks were

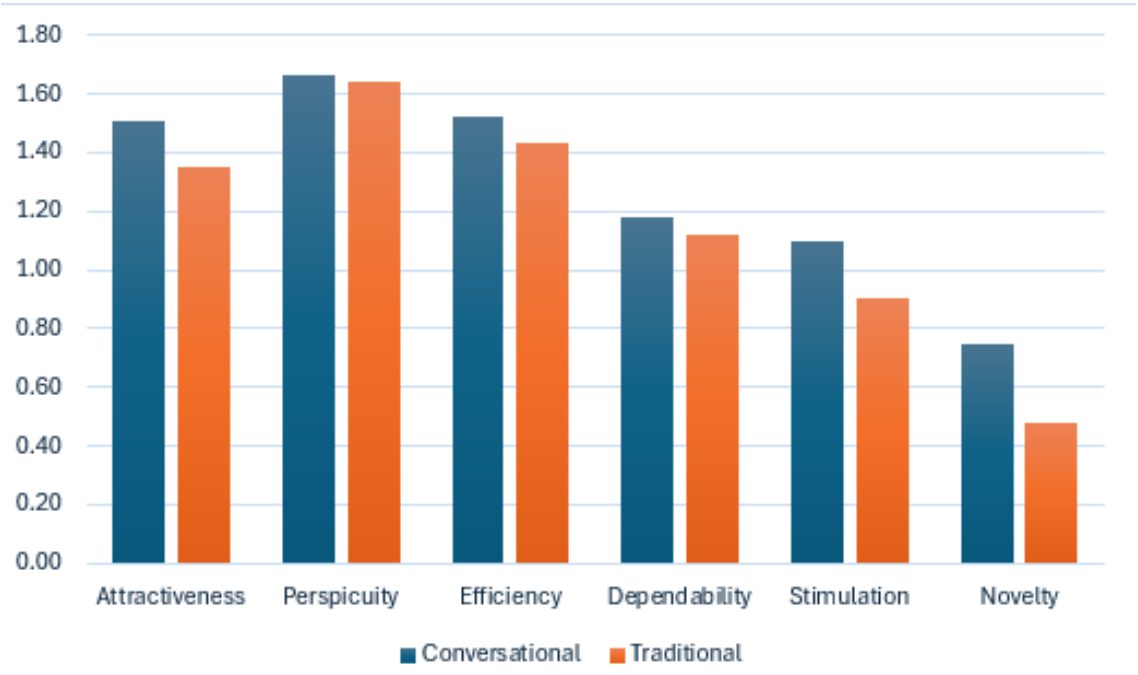

Figure 5.4: The comparative results of UEQ for conversational and traditional frameworks

evaluated by users. The results, illustrated in Figure 5.4, show that the conversational framework outperformed the traditional framework across all dimensions. It achieved higher scores in Attractiveness (overall appeal), Efficiency and Dependability (pragmatic qualities), as well as Stimulation and Novelty (hedonic qualities). Notably, Perspicuity (a pragmatic quality) received relatively high ratings for both frameworks, reflecting their ease of understanding and usability.

Scale	Conversational	Traditional
Attractiveness	1.51 (Above Average)	1.35 (Above Average)
Perspicuity	1.66 (Above Average)	1.64 (Above Average)
Efficiency	1.52 (Good)	1.43 (Above Average)
Dependability	1.18 (Above Average)	1.12 (Below Average)
Stimulation	1.1 (Above Average)	0.9 (Below Average)
Novelty	0.75 (Above Average)	0.48 (Below Average)

Table 5.1: Detailed results of comparison conversational and traditional frameworks on the UEQ benchmark

The measured scale means were compared against values from a benchmark UEQ dataset [8]. According to this benchmark, as presented in Table 5.1, the UEQ results for the conversational search framework are positioned within the 'Above Average—Good' range, while the results for the traditional framework fall within the 'Above Average—Below Average' range, particularly for hedonic qualities such as Stimulation and Novelty. These findings indicate that users perceived the conversational search approach as more effective and convenient for completing the video retrieval task.

5.4.2 Statistical analysis

As we discussed before one-way MANOVA (Multiple Analysis of Variance) was conducted to examine the variation between UEQ scales for both framework types, conversational and traditional.

The one-way MANOVA revealed a significant effect of the group on the scores across all six scales (5.2). Post hoc analyses indicated that participants rated the conversational search framework significantly higher than the traditional framework

Independent Variable	Dependent Variable	F-Test
Framework type	Attractiveness	F = 12.23, p << 0.001
	Perspicuity	F = 0.22, p = 0.92
	Efficiency	F = 3.4, p = 0.002
	Dependability	F = 7.3, p = 0.0003
	Stimulation	F = 13.12, p << 0.01
	Novelty	F = 31.08, p << 0.01

Table 5.2: Summary table of results

in terms of Attractiveness (p << 0.001), Efficiency (p = 0.002), Dependability (p = 0.0003), Stimulation (p << 0.001), and Novelty (p << 0.001). Although the difference for Perspicuity was not statistically significant (p = 0.92), it showed a trend consistent with the comparative results from the UEQ benchmark.

5.4.3 Search strategies and behaviour

In this section, we analyze user behaviour during the search sessions. Participants tended to submit a similar number of search requests to both the conversational and traditional search interfaces. However, conversational search scenarios typically required more time to complete. This extended duration can be attributed to participants spending additional time exploring the diverse features of the conversational interface, as emphasized in the feedback provided by users.

User's activity	Conversational	Traditional
Number of interactions	21.5	10.15
Number of queries	9.00	8.4
Restart	6.2	7.5
Reduce the search output	8.3	N/A
Rewrite the query	3	N/A
Ask about another image	6.1	N/A
Return to previous results	1.1	N/A
Time for the search task	20.2 (min)	21.6 (min)

Table 5.3: Summary table of average results

The quantitative results are shown in Table 5.3, which outlines the average number of interactions and time spent per search task per user.

In our study we were interested to analyze user behaviour with a standard user driven search tool in terms of the interactions they make and to seek to understand the reasons for their behaviour. In doing this, we divided search behaviour into four categories:

- **User type 1:** The user enters one text query and marks one video from the search output to fulfill the information need.
- **User type 2:** The user enters one text query and chooses several videos from the search output to fulfill the information need.
- **User type 3:** The user simplifies the text search query from sentence to a single word and interacts with the conversational agent by using image filter suggestions to automatically refine the search output, selecting several videos as relevant.
- **User type 4:** The user constructed more accurate text search query and scrolling the search output manually in conventional and conversational frameworks.
- Additionally there is the possibility of the case where the user issues one or more queries, but does not select any of the retrieved video, this may indicate that either the user retrieves no relevant items or cannot identify retrieved relevant items or is able to satisfy their information need from the provided videos.

5.4.4 Chatbot usability questionnaire results

In Table 5.4, the odd question numbers of the CUQ have statements that relate to the positive aspects of the conversational agent. On a scale of 1—Strongly Disagree to 5—Strongly Agree to the positive statements about usability, Question 15, which states 'The chatbot is very easy to use', had the highest average ranking of 4.4,

Positive aspects	Scale (1-5)
Q1: Realistic and engaging	3.4
Q3: Welcoming at setup	4.05
Q5: Explained purpose well	4.3
Q7: Easy to navigate	4.05
Q9: Understood me well	4
Q11: Useful informative responses	4
Q13: Coped well with mistakes	3.5
Q15: Very easy to use	4.4

Table 5.4: The average ranking for the positive aspects of the agent's usability.

corresponding to Agree. The lowest average ranking was 3.4 for Question 1, which states, 'The chatbot is realistic and engaging'.

In Table 5.5, the average ranking on a scale of 1—Strongly Disagree to 5—Strongly Agree of the CUQ even question numbers with statements related to the negative aspects of the chatbot are shown. Question 2, which states 'Too robotic' has an average ranking of 3.3. With an average ranking of 1.35, Question 4, which states 'The chatbot is very unfriendly', had the lowest ranking.

Negative aspects	Scale (1-5)
Q2: Too robotic	3.3
Q4: Very unfriendly	1.35
Q6: No purpose indication	1.7
Q8: Confusing	1.94
Q10: Failure to recognise inputs	2.05
Q12: Irrelevant responses	2.3
Q14: Unable to handle errors	2.1
Q16: Very complex	1.5

Table 5.5: The average ranking for the negative aspects of the agent's usability.

CUQ results indicate that participants found the conversational search framework to be clear, user-friendly, and intuitive to navigate. Despite these strengths, participants highlighted a limitation: the restricted flexibility of the suggested responses provided by the search assistant during the dialogue, which occasionally limited the scope for more dynamic interactions.

5.5 Summary analysis

Our evaluation demonstrates that the conversational assistant is an effective tool for video retrieval, providing interactive and engaging user experience. Through the study, we identified several key components critical to fostering successful user adoption of this technology:

- Prior Familiarity with the Interface: Some users had previously participated in our study involving conversational search with images. As a result, during this second study, they spent less time interacting with the conversational search interface, as they were already familiar with its functionality.

- Effectiveness and Ease of Use: Nearly all participants reported that the conversational search interface was both effective and easy to understand. They found the interface intuitive, with the conversational flow naturally guiding their search process compared to conventional methods.

- Adoption of Multimodal Clarification Questions: Participants widely adopted the use of multimodal clarification questions within the search dialogue. This feature significantly improved search efficiency by enabling users to quickly and effortlessly refine their search output. Several participants highlighted the importance of clarification questions in determining whether the suggested images accurately matched their information needs.

- Interest in Expanding to Video Searches: Some participants expressed interest in having a similar conversational assistant for searching short videos on social networks. They were convinced of the effectiveness of the approach and its potential for improving the in search experience.

5.5.1 Limitations

While this study provides valuable insights into the effectiveness of conversational search frameworks for multimedia information retrieval, certain limitations must be

acknowledged. These limitations arise from factors such as the experimental design, participant demographics, dataset constraints, and the data collection process. Addressing these challenges in future research will help to strengthen the findings and further refine the development of conversational search systems. Below, we discuss the key limitations identified in this study:

- Sample profile: the profile of many of the participants may be similar (e.g. age, background, current academic status), which may not accurately represent wider population groups and the results that may have been obtained from a more diverse sample.

- Data collection process: social desirability bias may be present in the questionnaire, as users may under-report bad experiences and over-report good experiences with the assistant in order to please the researcher.

- User experience: Several users already participated in the previous comparative study, so that experience could make an effect to the current results.

5.6 Concluding Remarks

This thematic user study compared conversational and non-conversational search frameworks, examining their impact on user experience during the video search process. Based on our results, users found the conversational search approach to be both helpful and effective, enhancing their overall interaction with the search system.

The core aim of our study was to investigate how multimodal conversational assistance could simplify and clarify the video search process for users. Through our experiment, we demonstrated that our multimodal dialogue-based search assistant makes the overall user experience more intuitive and engaging. The structured, interactive dialogue provided by the assistant enabled users to navigate the search more efficiently, reducing cognitive load and making the process more user-friendly.

Our findings underline the potential of conversational search to streamline multimedia retrieval tasks, offering a more flexible and personalized experience. We extended the capabilities of our multimodal conversational search assistant to video retrieval. This involves integrating more sophisticated conversational features to support the growing demand for effective video search, further advancing our approach to multimodal search interfaces and expanding the scope of our framework's application.

In the next chapter, we will explore the incorporation of textual labeling displayed to users as an augmentation to the conversational search framework.

Chapter 6

A Study of Augmented Labelling Methods in Conversational Image Search

6.1 Overview

We conduct a user study designed to examine the role of augmented labeling in the conversational search for image search problem. Augmented labeling introduces additional contextual information, such as object detection labels or visual tags, designed to enhance the user's ability to refine and navigate through search results effectively. Our study focuses on comparing user experiences across different search scenarios, evaluating the use of augmented reality-based labeling.

In the multimodal clarification scenario, users engage with the system by interacting with the search dialogue, as described in the Chapter 5. In contrast, the augmented reality labelling scenario presents users with visual labels placed onto the provided images, offering additional information intended to offer further support to filter and refine search outputs.

Through this comparative analysis, we aim to identify the strengths and limitations of each approach, shedding light on how multimodal interactions and aug-

mented labels can enhance user engagement and improve the overall effectiveness of image retrieval in conversational systems. Our study seeks to address the following research questions:

RQ3: Can augmenting media views with text object labels be effective in improving the user search experience in image retrieval?

RQ3 consists of the following subquestions:

1. Can augmented reality highlighted objects or textual labels in the search results make the user experience more convenient and efficient?

2. Which multimedia representation factors are important for a better user experience?

6.2 Experimental methodology

In this section, we provide a detailed description of the user study conducted for this investigation. The study emphasizes the advantages and limitations of visual labeling functionality when applied to image search scenarios. Specifically, it aims to gain deeper insights into the search behaviors of non-specialist users, whose techniques for using search engines are shaped by personal, informal experiences rather than formal training. Understanding these behaviours is crucial for designing intuitive systems that cater to a broader audience.

We begin by outlining the experimental framework designed for the image search application, detailing how the prototype was developed to incorporate multimodal features and augmented labeling. Following this, we describe the methodology employed to conduct the study, including the recruitment of participants, the tasks assigned to them, and the procedures followed to gather qualitative and quantitative data. This structured approach ensures a comprehensive evaluation of user interactions with the system, highlighting key areas for improvement and providing actionable insights for future developments in conversational search systems. The study was performed online, and participants used a setup of system deployed in

the Amazon Cloud.

6.2.1 Experimental framework

In this section, we present a detailed overview of a prototype conversational image search application that incorporates augmented labeling features. Our experimental study was conducted using the Flickr30k image dataset [31]. For the design of search tasks and object-level annotation, we used the MS COCO dataset [50], which provided a diverse and semantically rich collection of annotated visual content. To enhance visual interaction, objects within the images were automatically detected using the YOLOv8 model [89]. The model assigns a confidence score to each detected object, and in our implementation, only objects with a confidence level exceeding an empirically determined threshold of 0.4 were selected and displayed. This filtering ensures that the visual labels presented during search sessions are both relevant and reliable, contributing to a clearer and more effective user experience in the augmented interface.

Dialogue Flow

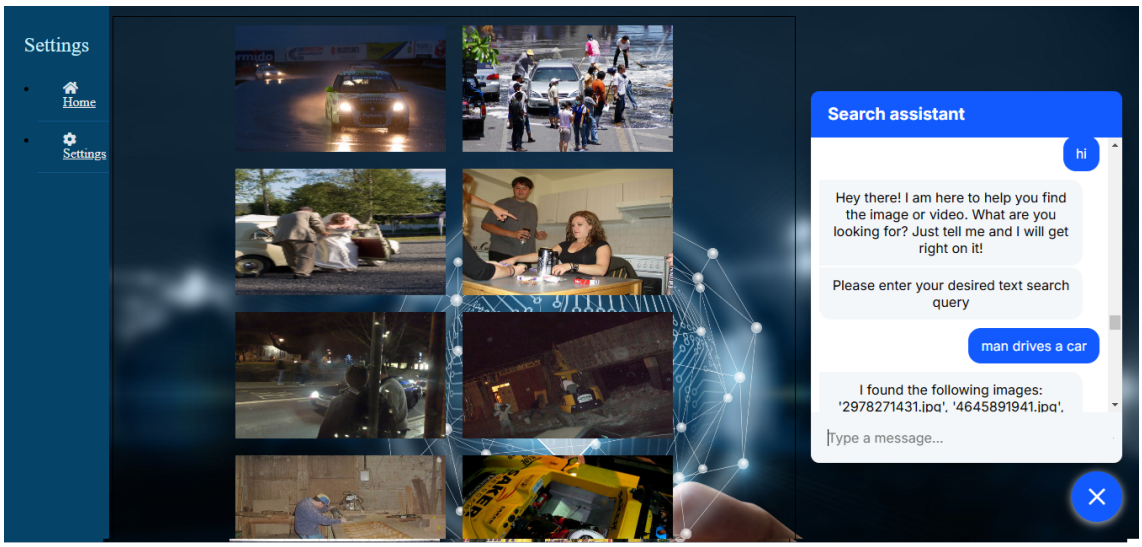

Figure 6.1: The interface of conversational search system

The search scenario starts as usual, from the submitting of text search query and obtaining initial search results, as demonstrated in Figure 6.1. When users enable the visual labeling mode, selecting the corresponding option in the search dialogue, these detected objects—including text labels and bounding boxes—are overlaid on the images in the multimedia gallery, enhancing the search experience and facilitating more precise interactions, as demonstrated in Figure 6.2.

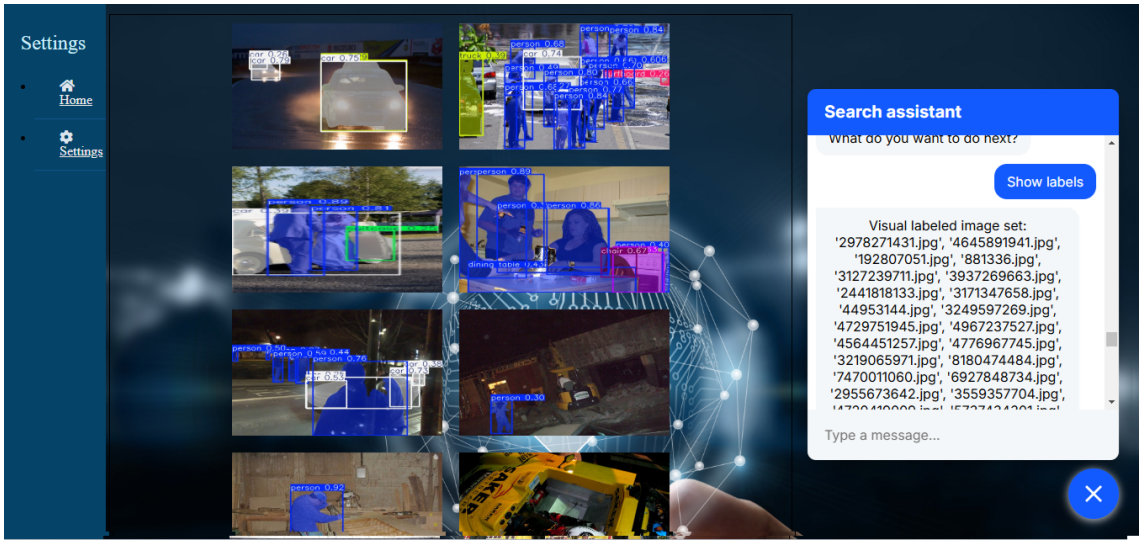

Figure 6.2: The interface of the conversational search system with added visual labels

If the initial search results appear too extensive or overwhelming, users can narrow the output by entering a specific hashtag into the search dialogue (e.g., "#truck"). Doing so filters the displayed results, retaining only images containing the specified labeled object, thus efficiently streamlining the search process, as demonstrated in Figure 6.3. Moreover, if this filtering approach does not fully satisfy the user's requirements, the system allows the user to revert to the previous interaction step and select an alternative tag for further refinement. Alternatively, they can restart their search entirely by submitting a new query. This iterative and flexible approach enables users to experiment with diverse filtering strategies, fostering a dynamic, adaptable, and user-driven search experience.

Figure 6.3: Updated image search output

If the refined output still does not meet the user's needs, they have the flexibility to return to the previous step and select a different tag to try an alternative filtering approach. Alternatively, users can restart the search altogether by typing a new query in the search dialogue. This iterative process ensures that users can experiment with different strategies to locate their desired images, making the search experience dynamic and adaptable.

6.2.2 Search task design

In experimental session, participants are required to complete designated multimedia search tasks and corresponding questionnaires. The session begins with a training task to familiarize participants with the search application and task requirements. Following the training, participants perform two search tasks. During these tasks, their search activities will be systematically logged for future analysis.

6.2.3 Experimental procedure

In this section, we outline the experimental procedure for comparing conversational image search with and without augmented labelling. The goal of this study is to gain insights into how augmented labelling can enhance user experience during the

image search process.

The study followed the previously described search task instructions to evaluate the time required to complete each section, analyze participant behavior, and collect feedback on their experiences with the conversational search approach for image retrieval. Participants first completed a pre-search survey to provide baseline information about their search habits and preferences. They then performed search tasks using two modes of the conversational image search system: one with augmented labelling and one without. Finally, participants completed a post-search survey to provide detailed feedback and compare their experiences across the two modes.

This experimental design enabled a thorough assessment of the impact of augmented labelling on user engagement, efficiency, and satisfaction during the image search process.

Pilot study

Prior to conducting our main study, a pilot study with two PhD students with STEM background was conducted using image search tasks to see how long it took them to complete the sections of the study, gain insights into the likely behaviour of participants, and to generally debug and refine the experimental setup.

Feedback from participants:

- Feature to Toggle Labels On/Off: Participants suggested adding an option to enable or disable labels on demand. This would give users greater control over the interface, allowing them to focus on either the visual content or additional metadata as needed.

- Additional Visual Labels Filter: Users proposed including several visual labels to provide additional flexibility in refining their search results, making the filtering process more comprehensive.

- Image Retrieval Issues with Labelling: Participants noted that sometimes images could not be located using the labelling feature, highlighting a potential

issue with how the system processes or matches labelled content.

- Clarification on Searching with Multiple Tags: It was unclear to some participants how to perform searches using multiple tags simultaneously or how to repeat a tag in the query effectively. Improved instructions or examples were suggested to address this confusion.

The instruction sheet was revised to improve clarity and ensure participants could easily follow the search task steps. Issues related to object detection were identified as limitations of the YOLOv8 model and could not be resolved within the scope of this project. Additionally, other suggestions from participants were not implemented due to time constraints.

6.3 User study

The main study involved 17 participants, all with STEM backgrounds, including MSc students, PhD candidates, and postdoctoral researchers. The participant group comprised 8 males and 9 females, with ages ranging from 23 to 41. All participants reported previous experience using image search engines, predominantly Google Images, for over 10 years. Additionally, 11 participants had experience using conversational assistants for general search tasks, which they described as user-friendly but not highly informative and 4 of them are using conversational search tools regularly. The average level of interest in the topic of conversational search, rated on a scale from 1 to 7, was 5.31, with 1 being the minimum and 7 the maximum. As demonstrated in previous studies this participant sample was expected to be sufficient to make meaningful insights and conclusions [3].

At the start of the experiment, participants were asked to complete a few initial steps to ensure informed and organized participation. First, they filled out consent forms to confirm their understanding and agreement with the study's procedures. Following this, participants completed a basic demographic pre-search survey to provide background information relevant to the study. Then, they reviewed a writ-

ten description of the search tasks, which gives them a clear understanding of the objectives.

Once the preliminary steps were completed, the investigator introduced both modes of search frameworks to the participants and demonstrated the core features. This demonstration was essential to help users understand how to interact with the tools effectively. Each search task sheet contained two distinct images, each representing a separate search scenario. For each scenario, participants described the content of the target image in their own words, generating a natural language text query based on what they observe.

The selected user scenarios were designed to reflect those commonly encountered by content creators, such as writers, illustrators looking for image references and individuals preparing presentations, and others who seek appropriate images to complement their work. These scenarios reflect typical use cases in the context of image search, making them highly relevant for our study.

Additionally, these scenarios gave users some flexibility in interpreting and constructing queries. This freedom was essential for exploring user behaviour in image search environments, as it allowed us to understand how users approach the search process, formulate queries, and navigate through the results.

The task was considered complete once the user had successfully found images that closely matched the content of the provided task images for both scenarios or user could declare that there are no relevant images in the search output. This process helped us collect valuable feedback on the framework's performance, highlighting how well they handle natural language queries and provide relevant results based on user input.

Since it is important to ensure that there are no sequence or order effects in an experiment of this type to ensure that results are not affected by potential sequencing effects. Sequencing effects can increase the chance that results are due to experimental conditions rather than genuine differences in user behaviour per task resulting from the experimental condition. To avoid any biases on the experimental

setup the search tasks and framework modes were rotated and counter balanced, according to Latin square design.

6.3.1 User experience questionnaire results

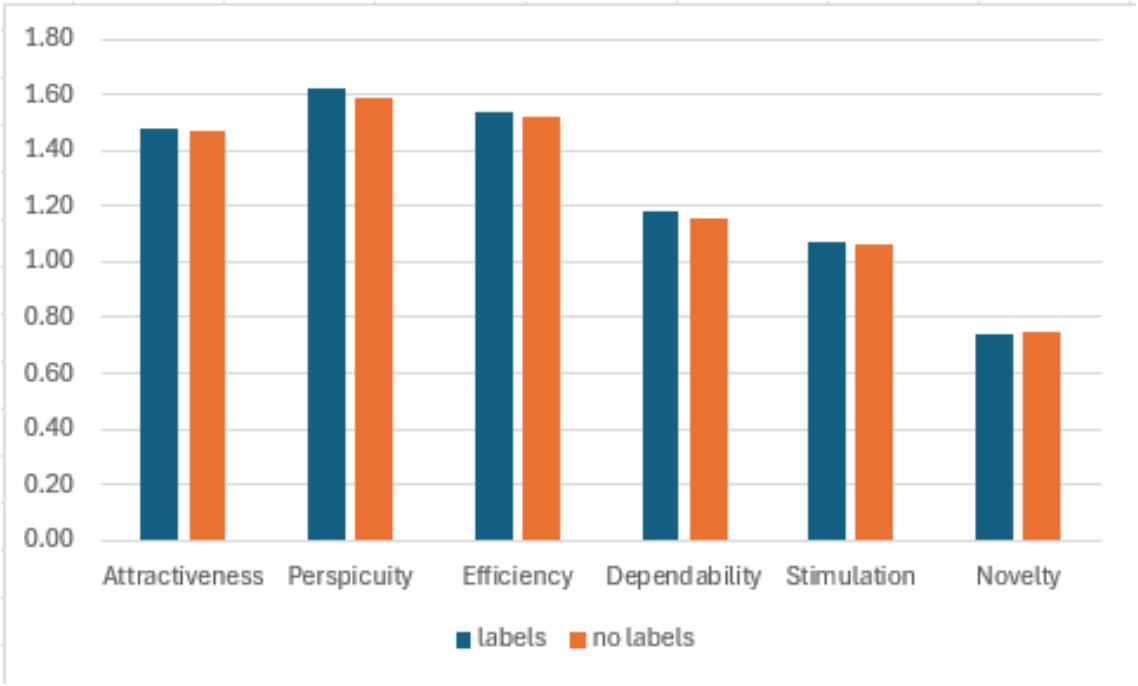

Figure 6.4: The comparative results of UEQ for conversational framework with and without labels

After completing all search tasks, users were asked to evaluate their experiences using post-search online questionnaires. Initially, participants assessed their experience with the conversational framework with and without visual labelling.

Using the UEQ metrics, both frameworks were evaluated by users. The results of comparison of user experience of usage of conversational framework with and without augmented labels, shown in Figure 6.4, indicate that the visual labelling mode outperformed the conversational interface in several key areas. It scored higher in Attractiveness (overall appeal), Perspicuity, Efficiency and Dependability (pragmatic qualities), demonstrating their ease of understanding and usability. Stimulation and Novelty (hedonic qualities) are relatively high for both conversational search modes.

When evaluated using the UEQ metrics, both framework modes were assessed by our users. The results, presented in Table 6.1 reveal that the conversational

Scale	With labels	Without labels
Attractiveness	1.50	1.44
Perspicuity	1.65	1.63
Efficiency	1.55	1.50
Dependability	1.21	1.15
Stimulation	1.11	1.03
Novelty	0.77	0.72

Table 6.1: Detailed results of comparison conversational interface with and without visual labelling the UEQ benchmark

framework with augmented labels scored higher than the non-labeled mode in terms of Attractiveness (overall characteristics), Efficiency and Dependability (pragmatic qualities), and Stimulation and Novelty (hedonic qualities). The Perspicuity (pragmatic quality) was relatively high for both frameworks.

These findings suggest that while participants considered both modes of conversational search effective, intuitive, and user-friendly, the inclusion of augmented visual labels significantly enhanced the perceived clarity, ease of interaction, and overall user satisfaction. The augmented labelling approach provided participants with additional contextual guidance, simplifying the task of locating relevant images and making the search process noticeably smoother and more efficient.

6.3.2 Search strategies and behaviour

In this section, we discuss user behaviour during the search session. Four participants explicitly reported that they were able to locate the relevant images only when using the augmented labeling functionality. These participants noted that the visual labels helped them quickly identify images containing specific content, especially in cases involving sports-related queries. In contrast, they were unable to find the same images using the non-labelled version of the conversational search framework.

This performance gain can be attributed to the effectiveness of the YOLOv8 object detection model, which accurately identified and tagged visual elements such as sports equipment and scenes. By surfacing this visual metadata as clickable or searchable labels, the system provided users with a clear and intuitive way to filter

the image search results. The availability of clearly defined object labels not only reduced the need for query reformulation but also streamlined the entire search process, particularly in visually complex domains, such as sports events.

User's activity	Conversational with labels	Conversational without labels
Number of interactions	19.3	22.15
Number of queries	2.5	3
Restart	2.5	2.7
Reduce the search output	N/A	1.13
Rewrite the query	0.1	0
Ask about another image	N/A	1.7
Visual labels usage	4.5	N/A
Return to previous results	0.6	1.1
Time for the search task	22.5 (min)	21.2 (min)

Table 6.2: Summary Table of Average Results

The quantitative results are shown in Table 6.2, which outlines the average number of interactions and time spent per search task per user. These results suggest that participants spent slightly more time and performed additional actions when using the conversational search mode without visual labels. This finding may indicate that incorporating visual labels simplifies the image search process, as it provides users with clear, identifiable elements for filtering. By allowing participants to focus on specific detected objects, the visual labels mode appears to make locating and selecting relevant images more straightforward and efficient.

We aimed to analyze user behavior while interacting with search tool, focusing on the types of interactions they performed and the underlying reasons for their behavior. To facilitate this analysis, we categorized search behavior into two distinct categories, enabling a structured approach to understanding user actions and motivations:

- **User type 1:** The user enters a full sentence as a text search query and then refines the search output through multimodal clarification questions or visual labels, depending on the search task. By the end of the task, the user selects one or more images as relevant.

- **User type 2:** The user inputs a single word as a text search query, focusing on the most prominent object in the image. Similar to User Type 1, the user leverages conversational search features to refine results, and ultimately chooses one or more images as relevant.

- There is also the possibility that a user may issue one or more queries without selecting any retrieved images. This could occur if the user finds no relevant items, cannot identify suitable results, or feels their information need is already met by viewing the images without making a selection. In this study, only one participant failed to find relevant images in one of the search tasks for the conversational search mode without visual labels.

6.3.3 Comparison with previous study

In our prior work, we developed and conducted a user study aimed at investigating how conversational search influences the user experience in the context of image retrieval. The results from this initial study indicated that users perceived the conversational search approach as both intuitive and effective.

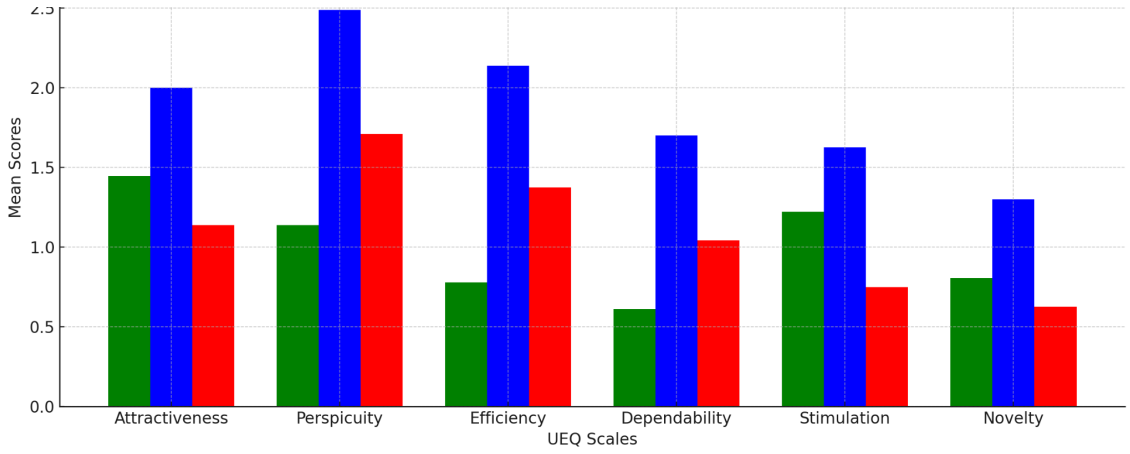

Figure 6.5: The comparison of various user groups results

A number of the participants in the current study participated in the earlier study. The results shown in Figure 6.5 show that interestingly experienced users who participated in our previous study rated our conversational search framework more

positively then those using it for the first time—particularly across the pragmatic quality dimensions of Perspicuity, Efficiency, and Dependability. This suggests that, in real-world scenarios where users engage with the tool on a regular basis, the system becomes increasingly intuitive and easier to use. Familiarity with the interface and search dialogue appears to enhance usability over time, reinforcing the long-term value and clarity of the proposed solution.

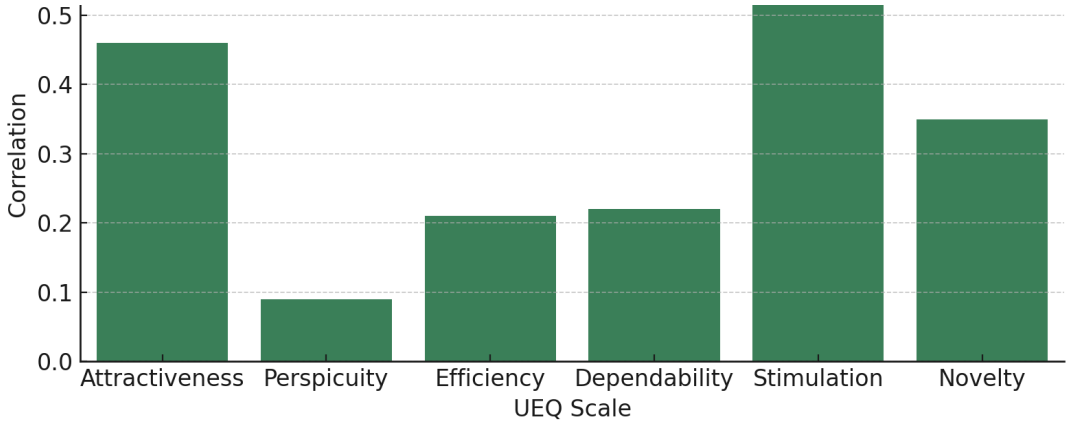

Figure 6.6: The correlation between interest in CS and UEQ scores for the AR study

Users who expressed a higher level of interest in conversational search systems evaluated our framework as more attractive and novel, as it shown on the Figure 6.6. This indicates that individuals more engaged with or curious about search technologies perceive the system as both engaging and innovative, suggesting that the framework resonates particularly well with those invested in emerging search paradigms. This extended behavioral analysis provides a more nuanced understanding of user interaction strategies.

6.4 Summary analysis

Our evaluation reveals that the conversational assistant can be successfully used for image retrieval with improved user experience when incorporating augmented labeling. We identified several key components crucial to fostering successful user adoption of this technology:

- Interactive engagement: Some users spent additional time exploring the var-

ious interactive features of the conversational interface. While this extended the duration of the search process, it significantly enhanced user engagement. The added complexity of navigating interactive elements contributed to a more immersive and user-friendly experience.

- Effectiveness of AR labels: Nearly all participants reported that the conversational search interface with augmented reality labels was intuitive and effective. The presence of visual labels made the process clearer and more transparent. Several users noted that they were able to identify the most relevant images using only the AR-labeled mode.

- Precision in object detection: When the system accurately detected objects within images, users reported an improvement in the relevance of search results. The ability to filter image outputs based on precisely labeled content was seen as a important contributor to the overall effectiveness of the search experience.

6.5 Concluding Remarks

This user study compared conversational search framework with particular emphasis on integrating visual labeling features within the conversational approach. Our findings indicate that the conversational search mode enhanced with visual labels was consistently perceived as clear, intuitive, and efficient. Users found that the presence of visual labels makes easier to locate and refine their desired results.

By incorporating multimodal conversational assistance alongside visual labels, we successfully simplified the image search process. The structured, interactive dialogue guided users in refining their queries and exploring various filtering options, while the visual labels provided immediate insights into the content of retrieved images. This combination fostered a more flexible interaction, ultimately streamlining the retrieval process and facilitating better user outcomes.

The study has some limitations which we will seek to address in future work.

The sample size of the student is relatively small, which limits the significance which can associated with the results. The experimental results our for very similar backgrounds, results from a broader set of users would be very interesting. Inaccuracies in object recognition with consequential errors in visual labels impacted negatively on some of the augmented label results. Some users requested a more flexible interface formed using a hybrid of the labeled and unlebeled settings.

Overall thought, the results from our current study underscore the substantial potential of integrating visual labeling into conversational search frameworks for image retrieval tasks. Building on these insights, future research may examine how these improvements can be extended to video retrieval scenarios.

Chapter 7

Conclusions and Future Directions

This chapter gives a summary of our work reported in this thesis, and revisits the research questions introduced in Chapter 3. Our analysis describes the conclusions that can be made and limitations of our investigation, and suggests directions for future work.

7.1 Key findings

In this PhD dissertation, we present our research on integrating conversational processes into MIR problems. This work investigates the challenges and opportunities associated with conversational search, analyzes user search behavior in both traditional IR and conversational search settings, and explores the development of an implicit evaluation framework for conversational search interfaces. The key contributions of this research are outlined below:

7.1.1 Exploring the Challenges in Current Conversational Systems

We investigated the challenges of conversational search in multimedia information retrieval, focusing on the issues users encounter while performing search tasks on such platforms. This investigation identified key problems and informed the design

and evaluation of an enhanced application to support exploratory search on these systems.

Additionally, we introduced a novel MIR framework that addresses various aspects of search, including image search, video search, and image search with augmented labels. To validate the framework, we employed statistical tests such as MANOVA to test hypotheses and compare the performance of alternative search interfaces, providing a robust evaluation of their effectiveness and usability.

7.1.2 Conversational Search Framework for Images and Videos

We have developed a web-based application designed to facilitate conversational search for multimedia content. This application serves as a platform to explore and implement a mixed-initiative interaction model through a dialogue-based framework, enabling dynamic collaboration between the system and users during search tasks.

Our research focuses on enhancing the interface to better support these interactions, with a specific emphasis on improving usability and functionality. Through user studies, we evaluated the framework's ability to improve the user experience while also examining the cognitive load imposed by the system. These studies aim to ensure that the interface is not only effective but also intuitive and user-friendly.

The current application supports mixed modalities, allowing users to interact with both images and text responses to fulfill their information needs. Dialogue-based interactions play a critical role in clarifying ambiguous queries and refining search processes. The integration of an interactive multimedia gallery, augmented with augmented reality features, provides users with a more immersive and intuitive way to engage with search results, offering options for filtering, refining, and exploring the retrieved content.

Moreover, this work stands out as one of the few efforts focused on conversational search for videos in open-domain archives. By addressing the challenges of video retrieval within conversational frameworks, our application pushes the boundaries of conversational search systems, catering to a wider range of multimedia content

and enhancing the overall search experience for users.

7.2 Research Questions Addressed in this Thesis

We investigated the research questions introduced in Chapter 1. The research questions examined user search behaviour, user interaction behaviour with conversational search in MIR, challenges and possible areas to support users in CS, investigating and developing dialogue strategies and the evaluation of CS. In this section we revisit these questions and summarize our findings from this thesis.

7.2.1 RQ1:User experience in conversational MIR: How does user experience in MIR compare between a standard MIR system and an equivalent one integrating a conversational search agent?

RQ1 focuses on investigation of conversational search effect on image information retrieval and user experience. Our first investigation studied user search behaviour in the conversational and conventional search systems. The investigation of this research question is explained in Chapter 5. We explored a conventional search system, analysed user search behaviour, and found potential opportunities for inclusion of conversational support in the search process.

How can the multimodal conversational search system could be compared with a conventional search system?

This user study compared conversational and non-conversational search frameworks to evaluate their impact on user experience during the image search process. The study focused on key aspects such as usability, efficiency, and the overall satisfaction derived from using each framework.

The results indicated that users perceived the conversational search approach as both helpful and effective. Participants highlighted that the structured dialogue

and interactive features provided by the conversational framework enabled them to refine their queries more easily and identify relevant results more quickly. Additionally, the conversational framework offered greater flexibility and personalization compared to the non-conversational approach. Features such as query expansion, clarification prompts, and visual labels empowered users to dynamically adjust their search strategies, resulting in a smoother and more engaging experience.

What aspects of using multimodal features in a search dialogue can be used effectively in multimedia information retrieval?

Our experiment demonstrated that the multimodal, dialogue-based search assistant significantly enhances the overall user experience in image search tasks, making it more intuitive and engaging. The assistant's structured and interactive dialogue facilitated efficient navigation of the search process, helping users clarify their information needs and refine their queries. This not only reduced cognitive load but also made the search process more accessible and user-friendly, encouraging seamless interaction with the system.

7.2.2 RQ2: Can clarifying questions be used effectively to resolve ambiguity and improve search effectiveness in conversational MIR?

RQ2 explores the impact of conversational search on video information retrieval and its influence on user experience. This investigation examines user search behavior in both conversational and conventional search systems.

The research addressing this question is detailed in Chapters 4 and 5, where we analyzed the conventional search system to understand user interactions and behaviors. Through this analysis, we identified opportunities to incorporate conversational support into the search process, enhancing the effectiveness and usability of video retrieval systems.

What are the opportunities and challenges for embedding clarifying questions into the conversational MIR framework?

Our experiment demonstrated that the multimodal dialogue-based search assistant significantly improved the overall user experience. Participants found the system to be intuitive, with the conversational interface offering a seamless way to engage with the search process. The structured and interactive dialogue enabled users to clarify ambiguous queries, refine their search criteria, and explore alternative search paths in a straightforward manner. This dynamic interaction allowed users to focus on their goals without being overwhelmed by technical details, making the search process both efficient and user-friendly.

Can multimodal clarification features advantage affect the user's search result preference and the user's perceived workload?

Our findings underline the potential of conversational search to streamline multimedia retrieval tasks, offering a more flexible and personalized experience. We extended the capabilities of our multimodal conversational search assistant to video retrieval. This involves integrating more sophisticated conversational features to support the growing demand for effective video search, further advancing our approach to multimodal search interfaces and expanding the scope of our framework's application.

7.2.3 RQ3: Can augmenting media views with text object labels be used to improve the conversational search process in MIR?

RQ3 investigates the impact of conversational search with augmented labelling on image information retrieval and its influence on user experience. This research focuses on examining user search behaviour within a conversational search framework enriched with augmented labelling functionality.

The findings addressing this research question are detailed in Chapter 6, where we analyzed user interactions and behaviors within both conventional and conversational search systems. By comparing the two approaches, we identified key opportunities to incorporate conversational support and augmented labelling into the search process. These enhancements were found to improve the effectiveness, efficiency, and usability of image retrieval systems, providing users with a more intuitive and engaging search experience.

Can augmented reality highlighted objects or textual labels in the search results make the user experience more convenient and efficient?

Integrating visual labelling elements—such as highlighted objects or textual labels—into search results can indeed make the image search process more convenient and efficient. Our user studies, which involved comparing both conversational search framework for multimedia retrieval tasks, consistently demonstrated that visual labelling and augmented overlays help users identify relevant content more quickly and with less effort. Participants in our studies noted the clarity, intuitive nature, and effectiveness of the interface when it included visual labeling features. The improved visibility and interactivity resulted in higher satisfaction, more streamlined search sessions, and an overall more enjoyable user experience.

Which multimedia representation factors are important for a better user experience?

An essential factor in effective multimedia representation is the accuracy of object detection within an image. When objects are clearly and correctly identified, the search process becomes more seamless, allowing users to easily locate the content they seek. In contrast, inaccurate detection may lead to the exclusion of potentially relevant images from the results, even if they contain valuable information.

Equally important is the visual presentation of these labels. Factors such as the number of displayed labels, their transparency, and the flexibility to enable or disable

labelling options were highlighted in both pilot and main user studies. Implementing this feedback can significantly enhance the interface's usability, ultimately improving the overall user experience.

7.2.4 Limitations and Opportunities

In this section, we illustrate the limitations of the work described in this thesis, and examine opportunities for future work:

- Sample Size: The relatively small sample size used in the experiment limits the statistical power of the data collected. This increases the margin of error and reduces the generalizability of the findings, making it difficult to draw robust conclusions about broader user populations.

- Sample Profile: The participant pool may lack diversity, with many individuals sharing similar characteristics such as age, educational background, and current academic status. This homogeneity may not accurately reflect the behavior and preferences of a wider population, potentially biasing the results and limiting their applicability to more diverse user groups.

- Object Detection Errors: Occasionally, the object detection model may misidentify or overlook certain objects, thereby reducing the overall accuracy of the image search. Future integration of more advanced object detection techniques and refined training methodologies can help mitigate these errors and improve the precision of retrieval results.

- Data Collection Process: The presence of social desirability bias in the questionnaire responses is another potential limitation. Participants may underreport negative experiences and over-report positive experiences with the conversational assistant in an effort to please the researcher or present themselves favorably. This bias could affect the reliability of the feedback collected and skew the evaluation of the system.

- User interactions: Users have to do more actions in the conversational search interface, especially during the first attempt of usage, because the conversational interface contains more functions than the traditional interface.

- User experience: Some users already participated in the previous comparative study, so that experience could make an effect to the current results.

- Data limitations: The size of the image and video archives used in the study is relatively limited, which can occasionally lead to inconveniences during the search process. A smaller dataset may restrict the system's ability to retrieve highly diverse or specific results, potentially impacting the user experience and the overall evaluation of the system's effectiveness. Expanding the dataset in future work would help address these limitations and provide a more comprehensive environment.

7.2.5 Future Directions

There are a number of potential directions for further work arising from this thesis.

- Enhanced Conversational Context and Personalized Recommendations: Develop deeper integration of conversational context to better understand user intent over extended interactions. This includes tailoring recommendations based on individual user preferences, past searches, and real-time dialogue analysis for a more personalized and intuitive search experience.

- Advancements in Augmented Labeling:

  - Implement a cloud of tags visualization to provide users with a comprehensive overview of detected labels, allowing for easier navigation and selection.

  - Organize tags into categories for improved clarity and usability, enabling users to refine searches more efficiently.

  - Extend augmented labeling to video search, incorporating dynamic labels for video frames to facilitate precise retrieval of relevant video segments.

- Integration of Audio Modality for Video Search: Introduce audio search capabilities, allowing users to input or refine queries through spoken commands. Additionally, incorporate audio analysis of video content to support search queries based on audio features, such as dialogue, background sounds, or music, enhancing the multimodal search experience.

These future directions aim to further advance conversational search systems in MIR, making them more versatile, user-friendly, and capable of addressing complex and diverse search needs.

Appendix A

Supplementary materials

Before performing the search tasks, participants were given details of the instructions for their search sessions and an interactive tutorial. Each search task was shared separately in a Google Docs protected folder. Each search task took around 20-30 minutes to complete. Feedback from the study was used to draw conclusions and answer the research questions described earlier. Approval was obtained from the DCU Research Ethics Committee before beginning the data collection in all the user studies conducted during this research.

A.1 Pre-search survey

This focuses only on contentment and contains questions on the searcher's demographic details, background knowledge about the search topic, interest in the search topic, and the searcher's experience of using a conversational system.

While searching, the user had to complete an online questionnaire in a Google form while undertaking their search activities.

Section	Questions
Pre-search survey	User Id Occupation Age Gender (M/F) Contact e-mail (for notifying about next study) For how many years have you been using video search engines? On average, how many web video searches do you make each week? Have you ever searched with the help of conversational systems? (Y/N) If Yes, how's your experience with it? Do you use a conversational search tool regularly? (Y/N) if Yes, how many times per week do you generally use this system? (Answers in numbers) How interested are you in learning more about conversational search topic? (low (1) - high (7))

Table A.1: Pre-search questionnaire

A.1.1 Post-search survey (CUQ)

Section	Questions
Post-search survey	The chatbot's personality was realistic and engaging (low(1)-high(5)) The chatbot seemed too robotic (low(1)-high(5)) The chatbot was welcoming during initial setup (low(1)-high(5)) The chatbot seemed very unfriendly (low(1)-high(5)) The chatbot explained its scope and purpose well (low(1)-high(5)) The chatbot gave no indication as to its purpose (low(1)-high(5)) The chatbot was easy to navigate (low(1)-high(5)) It would be easy to get confused when using the chatbot (low(1)-high(5)) The chatbot understood me well (low(1)-high(5)) The chatbot failed to recognise a lot of my inputs (low(1)-high(5)) Chatbot responses were useful, appropriate and informative (low(1)-high(5)) Chatbot responses were irrelevant (low(1)-high(5)) The chatbot coped well with any errors or mistakes (low(1)-high(5)) The chatbot seemed unable to handle any errors (low(1)-high(5)) The chatbot was very easy to use (low(1)-high(5)) The chatbot was very complex (low(1)-high(5))

Table A.2: Chatbot usability questionnaire

A.2 Post-search survey (UEQ)

Section	Questions
Post-search survey	While using the System your experience is like (annoying - enjoyable) While using the System your experience is like (not understandable - understandable) While using the System your experience is like (creative - dull) While using the System your experience is like (easy to learn - difficult to learn) While using the System your experience is like (valuable - inferior) While using the System your experience is like (boring - exciting) While using the System your experience is like (not interesting - interesting) While using the System your experience is like (unpredictable - predictable) While using the System your experience is like (fast - slow) While using the System your experience is like (inventive - conventional) While using the System your experience is like (obstructive - supportive) While using the System your experience is like (good - bad) While using the System your experience is like (complicated - easy) While using the System your experience is like (unlikable - pleasing) While using the System your experience is like (usual - leading edge) While using the System your experience is like (unpleasant - pleasant) While using the System your experience is like (secure - not secure) While using the System your experience is like (motivating - demotivating) While using the System your experience is like (meets expectations - does not meet expectations) While using the System your experience is like (inefficient - efficient) While using the System your experience is like (clear - confusing) While using the System your experience is like (impractical - practical) While using the System your experience is like (organized - cluttered) While using the System your experience is like (attractive - unattractive) While using the System your experience is like (friendly - unfriendly) While using the System your experience is like (conservative - innovative)

Table A.3: User experience questionnaire

A.3 Additional questions

Section	Questions
Post-search survey	Did you find the relevant images for the search task 1 in conversational interface? Did you find the relevant images for the search task 1 in traditional interface? Did you find the relevant images for the search task 2 in conversational interface? Did you find the relevant images for the search task 2 in traditional interface? Did you find the relevant videos for the search task 1 in conversational interface? Did you find the relevant videos for the search task 1 in traditional interface? Did you find the relevant videos for the search task 2 in conversational interface? Did you find the relevant videos for the search task 2 in traditional interface?

Table A.4: Relevant images and videos questionnaire

A.4 Search task instruction for user study 1

A.4.1 Introduction

In this session you are asked to complete assigned multimedia search tasks and associated questionnaires. You will first complete an example training task to gain familiarity with the search application and the requirements of the tasks. You will then complete a series of test tasks for which your search activities will be recorded.

A.4.2 Image-Search Session Instructions

Please, perform the tasks in the following order:

1. Conversational framework, image 1

2. Non-conversational framework, image 2

3. Conversational framework, image 2

4. Non-conversational framework, image 1

A.4.3 Instructions for conversational framework

To perform the particular image search tasks, please follow the instructions below:

- View the first image of the two images provided
- Formulate a text search query to search for each image: Describe the image you are looking for using natural language instructions in English. For example, you can include colours, objects, scenes, or relevant details
- Submit your query to the search interface: Enter your search query using the provided search interface on the right. Once you submit your query, the search framework will process it and retrieve the closest image based on your description
- Review the search results: Review the search results in the gallery and identify the image that best matches your query
- Rewrite the query: After reviewing the search results, you can filter the search output to reduce the number of images in the results. You can select one of two images from the conversational search dialogue, which is closest to your search query. You can repeat the process several times
- Expand the query: if you did not find any (or just a few) images, you may use the query expansion option to add more images to your search output
- Select expanded query: Framework suggests you two options of rewrited queries. You may select one of them or restart the search and type the new search query.
- Explore other options: If you wish to search for different images or explore alternative search queries, you can start a new search by following the same steps

- Repeat the same sequence of actions for the second image

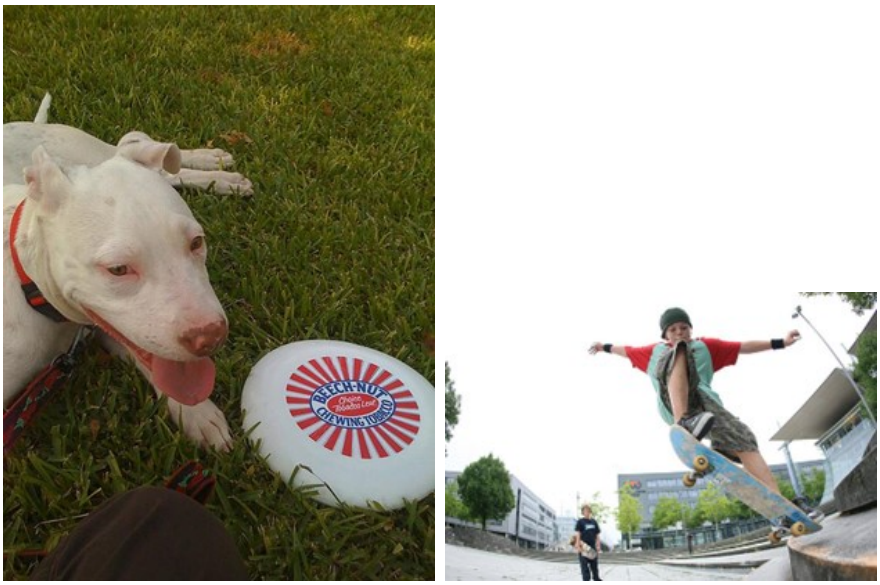

Figure A.1: Sample images for the search task

A.5 Search task instruction for user study 2

A.5.1 Introduction

In this session you are asked to complete assigned multimedia search tasks and associated questionnaires. You will first complete an example training task to gain familiarity with the search application and the requirements of the tasks. You will then complete a series of test tasks for which your search activities will be recorded.

A.5.2 Video-Search Session Instructions

Please, perform the tasks in the following order:

1. Conversational framework, query 1
2. Non-conversational framework, query 2
3. Conversational framework, query 2
4. Non-conversational framework, query 1

A.5.3 Instructions for conversational framework

To perform the particular video search tasks, please follow the instructions below:

- View the First Video Description: Begin by reviewing the provided description for the first video you need to search for.

- Formulate a Text Search Query: Create a text-based search query to search for the video. Describe the video you are looking for using natural language in English. Include relevant details such as objects, actions, scenes, or settings depicted in the video.

- Submit Your Query to the Search Interface: Enter your search query into the search interface provided on the right. Once submitted, the search framework will process your query and retrieve the closest videos matches based on your description.

- Review the Search Results: Examine the videos displayed in the search gallery and identify the ones that best matches your query. Play the video previews if needed to verify the content.

- Rewrite the Query: If the initial results are not satisfactory, you can refine your query to improve the search output. Alternatively, you may select one of the images suggested by the conversational search dialogue to reduce the search results. This process can be repeated multiple times for better accuracy.

- Expand the Query: If you are unable to locate the desired video or retrieve enough relevant results, use the query expansion option. This feature broadens the search scope by generating more video options for your query.

- Select an Expanded Query: The framework will suggest two reformulated query options based on your original input. You may select one of these expanded queries or restart the search and submit a new query altogether.

- Explore Other Options: If you want to search for different videos or try alternative search queries, you can begin a new search by following the same steps outlined above.

- Repeat for the Second Video: Follow the same sequence of actions to perform the search task for the second video.

Samples of video queries:

1. Search query 1: "A cartoon bear is running while a man sits at a bar."

2. Search query 2: "While playing a video game, someone is providing commentary."

A.6 Search task instruction for user study 3

A.6.1 Introduction

In this session you are asked to complete assigned multimedia search tasks and associated questionnaires. You will first complete an example training task to gain familiarity with the search application and the requirements of the tasks. You will then complete a series of test tasks for which your search activities will be recorded.

A.6.2 Image-Search Session Instructions

Please, perform the tasks in the following order:

1. Conversational framework without labelling, image 1

2. Conversational framework with labelling, image 2

3. Conversational framework without labelling, image 2

4. Conversational framework with labelling, image 1

A.6.3 Instructions for conversational framework

To perform the particular image search tasks, please follow the instructions below:

- View the first image of the two images provided
- Formulate a text search query to search for each image: Describe the image you are looking for using natural language instructions in English. For example, you can include colours, objects, scenes, or relevant details
- Submit your query to the search interface: Enter your search query using the provided search interface on the right. Once you submit your query, the search framework will process it and retrieve the closest image based on your description
- Review the search results: Review the search results in the gallery and identify the image that best matches your query
- Rewrite the query: After reviewing the search results, you can filter the search output to reduce the number of images in the results. You can select one of two images from the conversational search dialogue, which is closest to your search query. You can repeat the process several times
- Expand the query: if you did not find any (or just a few) images, you may use the query expansion option to add more images to your search output
- Select expanded query: Framework suggests you two options of rewrited queries. You may select one of them or restart the search and type the new search query.
- Explore other options: If you wish to search for different images or explore alternative search queries, you can start a new search by following the same steps
- Repeat the same sequence of actions for the second image

Instructions for Using the Visual Labels Mode:

- Activate the Visual Labels Mode: After performing the image search, activate the visual labels mode by pressing the "Show Labels" button in the search dialogue.
- Filter Search Results Using Tags: To reduce the search output, enter a tag (e.g., "#truck") in the search dialogue. Images that do not contain the specified object will be automatically removed from the search results.
- Review the updated results: If you found the results unsatisfactory you may return to previous step and try again or you may restart the search from the beginning.

Figure A.2: Sample images for the search task 3

Appendix B

List of Publications

- Anastasia Potyagalova, Gareth J.F. Jones. DCU ADAPT at TRECVID 2021: Video Summarization-Keeping It Simple. TRECVID 2021 Workshop, 2021.
- Anastasia Potyagalova, Gareth J.F. Jones. DCU ADAPT at TRECVID 2022: Deep Video Understanding challenge. TRECVID 2022 Workshop, 2022.
- Anastasia Potyagalova. Conversational Search for Multimedia Archives. In European Conference on Information Retrieval, pp (462-467), Springer, 2023
- Anastasia Potyagalova, Gareth J.F. Jones. A Conversational Search Framework for Multimedia Archives. In European Conference on Information Retrieval, pp (241-245), Springer Nature Switzerland, 2024.

Bibliography

[1] Mohammad Aliannejadi et al. “Analysing mixed initiatives and search strategies during conversational search”. In: *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 2021, pp. 16–26.

[2] Mohammad Aliannejadi et al. “Asking clarifying questions in open-domain information-seeking conversations”. In: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2019, pp. 475–484.

[3] Tayfun Alpay et al. “Multimodal video retrieval with CLIP: a user study”. In: *Information Retrieval Journal* 26.1 (2023), p. 6.

[4] Sungeun An et al. “Recipient design for conversational agents: Tailoring agent’s utterance to user’s knowledge”. In: *Proceedings of the 3rd Conference on Conversational User Interfaces*. 2021, pp. 1–5.

[5] Werner Bailer et al. “Improving Query and Assessment Quality in Text-Based Interactive Video Retrieval Evaluation”. In: *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*. ICMR ’23. Thessaloniki, Greece: Association for Computing Machinery, 2023, pp. 597–601. ISBN: 9798400701788. DOI: 10.1145/3591106.3592281. URL: https://doi.org/10.1145/3591106.3592281.

[6] Xigang Bao et al. “MPMRC-MNER: A unified MRC framework for multimodal named entity recognition based multimodal prompt”. In: *Proceedings*

of the 32nd ACM International Conference on Information and Knowledge Management. 2023, pp. 47–56.

[7] Keping Bi, Qingyao Ai, and W Bruce Croft. "Asking clarifying questions based on negative feedback in conversational search". In: *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*. 2021, pp. 157–166.

[8] John Brooke. "SUS: A quick and dirty usability scale". In: *Usability Eval. Ind.* 189 (Nov. 1995).

[9] Tanja Bunk et al. *DIET: Lightweight Language Understanding for Dialogue Systems*. 2020. arXiv: 2004.09936 [cs.CL]. URL: https://arxiv.org/abs/2004.09936.

[10] Francisco Caldeira et al. "Towards Multimodal Search and Visualization of Movies Based on Emotions". In: *ACM International Conference on Interactive Media Experiences*. IMX '22. Aveiro, JB, Portugal: Association for Computing Machinery, 2022, pp. 349–356. ISBN: 9781450392129. DOI: 10.1145/3505284.3532987. URL: https://doi.org/10.1145/3505284.3532987.

[11] Jiaxun Cao et al. "DreamVR: curating an interactive exhibition in social VR through an autobiographical design study". In: *Proceedings of the 2023 CHI conference on human factors in computing systems*. 2023, pp. 1–18.

[12] Zhiguo Chen et al. "Joint Searching and Grounding: Multi-Granularity Video Content Retrieval". In: *Proceedings of the 31st ACM International Conference on Multimedia*. 2023, pp. 975–983.

[13] Krishna Choudhari and Vinod K Bhalla. "Video search engine optimization using keyword and feature analysis". In: *Procedia Computer Science* 58 (2015), pp. 691–697.

[14] Yung-Sung Chuang et al. *Expand, Rerank, and Retrieve: Query Reranking for Open-Domain Question Answering*. 2023. arXiv: 2305.17080 [cs.CL].

[15] Benjamin R Cowan et al. “” What can i help you with?” infrequent users’ experiences of intelligent personal assistants”. In: *Proceedings of the 19th international conference on human-computer interaction with mobile devices and services*. 2017, pp. 1–12.

[16] Anita Crescenzi, Diane Kelly, and Leif Azzopardi. “Impacts of time constraints and system delays on user experience”. In: *Proceedings of the 2016 acm on conference on human information interaction and retrieval*. 2016, pp. 141–150.

[17] Yashar Deldjoo, Johanne R Trippas, and Hamed Zamani. “Towards multimodal conversational information seeking”. In: *Proceedings of the 44th International ACM SIGIR conference on research and development in Information Retrieval*. 2021, pp. 1577–1587.

[18] Jianfeng Dong et al. “Partially relevant video retrieval”. In: *Proceedings of the 30th ACM International Conference on Multimedia*. 2022, pp. 246–257.

[19] Pei Dong et al. “Disentangled representations and hierarchical refinement of multi-granularity features for text-to-image synthesis”. In: *Proceedings of the 2022 International Conference on Multimedia Retrieval*. 2022, pp. 268–276.

[20] Philip R Doyle et al. “Mapping perceptions of humanness in intelligent personal assistant interaction”. In: *Proceedings of the 21st international conference on human-computer interaction with mobile devices and services*. 2019, pp. 1–12.

[21] Philip R Doyle et al. “The Partner Modelling Questionnaire: A validated self-report measure of perceptions toward machines as dialogue partners”. In: *arXiv preprint arXiv:2308.07164* (2023).

[22] Erlangga, Yaya Wihardi, and Eki Nugraha. “User Experience Evaluation by Using a User Experience Questionnaire (UEQ) Based on an Artificial Neural Network Approach”. In: *2021 3rd International Conference on Research and*

Academic Community Services (ICRACOS). 2021, pp. 17–22. DOI: 10.1109/ICRACOS53680.2021.9702096.

[23] Alex Falcon et al. "Relevance-based margin for contrastively-trained video retrieval models". In: *Proceedings of the 2022 international conference on multimedia retrieval*. 2022, pp. 146–157.

[24] Rafael Ferreira et al. "TWIZ: The wizard of multimodal conversational-stimulus". In: *arXiv preprint arXiv:2310.02118* (2023).

[25] Mauajama Firdaus, Nidhi Thakur, and Asif Ekbal. "Aspect-aware response generation for multimodal dialogue system". In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 12.2 (2021), pp. 1–33.

[26] Dehong Gao et al. "Fashionbert: Text and image matching with adaptive loss for cross-modal retrieval". In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2020, pp. 2251–2260.

[27] Lei Gao. "Latin squares in experimental design". In: *Michigan State University* (2005).

[28] Noa Garcia. "Temporal aggregation of visual features for large-scale image-to-video retrieval". In: *Proceedings of the 2018 ACM on international conference on multimedia retrieval*. 2018, pp. 489–492.

[29] Rafal Grycuk and Rafal Scherer. "Software framework for fast image retrieval". In: *Proceedings of the 24th International Conference on Methods and Models in Automation and Robotics (MMAR 2019)*. IEEE. 2019, pp. 588–593.

[30] Marti Hearst and Melanie Tory. "Would you like a chart with that? Incorporating visualizations into conversational interfaces". In: *2019 IEEE Visualization Conference (VIS)*. IEEE. 2019, pp. 1–5.

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

[32] Samuel Holmes et al. “Usability testing of a healthcare chatbot: Can we use conventional methods to assess conversational user interfaces?” In: *Proceedings of the 31st European Conference on Cognitive Ergonomics*. 2019, pp. 207–214.

[33] Rolf Jagerman et al. *Query Expansion by Prompting Large Language Models*. 2023. arXiv: 2305.03653 [cs.IR].

[34] Xun Jiang et al. “Faster Video Moment Retrieval with Point-Level Supervision”. In: *Proceedings of the 31st ACM International Conference on Multimedia*. 2023, pp. 1334–1342.

[35] Minnu Helen Joseph and Sri Devi Ravana. “Generation of High-Quality Relevant Judgments through Document Similarity and Document Pooling for the Evaluation of Information Retrieval Systems”. In: *2022 14th International Conference on Software, Knowledge, Information Management and Applications (SKIMA)*. 2022, pp. 261–265. DOI: 10.1109/SKIMA57145.2022.10029459.

[36] Abhishek Kaushik, Billy Jacob, and Pankaj Velavan. “An Exploratory Study on a Reinforcement Learning Prototype for Multimodal Image Retrieval Using a Conversational Search Interface”. In: *Knowledge* 2.1 (2022), pp. 116–138.

[37] Diane Kelly et al. “Methods for evaluating interactive information retrieval systems with users”. In: *Foundations and Trends® in Information Retrieval* 3.1–2 (2009), pp. 1–224.

[38] Diane Kelly, David J Harper, and Brian Landau. “Questionnaire mode effects in interactive information retrieval experiments”. In: *Information processing & management* 44.1 (2008), pp. 122–141.

[39] Diane Kelly et al. “Development and evaluation of search tasks for IIR experiments using a cognitive complexity framework”. In: *Proceedings of the*

2015 international conference on the theory of information retrieval. 2015, pp. 101–110.

[40] Antti Keurulainen et al. "Amortised experimental design and parameter estimation for user models of pointing". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 2023, pp. 1–17.

[41] Omar Shahbaz Khan et al. "Exquisitor at the Video Browser Showdown 2024: Relevance Feedback Meets Conversational Search". In: *International Conference on Multimedia Modeling*. Springer. 2024, pp. 347–355.

[42] Hyounghun Kim et al. "CAISE: Conversational Agent for Image Search and Editing". In: (Feb. 2022).

[43] Hyunchul Kim, Kasper Hornbæk, and Byungjoo Lee. "Quantifying Proactive and Reactive Button Input". In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 2022, pp. 1–18.

[44] Ahmet Baki Kocaballi, Liliana Laranjo, and Enrico Coiera. "Understanding and measuring user experience in conversational interfaces". In: *Interacting with Computers* 31.2 (2019), pp. 192–207.

[45] Guy Laban. "Perceptions of anthropomorphism in a chatbot dialogue: the role of animacy and intelligence". In: *Proceedings of the 9th international conference on human-agent interaction*. 2021, pp. 305–310.

[46] Paula Lauren and Paul Watta. "A conversational user interface for stock analysis". In: *2019 IEEE International Conference on Big Data (Big Data)*. IEEE. 2019, pp. 5298–5305.

[47] Yikang Li, Jenhao Hsiao, and Chiuman Ho. "Videoclip: A cross-attention model for fast video-text retrieval task with image clip". In: *Proceedings of the 2022 International Conference on Multimedia Retrieval*. 2022, pp. 29–33.

[48] Lizi Liao et al. "MMConv: an environment for multimodal conversational search across multiple domains". In: *Proceedings of the 44th International*

ACM SIGIR Conference on Research and Development in Information Retrieval. 2021, pp. 675–684.

[49] Hongpeng Lin et al. "TikTalk: A Video-Based Dialogue Dataset for Multi-Modal Chitchat in Real World". In: *Proceedings of the 31st ACM International Conference on Multimedia*. 2023, pp. 1303–1313.

[50] Tsung-Yi Lin et al. "Microsoft COCO: Common Objects in Context". In: *CoRR* abs/1405.0312 (2014). arXiv: 1405.0312. URL: http://arxiv.org/abs/1405.0312.

[51] Caitlin Lustig, Artie Konrad, and Jed R Brubaker. "Designing for the bittersweet: Improving sensitive experiences with recommender systems". In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 2022, pp. 1–18.

[52] Chenyang Lyu et al. "Dialogue-to-Video Retrieval". In: *European Conference on Information Retrieval*. Springer. 2023, pp. 493–501.

[53] Changyi Ma et al. "Large-scale image retrieval with sparse binary projections". In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2020, pp. 1817–1820.

[54] Zhixin Ma and Chong Wah Ngo. "Interactive video corpus moment retrieval using reinforcement learning". In: *Proceedings of the 30th ACM International Conference on Multimedia*. 2022, pp. 296–306.

[55] Iain Mackie, Shubham Chatterjee, and Jeffrey Dalton. *Generative and Pseudo-Relevant Feedback for Sparse, Dense and Learned Sparse Retrieval*. 2023. arXiv: 2305.07477 [cs.IR].

[56] Iain Mackie et al. *GRM: Generative Relevance Modeling Using Relevance-Aware Sample Estimation for Document Retrieval*. 2023. arXiv: 2306.09938 [cs.IR].

[57] Joao Magalhaes et al. “The Next Generation Multimodal Conversational Search and Recommendation”. In: *Proceedings of the 29th ACM International Conference on Multimedia*. New York, NY, USA: Association for Computing Machinery, 2021, pp. 953–954. ISBN: 9781450386517. URL: https://doi.org/10.1145/3474085.3480025.

[58] Louis Mahon et al. “Knowledge Graph Extraction from Videos”. In: *CoRR* abs/2007.10040 (2020). arXiv: 2007.10040. URL: https://arxiv.org/abs/2007.10040.

[59] Foteini Markatopoulou et al. “Query and keyframe representations for ad-hoc video search”. In: *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*. 2017, pp. 407–411.

[60] Maristella Matera, Francesca Rizzo, and Giovanni Toffetti Carughi. “Web usability: Principles and evaluation methods”. In: *Web engineering* (2006), pp. 143–180.

[61] Michael F McTear. “The rise of the conversational interface: A new kid on the block?” In: *International workshop on future and emerging trends in language technology*. Springer. 2016, pp. 38–49.

[62] Hee-Seung Moon, Antti Oulasvirta, and Byungjoo Lee. “Amortized inference with user simulations”. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 2023, pp. 1–20.

[63] Pan Mu et al. “Little Strokes Fell Great Oaks: Boosting the Hierarchical Features for Multi-exposure Image Fusion”. In: *Proceedings of the 31st ACM International Conference on Multimedia*. 2023, pp. 2985–2993.

[64] Meenaakshi N Munjal and Shaveta Bhatia. “A novel technique for effective image gallery search using content based image retrieval system”. In: *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*. IEEE. 2019, pp. 25–29.

[65] Liqiang Nie et al. “Conversational Image Search”. In: *IEEE Trans. Image Process.* 30 (2021), pp. 7732–7743. DOI: 10.1109/TIP.2021.3108724. URL: https://doi.org/10.1109/TIP.2021.3108724.

[66] Young Hoon Oh, Kyungjin Chung, and Da Young Ju. “Differences in interactions with a conversational agent”. In: *International journal of environmental research and public health* 17.9 (2020), p. 3189.

[67] OpenAI. “GPT-4 Technical Report”. In: *ArXiv* abs/2303.08774 (2023). URL: https://arxiv.org/abs/2303.08774.

[68] Paul Owoicho et al. “Exploiting simulated user feedback for conversational search: Ranking, rewriting, and beyond”. In: *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2023, pp. 632–642.

[69] Jiancheng Pan, Qing Ma, and Cong Bai. “Reducing semantic confusion: Scene-aware aggregation network for remote sensing cross-modal retrieval”. In: *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*. 2023, pp. 398–406.

[70] Sandesh Keshav Pawaskar and SB Chaudhari. “Web image search engine using semantic of Images’s meaning for achieving accuracy”. In: *2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT)*. IEEE. 2016, pp. 99–103.

[71] Maria Pegia et al. “MuseHash: supervised bayesian hashing for multimodal image representation”. In: *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*. 2023, pp. 434–442.

[72] Maxime Portaz et al. “QISS: an open source image similarity search engine”. In: *European Conference on Information Retrieval*. Springer. 2020, pp. 486–490.

[73] Anastasia Potyagalova and Gareth JF Jones. “DCU ADAPT at TRECVID 2021: Video Summarization-Keeping It Simple”. In: ().

[74] Anastasia Potyagalova and Gareth JF Jones. "DCU ADAPT at TRECVID 2022: Deep Video Understanding challenge". In: ().

[75] Emanuele Pucci et al. "Defining Patterns for a Conversational Web". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 2023, pp. 1–17.

[76] Filip Radlinski and Nick Craswell. "A theoretical framework for conversational search". In: *Proceedings of the 2017 conference on conference human information interaction and retrieval*. 2017, pp. 117–126.

[77] Shwetha Rajaram et al. "Eliciting security & privacy-informed sharing techniques for multi-user augmented reality". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 2023, pp. 1–17.

[78] Amon Rapp et al. "Collaborating with a Text-Based Chatbot: An Exploration of Real-World Collaboration Strategies Enacted during Human-Chatbot Interactions". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 2023, pp. 1–17.

[79] Luca Rossetto et al. "vitrivr: A flexible retrieval stack supporting multiple query modes for searching in multimedia collections". In: *Proceedings of the 24th ACM international conference on Multimedia*. 2016, pp. 1183–1186.

[80] Nazmus Saquib, Faria Huq, and Syed Arefinul Haque. "graphiti: Sketch-based Graph Analytics for Images and Videos". In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 2022, pp. 1–15.

[81] R Rani Saritha, Varghese Paul, and P Ganesh Kumar. "Content based image retrieval using deep learning process". In: *Cluster Computing* 22.2 (2019), pp. 4187–4200.

[82] Andrea Schankin et al. "Psychometric properties of the user experience questionnaire (UEQ)". In: *Proceedings of the 2022 Chi conference on human factors in computing systems*. 2022, pp. 1–11.

[83] Martin Schrepp, Jörg Thomaschewski, and Andreas Hinderks. "Construction of a Benchmark for the User Experience Questionnaire (UEQ)". In: *International Journal of Interactive Multimedia and Artificial Intelligence* 4.4 (June 2017), pp. 40–44. ISSN: 1989-1660. DOI: 10.9781/ijimai.2017.445. URL: http://www.ijimai.org/journal/sites/default/files/files/2016/12/ijimai20174_4_5_pdf_94297.pdf.

[84] Kihoon Son, Kyungmin Kim, and Kyung Hoon Hyun. "BIGexplore: Bayesian information gain framework for information exploration". In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 2022, pp. 1–16.

[85] Florian Spiess et al. "A comparison of video browsing performance between desktop and virtual reality interfaces". In: *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*. 2023, pp. 535–539.

[86] Wenliang Tang et al. "OCR-oriented master object for text image captioning". In: *Proceedings of the 2022 International Conference on Multimedia Retrieval*. 2022, pp. 39–43.

[87] Jialin Tian et al. "Zero-shot sketch-based image retrieval with adaptive balanced discriminability and generalizability". In: *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*. 2023, pp. 407–415.

[88] Svitlana Vakulenko et al. "Question rewriting for conversational question answering". In: *Proceedings of the 14th ACM international conference on web search and data mining*. 2021, pp. 355–363.

[89] Rejin Varghese and Sambath M. "YOLOv8: A Novel Object Detection Algorithm with Enhanced Performance and Robustness". In: *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*. 2024, pp. 1–6. DOI: 10.1109/ADICS58448.2024.10533619.

[90] Somin Wadhwa and Hamed Zamani. "Towards System-Initiative Conversational Information Seeking." In: *DESIRES*. 2021, pp. 102–116.

[91] Junke Wang et al. “Omnivl: One foundation model for image-language and video-language tasks”. In: *arXiv preprint arXiv:2209.07526* (2022).

[92] Liang Wang, Nan Yang, and Furu Wei. *Query2doc: Query Expansion with Large Language Models*. 2023. arXiv: 2303.07678 [cs.IR].

[93] Yi Wang et al. “InternVideo: General Video Foundation Models via Generative and Discriminative Learning”. In: *arXiv preprint arXiv:2212.03191* (2022).

[94] Zhenduo Wang et al. “Zero-shot Clarifying Question Generation for Conversational Search”. In: *Proceedings of the ACM Web Conference 2023*. 2023, pp. 3288–3298.

[95] Hao Wei et al. “Conversational Composed Retrieval with Iterative Sequence Refinement”. In: *Proceedings of the 31st ACM International Conference on Multimedia*. 2023, pp. 6390–6399.

[96] Jiaxin Wu, Phuong Anh Nguyen, and Chong-Wah Ngo. “VIREO@ TRECVid 2021 ad-hoc video search”. In: (2021).

[97] Yaxiong Wu, Craig Macdonald, and Iadh Ounis. “Multi-Modal Dialog State Tracking for Interactive Fashion Recommendation”. In: *Proceedings of the 16th ACM Conference on Recommender Systems*. 2022, pp. 124–133.

[98] Fangxiong Xiao et al. “From abstract to details: A generative multimodal fusion framework for recommendation”. In: *Proceedings of the 30th ACM International Conference on Multimedia*. 2022, pp. 258–267.

[99] Ziang Xiao et al. “Inform the uninformed: improving online informed consent reading with an AI-powered Chatbot”. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 2023, pp. 1–17.

[100] Haiyang Xu et al. “mPLUG-2: A Modularized Multi-modal Foundation Model Across Text, Image and Video”. In: *arXiv preprint arXiv:2302.00402* (2023).

[101] Jun Xu et al. “Msr-vtt: A large video description dataset for bridging video and language”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 5288–5296.

[102] Guoxing Yang et al. “Shot Retrieval and Assembly with Text Script for Video Montage Generation”. In: *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*. 2023, pp. 298–306.

[103] Yifei Yuan et al. “Asking Multimodal Clarifying Questions in Mixed-Initiative Conversational Search”. In: *arXiv preprint arXiv:2402.07742* (2024).

[104] Arun Zachariah and Praveen Rao. “Video Retrieval for Everyday Scenes With Common Objects”. In: *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*. 2023, pp. 565–570.

[105] Hamed Zamani et al. “Analyzing and learning from user interactions for search clarification”. In: *Proceedings of the 43rd international acm sigir conference on research and development in information retrieval*. 2020, pp. 1181–1190.

[106] Hamed Zamani et al. “Conversational information seeking”. In: *arXiv preprint arXiv:2201.08808* (2022).

[107] Hamed Zamani et al. “Generating clarifying questions for information retrieval”. In: *Proceedings of the web conference 2020*. 2020, pp. 418–428.

[108] Hamed Zamani et al. “Mimics: A large-scale data collection for search clarification”. In: *Proceedings of the 29th ACM international conference on information & knowledge management*. 2020, pp. 3189–3196.

[109] Peng-Fei Zhang et al. “Machine unlearning for image retrieval: A generative scrubbing approach”. In: *Proceedings of the 30th ACM International Conference on Multimedia*. 2022, pp. 237–245.

[110] Yida Zhao et al. “RUC_AIM3 at TRECVID 2020: Ad-hoc Video Search & Video to Text Description.” In: *TRECVID*. Vol. 1. 2020, p. 2.

[111] Yaoxin Zhuo et al. “Clip4hashing: unsupervised deep hashing for cross-modal video-text retrieval”. In: *Proceedings of the 2022 international conference on multimedia retrieval*. 2022, pp. 158–166.