Toward Facilitating Search in VR With the Assistance of Vision Large Language Models

Chao Liu
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
cliu009@connect.hkust-gz.edu.cn

Clarence Chi San Cheung
The Hong Kong University of Science and Technology
Hong Kong SAR, China
cscheungah@connect.ust.hk

Mingqing Xu
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
mxu097@connect.hkust-gz.edu.cn

Zhongyue Zhang
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
zzhang837@connect.hkust-gz.edu.cn

Mingyang Su
Tsinghua Shenzhen International Graduate School
Shenzhen, China
sumy22@mails.tsinghua.edu.cn

Mingming Fan*
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
The Hong Kong University of Science and Technology
Hong Kong SAR, China
mingmingfan@ust.hk

ABSTRACT

While search is a common need in Virtual Reality (VR) applications, current approaches are cumbersome, often requiring users to type on a mid-air keyboard using controllers in VR or remove VR equipment to search on a computer. We first conducted a literature review and a formative study, identifying six common search needs: knowing about one object, knowing about the object's partial details, knowing objects with environmental context, knowing about interactions with objects, and finding objects within field of view (FOV) and out of FOV in the VR scene. Informed by these needs, we designed technology probes that leveraged recent advances in Vision Large Language Models and conducted a probe-based study with users to elicit feedback. Based on the findings, we derived design principles for VR designers and developers to consider when designing a user-friendly search interface in VR. While prior work about VR search tended to address specific aspects of search, our work contributes design considerations aimed at enhancing the ease of search in VR and potential future directions.

CCS CONCEPTS

• **Human-centered computing → Participatory design; Virtual Reality**.

KEYWORDS

Virtual reality, VR search, participatory design, vision large language model

ACM Reference Format:
Chao Liu, Clarence Chi San Cheung, Mingqing Xu, Zhongyue Zhang, Mingyang Su, and Mingming Fan. 2024. Toward Facilitating Search in VR With the Assistance of Vision Large Language Models. In *Proceedings of ACM Symposium on Virtual Reality Software and Technology (VRST).* ACM, New York, NY, USA, 14 pages. https://doi.org/3641825.3687742

1 INTRODUCTION

With the ongoing development of search engine technology, online information retrieval and photo recognition searches for target objects have become ubiquitous. In the VR field, with the popularization of VR devices and the development of the metaverse concept, an expanding number of VR applications containing a multitude of objects have been developed. Many metaverse open-world platforms allow users to upload and build models, further complicating the VR space with diverse visual and textual information. When users encounter this information, they seek to understand the object in front of them, determine how to interact with it, or locate the target object. This trend has led to searching in VR becoming an increasingly common requirement, akin to the real world.

Searching in VR generally refers to the act of looking for information, objects, or locations within a virtual environment using VR technology [33, 41]. This can involve the use of gestures, gaze, text input, voice commands, or other forms of interaction to navigate and explore the VR environment and find what one is looking for [48]. The goal of searching in VR is to provide an immersive and intuitive experience that enables users to easily find what they need without disrupting the sense of presence and immersiveness. Previous literature has focused on specific aspects of VR search, such as information retrieval or searching for specific targets within a 3D space [8, 41, 50]. However, the works on information retrieval in VR have mainly focused on adapting 2D search strategies within 3D environments, in the form of interacting with a 2D search web window in 3D space, which fails to address the nature and possibilities of VR spaces. Consequently, no comprehensive work presents a VR search interface across different scenarios. VR users typically interact with 3D objects that may convey different meanings from

[*] Corresponding Author

VRST, October 09–11, 2024, Trier, Germany

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/3641825.3687742

various angles and can perform actions that are impossible in the real world. It is essential to develop a VR search interface that takes into account the unique features of VR technology.

In practical applications, the closed nature of individual VR programs and the absence of a cross-application VR search interface limit users to searching for information within the confines of the current application's search functionality. Often, users must exit the application to search—either by switching to a VR browser search engine or removing the headset to use other devices—thereby disrupting the immersive experience and significantly affecting user experience [23].

Recently, the emergence of large language models, such as ChatGPT, has promoted the development of LLM-based conversational search tools. Unlike traditional search engines that only provide a list of websites, LLM-based search tools first understand the query, then absorb information from relevant websites, and finally provide a coherent response that integrates reference information from multiple sources for verification. From a user experience perspective, LLM-based search systems engage in conversational interactions with users, rather than simply presenting a list of search results. They offer more intelligent responses by comprehending the user's language and context [22, 36].

With the emergence of Vision Large Language Models (Vision LLMs) and their ability to process multi-modal data, LLMs can play a key role in the collection of user input, as well as the retrieval of information that is not pre-programmed into the virtual world [42]. However, current research on LLM-driven search systems remains predominantly focused on 2D interfaces, with limited exploration of virtual 3D scenes.

In summary, although searching is a common requirement in virtual reality applications, the current method remains cumbersome. Leveraging the visual understanding capabilities of Vision LLM in the design of a VR search interface can address this problem. However, the current research on user needs in VR search and interface interaction design is not well understood. Based on this, we sought to answer the following two research questions (RQs):

- **RQ 1:** What search difficulties do users meet in the current VR experience, and how do they solve them?
- **RQ 2:** What elements should be considered when designing a VR search interface leveraging Vision LLMs?

In this paper, we first performed the literature review of the work that identifies critical issues in VR search; then, we conducted a formative study (N=10) to gain a deeper understanding of how people search, and the pain points they face. Based on the findings of the formative study, we further examined literature in related fields to identify potential solutions and designed six probes leveraging Vision LLMs. We then conducted a probe-based participatory design workshop (N=11) to explore people's preferences and expectations for VR search interactions, gather feedback, and co-design new solutions. Finally, we present the findings and propose design principles.

The contributions of this research are:

- We present users' VR search practices, their challenges during the VR search activities, and their insights on potential solutions.
- We present a series of design probes leveraging Vision LLMs for each search intention.
- Using the probes, we present VR search interface design principles based on feedback.

2 RELATED WORKS

We first demonstrate that current research on VR search primarily focuses on transferring 2D interfaces into 3D spaces, with limited exploration of how search tasks interact with 3D objects and the direct presentation of information within 3D spaces. Subsequently, we highlight the advantages of Vision LLM, in performing visual understanding tasks and the possibility that they can be applied to VR search tasks. Finally, we illustrate the great potential of LLM-powered search systems in enhancing VR search capabilities.

2.1 Challenges in VR Search

The VR search task consists of two main components: formulating search queries and presenting search engine results. While existing studies have investigated methods of presenting search engine results in VR [32], such as techniques in shopping scenarios and the arrangement of search results [42, 43], there is a notable gap in understanding how to craft search queries within the VR space. Yang et al. [48] highlighted the differences between information-oriented web search and space-oriented VR search in terms of components, interfaces, and data representation. Unlike web search, VR search needs to accommodate 3D models instead of hyperlinks, with data presented in ways that support 3D representations derived from 2D input. Another challenge in VR search, which Yang et al. did not address, is spatial search—locating targets within the VR space. Gao et al. [8] introduced a method using bimanual haptic feedback for spatial search, demonstrating improvements in task completion time, accuracy, and user perception compared to existing methods like spatial audio. However, these explorations have predominantly focused on textual queries. Interacting with diverse objects within the immersive VR environment and leveraging these objects or contextual information for search queries remain underexplored. Therefore, our research aim to explore how to design user-centered search interaction interfaces in VR that address these challenges.

2.2 Vision Large Language Models

Visual Large Language Models like Shikra [5], VisionLLM [40], and Qwen-VL[2] are advanced LLMs that excel in visual tasks such as object recognition, contextual question answering, image captioning, and task instruction [40]. Some studies have attempted to use Vision LLMs for visual search, helping users find information more efficiently within 2D graphical interfaces [45]. Specifically, Vision LLMs have assisted visually impaired individuals in accessing visual information, using semantic segmentation and visual understanding, and providing auditory prompts [49]. However, despite the significant potential of Vision LLMs in these tasks, their application remains largely confined to processing 2D images. In real-world and virtual reality (VR) environments, search tasks often involve recognizing and interacting with 3D objects, yet current Vision LLMs are not equipped to handle 3D objects, presenting a new challenge for these models.

To address these challenges, some research has begun exploring the integration of 3D worlds into large language models. For example, by using 3D point clouds to represent 3D environments and fine-tuning the Vision LLMs, they can perform more complex tasks such as 3D grounding, 3D-assisted dialogue, and navigation [12]. Additionally, other studies have focused on capturing and rendering scenes from virtual worlds to feed into Vision LLMs. For instance, in architectural design, Vision LLMs have been used to help architects predict how users might experience a space in 3D [1]. Although these models still rely on static images as input, their application in 3D environments demonstrates the potential for Vision LLMs to expand into more complex scenarios. In summary, while Vision LLMs have made significant progress in 2D tasks, their application in 3D virtual spaces is still in its early stages. Particularly in the context of VR, conducting searches remains an open question, and we will further explore how to extend the capabilities of Vision LLMs to support 3D object searches within VR environments.

2.3 LLM-Powered Search Systems

The popularization of ChatGPT has led to the emergence of conversational search interfaces [22, 39]. Since the public release of Microsoft Bing Chat and Google Bard in 2023, the number of monthly users of LLM-powered search systems has exceeded hundreds of millions [20, 22, 34]. Unlike traditional search engines, which only provide lists of websites, LLM-powered search engines first understand the query, then assimilate information from relevant websites, and finally provide a coherent response that integrates references from multiple sources for verification. From a user experience perspective, LLM-powered search systems engage in interactive dialogues with users rather than merely presenting a list of search results. They deliver smarter and faster responses by understanding the user's language and contextual content [22, 36]. However, current research on LLM-powered search systems still primarily focuses on 2D interfaces, with limited exploration in virtual 3D scenes, this paper aims to provide design considerations for VR search interfaces through Vision LLMs.

3 METHOD

To address our research questions, we conducted two studies. The first was a formative study to explore people's current search practices and challenges in VR scenes. Then, a probe-based participatory design workshop will be held to explore people's preferences and expectations for VR search interactions. We have summarized our findings as design insights and considerations for future VR search interaction designs.

3.1 Formative Study

3.1.1 Participants. We recruited 10 participants through personal networks and snowball sampling. Inclusion criteria required participants to have experience with VR and be willing to share their experiences. Participants' ages ranged from 18 to 42 (mean = 26.7, median = 24.0, SD = 7.82), with VR experience spanning from 1 year to 8 years (mean = 4.3, median = 4.0, SD = 2.33). Tab. 1 in Appendix A.1 shows more details of the participants. Primarily, participants reported using VR for gaming, development, and social purposes. We conducted a total of 10 individual interviews, each lasting 30 minutes.

3.1.2 Formative Study Process. The semi-structured interviews were conducted both online and in-person. Initially, we provided participants with an overview of the research context and session structure, addressing any questions or concerns they had, and Appendix A.2 provides interview questions. Consent for audio recordings was obtained before each session. Our objective was to delve into participants' current VR search practices, their overall experiences with searching in VR, any challenges they faced, and their viewpoints on potential solutions.

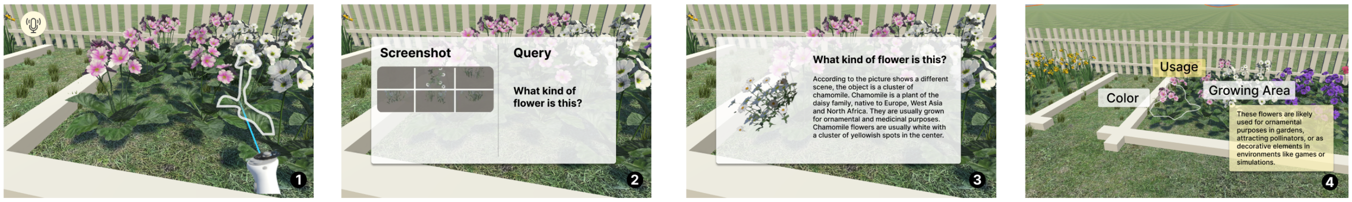
3.1.3 Needs and Challenges. We categorized the probes based on two search needs identified from both the formative study and relevant literature: **Knowing Object (KO)**: these behaviors include reading information relevant to the object to enhance people's understanding of the object [24, 30], and learning about the tutorials or interaction methods of the virtual object [17, 29]. **Finding Object (FO)**: these behaviors involve two task types: in-view searching and out-of-view searching. In-view searching involves objects or interface elements that users can see directly within the current field of view (FOV) without head movement. Out-of-view searching involves targets outside the current FOV that require head or body movement or interaction methods like controllers to be seen [6, 11, 18, 25]. Previous research has shown that two kinds of searching have a significant effect on users' search strategies; in-view targets are typically easier to recognize and select quickly, whereas out-of-view targets may require users to engage in more complex search strategies and spatial memorization, and it may be necessary for the designer to provide additional visual or auditory cues to help users locate these targets [11, 13].

3.2 Probe Design

The probe is developed to address challenges from the formative study, seeks participant feedback in a participatory design workshop. We chose Vision LLM as the technical foundation for two main reasons: it intuitively interacts with users through natural language, and unlike text-based LLMs, it comprehends contextual information from images, enabling us to address a wider range of search challenges in VR. Based on our formative study and previous literature, we designed six VR search probes leveraging Qwen-VL[1] Vision LLM, and we placed these probes in three typical scenarios: the garden, the hospital and the supermarket (Fig. 1- 6), which are often seen in VR entertainment, simulation training and VR shopping. The purpose of displaying our probes to participants was to familiarize them with Vision LLM's capabilities and fundamental VR search interaction scenarios, thereby collecting preliminary user feedback on our designs and stimulating further ideas during the sketching phase. Probes (1)-(4) assist users in knowing objects in VR, while probes (5)-(6) aid in finding objects in VR:

Knowing about one object (KO). As shown in Fig. 1, Probe 1 is designed to help users know about unfamiliar objects in VR. The user needs to select the object and use audio to input the query. Then, the system automatically screenshots six view pictures and sends

[1] https://github.com/QwenLM/Qwen-VL



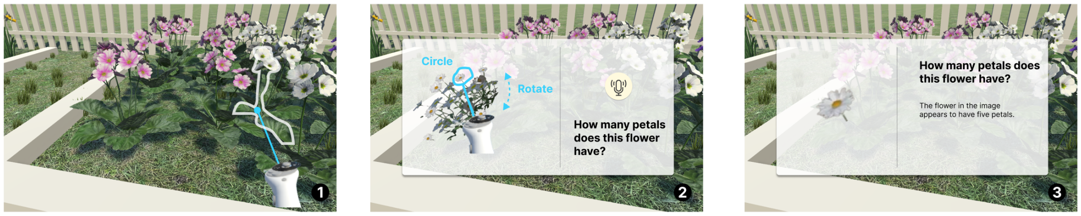
What kind of flower is this?

What kind of flower is this?

According to the picture shows a different scene, the object is a cluster of chamomile. Chamomile is a plant of the daisy family, native to Europe, West Asia and North Africa. They are usually grown for ornamental and medicinal purposes. Chamomile flowers are usually white with a cluster of yellowish spots in the center.

These flowers are likely used for ornamental purposes in gardens, attracting pollinators, or as decorative elements in environments like games or simulations.

Figure 1: Probe 1: (1) Select the object and use audio to input the query; (2) System automatically screenshots six view pictures and sends them to Vision LLM along with the query; (3) Display 1: Vision LLM's text results show along with the 3d object; (4) Display 2: Label Display.

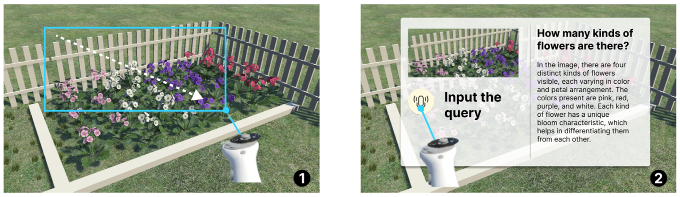


How many petals does this flower have?

How many petals does this flower have?

The flower in the image appears to have five petals.

Figure 2: Probe 2: (1) Select the object; (2) Rotate the model and circle the detail to search; (3) Show the text result.



How many kinds of flowers are there?

In the image, there are four distinct kinds of flowers visible, each varying in color and petal arrangement. The colors present are pink, red, purple, and white. Each kind of flower has a unique bloom characteristic, which helps in differentiating them from each other.

Figure 3: Probe 3: (1) Drag to screenshot the picture of the scene; (2) Use audio to input the query and the system gives the text feedback.

them to Vision LLM along with the query. The views are generated to address the Vision LLM's inability to process 3D objects, so to improve results multiple views of the 3D object are submitted as 2D images. The Vision LLM returns textual results alongside the 3D object or in a label format. The label display method is inspired by previous research [24] that uses label annotations associated with specific objects to provide relevant information to the user. In our probe, label display can reduce the reading burden of Vision LLM's text answer.

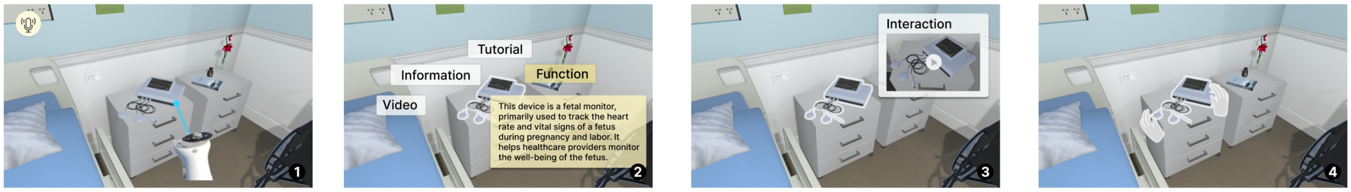
Knowing about the object's partial details (KO). As shown in Fig. 2, in probe 2, users can further investigate an object by rotating it and circling the detail they wish to explore. The Vision LLM then displays text results alongside a detailed picture. This detail search functionality is inspired by the "Circle to Search" feature originally proposed by Google and Samsung [7] for detail searches in 2D images. In our probe, this function can meet the search needs about the model's details.

Knowing objects with environmental context (KO). As shown in Fig. 3, Probe 3 allows users to take screenshots in VR and send the image along with a query to the Vision LLM for results. Designed based on findings from the formative study, this probe enables context-sensitive searches in VR. Unlike Probes 1 and 2, screenshot searches provide more contextual background information.

Knowing about interactions with objects (KO). As illustrated in Fig. 4, users are able to select the equipment and enquire about its usage in Probe 4. The results are displayed in three formats: Label Display, Video Display and VR Animation Display.The label display is inspired by [24], which is similar to Probe 1. The video display method is inspired by the video tutorials commonly used by people in the real world [14, 44]. In this probe, we display the video next to the object to help people learn how to use it. In this probe, we display the video next to the object to help people learn how to use it. The animation display method is inspired by previous research on VR tutorials, which use the animation of objects or virtual hands to show the VR tutorial.[21, 51].

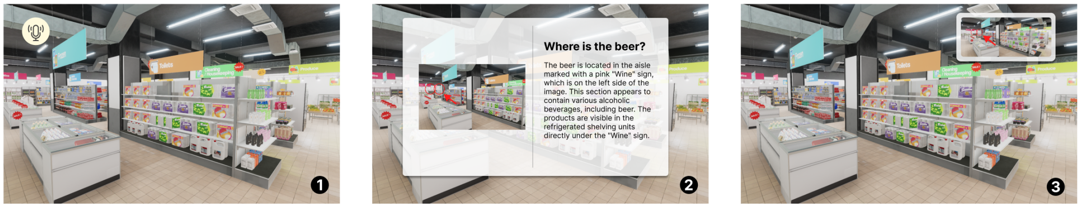
Finding objects within FOV (FO). As shown in Fig. 5, probe 5 is designed based on Vision LLM's ability to comprehend picture content. When users need to locate an object in view, the system automatically captures a screenshot and sends it to the vision LLM, which then displays the text results with picture annotations or shows the annotations in a small window. This probe design is based on the current Vision LLM's ability to return annotation results circled on the picture.

Find objects out of FOV (FO). As shown in Fig. 6, probe 6 is designed to help users find objects that may be outside sight. Vision LLM searches for the location in this probe based on the developer's prior knowledge given to the vision LLM. Then, the results will be displayed using three methods: Penetration highlight, text along with the model, and teleportation. This probe is designed based on the users' need in the formative study, which is that they sometimes want to find an unseen object in the scene.



This device is a fetal monitor, primarily used to track the heart rate and vital signs of a fetus during pregnancy and labor. It helps healthcare providers monitor the well-being of the fetus.

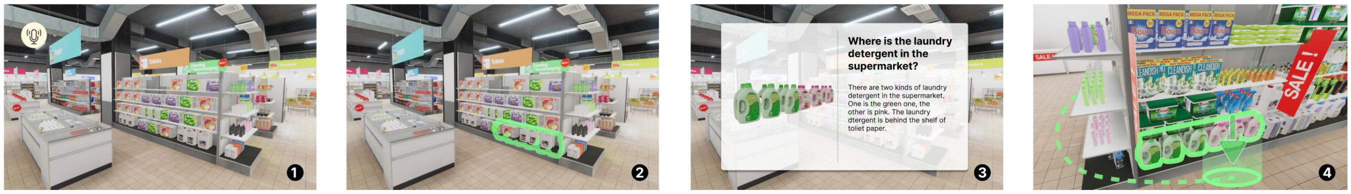
Figure 4: Probe 4: (1) Select the equipment and ask how to use it; (2) Display 1: Label display; (3) Display 2: Show the teaching video; (4) Display 3: Animation Demonstration.



Where is the beer?

The beer is located in the aisle marked with a pink "Wine" sign, which is on the left side of the image. This section appears to contain various alcoholic beverages, including beer. The products are visible in the refrigerated shelving units directly under the "Wine" sign.

Figure 5: Probe 5: (1) Audio ask where is the beer in the supermarket; (2) Display 1: Text + Picture; (3) Display 2: Audio feedback + small window picture.



Where is the laundry detergent in the supermarket?

There are two kinds of laundry detergent in the supermarket. One is the green one, the other is pink. The laundry dtergent is behind the shelf of toliet paper.

Figure 6: Probe 6: (1) Audio asks where is the laundry detergent in the supermarket which is unseen. The Vision LLM will check the prior knowledge of the scene and display the result; (2) Display 1: Penetrate Highlight; (3) Display 2: Model + text description; (4) Display 3: Teleport the user in front of the laundry detergent and highlight the laundry detergent.

3.3 Participatory Design Workshop

This section aims to generate design considerations for VR search interfaces leveraging Vision LLMs. We adopted a probe-based study method instead of a full-functional prototype to involve participants in the initial design phase for qualitative feedback [4]. The participatory design workshop consisted of two phases (Fig. 7). In the first phase, we introduced six probes to participants and gathered their feedback. In the second phase, participants engaged in a co-design workshop to sketch their ideal search interactions and scenarios, which helped generate implications for future technologies [37].

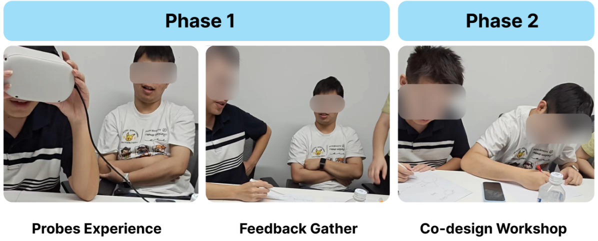


Figure 7: Procedure of the participatory study

3.3.1 Participants. We recruited another 11 participants through personal networks and snowball sampling. The participants ranged in age from 18 to 45 years, with VR experience spanning from 0 to 8 years (mean = 3.45, median = 4.0, SD = 2.39), and Tab. ?? in Appendix B.1 shows more details of the participants. Seven participants had more than three years of VR experience, while two were novices, having used VR only once or twice. Each group of 1-2 participants was then divided into eight groups to participate in the workshop.

3.3.2 Participatory Design Process. Our participatory design process lasted approximately 40 minutes and included two phases. **Phase 1, Probe experience and feedback gathering:** In this part, we first let the participants watch an introduction video about our research. The video introduces our research's goal, Vision LLM's abilities, and some of our probes in the video format. The video is intended to efficiently convey the research background and demonstrate the capabilities of Vision LLM to the participants. After watching the video, we demonstrate and explain the design probes one by one to them. For each probe, we ask for their feedback, what they liked, what they did not, and what improvements they thought could be made. **Phase 2, Co-design session:** we held a co-design session with the participants. We engaged participants by reminding them of the feedback and ideas in the probe experience process of phase 1. This encourages them to propose modifications and new functions that would improve the search experience in VR. Additionally, we invited them to sketch their ideal use scenarios.

4 DATA ANALYSIS

In our research, we conducted semi-structured interviews with participants across two phases: the formative study and the participatory design workshop. For the formative study, we interviewed 10 participants, and for the participatory design workshop, we interviewed 11 participants. All sessions were recorded on video, and the content was transcribed using a commercial automatic speech recognition (ASR) system, (iFlyrec[2]). The research team verified the accuracy of the transcriptions to ensure reliability.

For both phases, four researchers independently applied open coding to the transcripts, after initially familiarizing themselves with the data. The research team then collaboratively discussed the coding results, resolving any disagreements through iterative refinement. This process of grouping the codes allowed us to identify overarching themes and subthemes that structured our findings.

Additionally, during the participatory design workshop, we collected digital copies of sketches created by participants, which were also analyzed in conjunction with the transcripts. The iterative structuring of the codes into themes was conducted during weekly meetings, where conflicts were resolved through discussion until consensus was achieved.

Selected quotes from both phases were translated by the first author and subsequently reviewed by co-authors to ensure accuracy.

5 FINDINGS

We present our key findings of the two parts of our study to answer two research questions, **RQ 1**: Based on part 1 of our formative study, we show users' needs and challenges during their VR search activities and present their insights on potential solutions. **RQ 2**: Based on the part 2 of our participatory design workshop, we summarized the design elements from five aspects.

5.1 The search difficulties users meet in the VR experience and how they solve them

This section summarizes the findings from the formative study, primarily answering **RQ 1**. We show VR search practices, the overall experiences of users, the challenges they face during their search activities, and their insights on potential solutions. These findings provided the necessary basis for the design of probes.

5.1.1 Search Needs in VR. Participants in the study indicated various scenarios in which they felt a need to search for information while using VR. Common situations include encountering unknown items or terminology within VR applications, needing assistance with VR hardware or software issues, and seeking deeper insights into VR content or gameplay strategies. For instance, participants frequently mentioned the necessity to search when they come across unfamiliar objects or when they need clarifications during VR experiences such as games, scientific research, and experimental setups. Additionally, the need for real-time information retrieval becomes critical when participants face technical challenges, such as hardware recognition errors or software malfunctions. The demand for search capabilities also extends to social interactions within VR, where participants wish to quickly find information relevant to their conversations without having to disrupt the immersive experience by removing the headset or switching applications.

5.1.2 Current Search Practices in VR. Participants described a range of search practices within VR environments, including using the virtual keyboard within the VR browsers and removing the VR headset to use other devices, such as mobile phones, for searching. In both cases, searching requires temporarily exiting the VR immersion to launch a separate search engine interface. The most prevalent method among our participants was physically removing the VR headset to use external devices for searching. For example, P1 stated:

> "If it's a standalone application, I take off the VR HMD and use my phone; I do not use VR's browser because typing is very inconvenient."

Similarly, P2 mentioned:

> "The only solution is to remember the content inside the VR space and then take off the HMD to search on a phone or computer."

5.1.3 Challenges with Current VR Search Interfaces. Participants reported significant disruptions to their immersive experience due to the inadequacies of current VR search interfaces and input methods. The necessity to alternate between VR and real-world devices is a major impediment. P3 highlighted the disruptive nature of this practice:

> "Taking off the helmet is a physically and mentally draining process; it affects focus and repeatedly doing so breaks immersion."

Participants also expressed dissatisfaction with the available VR browser interfaces and input methods, which they found cumbersome and inefficient. P4 criticized the current state of VR search tools:

> "I don't use VR's browser for searches; it's cumbersome and breaks my focus on the task at hand."

5.1.4 Desired Improvements and Potential Solutions. Participants provided valuable insights into potential improvements that could enhance the search experience within VR environments. There is a unanimous demand for more seamless and user-friendly search tools that integrate directly into the VR interface, minimizing or eliminating the need to disrupt immersion. P5 sees potential in voice-activated searches:

> "I would prefer to use a voice command to initiate searches without interrupting my activity or removing the headset."

P6 suggested an integration of context-sensitive searches:

> "It would be beneficial if the search could be context sensitive and provide information based on the specific activity I am engaged in within the VR environment."

P3 emphasized the need for advanced input methods:

[2] https://www.iflyrec.com/zhuanwenzi.html

> "We should explore touchless input methods such as eye tracking or gesture recognition that allow for more natural interactions within VR."

5.1.5 Summary. The findings reveal a critical gap in the VR technology pertaining to user-friendly and efficient search functionalities. The current need to frequently switch between virtual and real-world interfaces significantly hampers the immersive VR experience. Participants advocate for the development of integrated search tools that are both efficient and capable of maintaining immersion, with suggestions favoring voice commands, context-sensitive searches, and improved interaction methods. Acknowledging users' current challenges in searching, we designed probes based on literature to further explore people's preferences and expectations for VR search interactions.

5.2 Elements that should be considered when designing VR search interfaces leveraging Vision LLMs

This section summarizes the findings from the participatory design workshop, primarily answering **RQ 2**. From the probes presentation phase, through participants' feedback, we confirmed that six types of VR search probes can help users address search needs and challenges in the formative study. Based on the sketches and interview results from the co-design workshop phase, we summarized the design elements from five aspects: input interaction, output display, context-sensitive and intelligent recommendation, user participation in editing search results, and search tool anthropomorphization.

5.2.1 Input Interaction. Habit Transfer and Consistent User Interface: Habit transfer emphasizes how users apply familiar behaviors and operations in a new interface, while cross-device consistency focuses on providing a consistent user experience across different devices and platforms, making it easier for users to adapt to new systems. Participants (N=5) suggested that in addition to using a controller for selection, search tasks could also be performed through gestures, such as pointing at the target object with an index finger and combining it with voice queries. Other methods include using center-of-vision positioning or eye-tracking to initiate queries. P2 highlighted the potential of gesture and eye-tracking input:

> "Can input be with the finger or with eye tracking."

Some participants (N=3) highly endorsed using a search method similar to taking a screenshot on a smartphone and circling the area of interest. This approach is consistent with smartphone operations.

Efficiency is Important: Most participants (N=7) emphasized the importance of interaction efficiency. Some participants (N=4) preferred fewer interactions and faster system response times. Some participants believed the search should already have been preloaded and not initiated by the user.

5.2.2 Output Display. Integration of Search Results with the VR Scenes: As shown in Fig. 1, in the Probe 1 garden scene, there are two ways to display the text results: Display mode 1 is shown in the form of a window UI; Display mode 2 is to classify the retrieved text and finally display it in the form of expandable labels. According to our survey results with participants (N=8), text visualization should be harmonized with the 3D environment to reduce users' cognitive and reading burdens. For instance, users can easily identify and understand key data by integrating information labels in a non-intrusive manner within the scene while maintaining situational awareness. P5 and P6 expressly stated their preference for labels over text:

> "I like the label better. Because it is clearer than a paragraph of text, users can choose the information they want."

P11 added the importance of organized labels:

> "Labels are an ideal display method as they do not obstruct the view significantly. Labels are more organized, allowing users to grasp useful information quickly."

Dynamic Presentation of Generated Results: In the hospital scenario, we showed the participants the probe 4: Select the equipment and ask how to use it (Fig. 4). Some participants (N=3) Highlighted that presenting related content, such as images or videos, like traditional search engines, can disrupt the immersive experience in VR spaces. Instead, dynamically generating and displaying animations within the scene can significantly enhance immersion. P4 noted the immersive advantage of animations:

> "Animation and video are similar. However, users like animation more because it is more immersive and direct."

5.2.3 Context-sensitive and intelligent recommendation. Participants (N=5) believe the search interface should be designed to be contextually aware, offering relevant suggestions and results based on the current VR environment and application. This ensures users receive information tailored to their immediate context, enhancing their interaction experience, and the Fig. 9 Redrawn co-design workshop sketches in Appendix B.2 show this. P2 highlighted the need for flexible layout and accuracy verification:

> "Does not prefer a single, unified layout. It is useful to consider different types of inputs and scenarios and provide context-relevant outputs. When using LLMs, should be able to determine the accuracy of the output."

P3 emphasized the need for advanced input methods:

> "Desires different information presentation methods for different search goals. For example, provide interactive animations for 'how to use' queries and use labels for detailed information about target models."

P7 and P8 stressed the importance of adaptive labeling:

> "The system should continuously learn from users to improve labels. Different objects and users should have different labels. Incorrect labels can lead to user distrust."

5.2.4 Users participate in editing search results. Some participants (N=5) believe an effective VR search system should incorporate user feedback, collaborative learning, and participatory content refinement mechanisms. Also, the Fig. 9 Redrawn co-design workshop sketches in Appendix B.2 could show this. The system may enhance the overall user experience and foster a more engaged and interactive community by enabling users to contribute actively to content accuracy and relevance. P7 and P8 suggested a continuous improvement through user feedback:

> "There needs to be some feedback channel for users to submit, and the system needs to learn from the users to improve the label constantly."

P9 and P10 highlighted the importance of multi-user interest and recommendations:

> "Other users will be interested in this object. Can you recommend what other search?"

P11 emphasized the role of user involvement in ensuring description accuracy:

> "If the system is Internet accessible, it should involve the users in the precision of the description."

5.2.5 Search tool anthropomorphization. Some participants (N=3) mentioned that, based on the immersive characteristics of VR spaces and Vision LLMs' capabilities, the VR environment search tool could be anthropomorphized into a guide avatar. Users could ask questions in natural language and receive direct answers from the guide avatar, which aligns more closely with real-world intuition and perception. Additionally, giving the search tool a persona encourages users to build trust with the system and increases their willingness to interact. This is particularly beneficial in VR environments where immersion is crucial. P3 suggested envisioning the search tool as an intelligent voice assistant: "Think about Jarvis, the Ironman assistant." P11 supported the idea of representing the VR search tool in the form of avatar:

> "It is possible to think of this search function as a person you just have to ask for, just like in the real world."

6 DISCUSSION

We first present the key takeaways of our research and our key contributions in Section 6.1, and then we further discuss the design considerations based on our findings in Section 6.2. Finally, we present two demonstration scenarios in Appendix C Fig. 10 that illustrate the application of our design implications and considerations in practice.

6.1 Key Takeaways

Based on our formative study, we found that while search is a common need in VR applications, current approaches are often time-consuming and labour-intensive. Developing integrated search tools that maintain immersion while favoring voice commands, context-sensitive searches, and improved interaction methods is essential. The emergence of Vision LLMs has significantly enhanced search capabilities in 2D spaces, bringing great convenience to users. This advancement inspired us to explore the possibility of combining Vision LLMs with VR 3D spaces for search tasks. However, the design of a VR search interface leveraging Vision LLMs remains largely unexplored. Therefore, we combined the findings from our formative study with previous research on search interfaces to identify six common search needs in VR environments and designed corresponding solution probes using the Qwen-VL Vision LLM. We then recruited participants to explore these probes, gathered their feedback, and engaged in co-design sessions to discuss VR Search Interface Design Suggestions and Ideas. These include: input interaction, where participants emphasized the need for consistent search operations across devices, suggesting voice queries and gesture-based searches (e.g., pointing and speaking); output display, where search results should be seamlessly integrated into the scene (e.g., labels around objects) without disrupting immersion; context-sensitive and intelligent recommendations, where the search interface should offer relevant suggestions and results based on the current VR environment; user participation in editing search results, incorporating feedback, collaborative learning, and participatory content refinement mechanisms; and VR interface morphology, where anthropomorphizing the search tool as a wizard avatar aligns with real-world intuition (e.g., asking questions and receiving answers or demonstrations directly from the avatar). These insights can guide the design of VR search interfaces, enhance user experience, and maintain immersion.

6.2 Design Implications and Considerations

6.2.1 Scene Context Construction: Aligning Objects, Scenes, and Text Descriptions. Our research indicates that to improve search accuracy in VR, developers can create a comprehensive scene knowledge database to provide Vision LLMs with sufficient context. As LLMs are often black-box models that may struggle to capture factual knowledge accurately, they can produce incorrect results when given isolated data [3, 15]. Therefore, constructing a well-defined scene context is essential to ensure reliable information. For example, developers can use knowledge graphs, which can be used to store scene information and objects within the scene and their textual descriptions [26]. This database may facilitate contextually relevant search tasks within the scene, enhancing the user's search experience by supporting associative search queries.

6.2.2 Context-Sensitive Information Retrieval . As indicated by Section 5.2.3, participants believe the search interface should be context-aware, offering relevant suggestions and results based on the current state and environmental information. So our findings suggested recording context-sensitive information in VR, which could enhance the relevance of search results through real-time contextual and query contextual relationships [35], and it can help LLMs more accurately understand and process user input and thus respond more aligned with user needs [16, 46]. Here, we borrow the key information factors in the XAIR interpretive AR framework to categorize context-sensitive information into three categories for VR [47]: **User State**: This includes tracking user behaviour, attention, and potential intent. Understanding the user's focus and actions within the VR environment helps tailor search results to their current needs and interests. **Scene Information**: This encompasses spatial coordinates, the current scene's semantic meaning, and the user's field of view semantics. Accurate scene information ensures that the search results are relevant to the user's current context within the VR environment. **User intent**: The search mechanism should understand user intent so the system can determine the next best action based on the intent.

6.2.3 Executing Search Paths Based on User Input Modalities and Executing Display Actions Based on the type of Search. Our findings reveal that participants have high expectations for the VR search system's ability to understand user intentions. While users may not always actively input prompts when handling visual search tasks, the system could intelligently assess the input content to determine the appropriate course of action, illustrated by the examples below: **Visual Information Only, No Prompt Input**: When the user provides only visual information without an accompanying prompt, the system leverages context-sensitive information to generate a relevant prompt via Vision LLMs, which then process this prompt to produce text-based search results. **Visual Information with Prompt Input**: When the user provides both visual information and a prompt, the system integrates these inputs with context-sensitive information to refine the prompt. Vision LLMs then process the refined prompt to deliver precise text-based search results. **Text and Audio Prompt**: In scenarios where the object is out of view, making visual input impossible, users should be able to initiate the search using text or audio input alone.

Depending on the question types, different disply actions should be triggered: **Object Identification Questions**: If the user asks about the identity of an object, the system triggers the labels display action. This action generates multiple labels around the target object and arranges them in order. **Object Interaction Questions**: If the user inquires about how to interact with an object, the system triggers the animation display action. Vision LLMs generate interaction animations and logic to demonstrate how to interact with the object. By establishing a mapping between input content and action paths to understand user intent and execute tasks, the complexity of these actions can be concealed, making them easily accessible to non-skilled users [9].

6.2.4 Enhancing Search Results through User Feedback. As shown in Section 5.2.4, participants emphasized the importance of the feedback function for search results. This feature is especially vital because LLMs can sometimes generate hallucinations that distort accuracy, and pre-trained models may include outdated content [3, 15]. So the system could provide users with the ability to review the search results, make corrections to mistakes or add information via audio or text. For example, setting up direct feedback buttons, interactive scoring systems or pop-up feedback forms, allows users to quickly assess the quality of the answers and collect more detailed user feedback by evaluating the search results as "useful" or "unhelpful," rating the results (e.g. 1-5 stars), and providing detailed comments after a certain number of interactions. In addition, implicit feedback mechanisms can also be used, such as providing two or more results for each search task, using logs to record the user's selection order and reading or listening time, then building a planning model to improve the relevance of search results [19, 38].

6.2.5 No disruptions to immersive experience. Our findings revealed that keyboard input or jumping other windows to make a search action may affect the user's ability to perform other tasks in VR. Therefore, using a natural user interface (NUI), such as voice commands, eye movements combined with the controllers, or specific gestures to initiate search commands can significantly enhance the user's immersion [27, 28]. For example, the method of selecting an area by drawing a circle, which is commonly used in mobile phones, such as TapTell and Google Circle to Search [10], has been verified in mobile visual search tasks and can effectively improve the user experience [31]. This interaction method can also be migrated to VR search tasks, using a controller or gestures combined with voice commands to prompt the visual information to the Vision LLMs. Moreover, our findings suggested that the display of search results should not obstruct the user's FOV. The results can be divided into labels and surrounded by the search object target. In addition, the display method should also be determined based on the type of search and the content of the search. For example, in the task of knowing about interactions with objects (KO), the interaction methods can be displayed in the form of a model animation, so that the search task does not destroy the immersive experience of VR.

6.3 Limitations and Future Work

Our study has several limitations that future work should address. We did not evaluate the effectiveness of Vision LLMs in handling context-sensitive questions within VR, partly due to technical constraints that prevented the full deployment of their functions in our probes. Additionally, while we based our assumptions on the capabilities of Vision LLMs as described in peer-reviewed papers, their full potential still needs to be explored. The issue of hallucinations in LLMs, which can result in errors, underscores the need for multiuser participation to help correct inaccuracies. Future research should focus on developing working prototypes and conducting both qualitative and quantitative evaluations. We have identified a gap in the current research, and addressing this will be crucial for defining how VR search mechanisms can be effectively implemented, whether as a plugin or a core system feature.

7 CONCLUSION

In this study, we summarised the current VR search needs and challenges faced by users through a formative study, and designed six probes to address these challenges based on previous research and the capabilities of Vision LLMs. We then conducted a participatory design workshop, and based on the participant's feedback, we obtained five design considerations for VR search interfaces. In summary, our work represents a foundational step in exploring the design of VR search interfaces and could offer guidance for designers and developers moving forward.

ACKNOWLEDGMENTS

This work is partially supported by Guangzhou-HKUST(GZ) Joint Funding Project (No.: 2024A03J0617), HKUST Practice Research with Project title "RBM talent cultivation Exploration" (No.: HKUST(GZ)-ROP2023030), Guangzhou Science and Technology Program City-University Joint Funding Project (No.: 2023A03J0001), and Guangdong Provincial Key Lab of Integrated Communication, Sensing and Computation for Ubiquitous Internet of Things (No.2023B1212010007).

REFERENCES

[1] Bon Adriel Aseniero, Michael Lee, Yi Wang, Qian Zhou, Nastaran Shahmansouri, and Rhys Goldstein. 2024. Experiential Views: Towards Human Experience Evaluation of Designed Spaces using Vision-Language Models. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems (CHI EA '24).*

Association for Computing Machinery, New York, NY, USA, Article 136, 7 pages. https://doi.org/10.1145/3613905.3650815

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. arXiv:2308.12966 [cs.CV] https://arxiv.org/abs/2308.12966

[3] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiehzheng Yu, Willy Chung, Quyet Do, Xu Yan, and Pascale Fung. 2023. A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. 675–718. https://doi.org/10.18653/v1/2023.ijcnlp-main.45

[4] Kirsten Boehner, Janet Vertesi, Phoebe Sengers, and Paul Dourish. 2007. How HCI interprets the probes. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '07)*. Association for Computing Machinery, New York, NY, USA, 1077–1086. https://doi.org/10.1145/1240624.1240789

[5] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023. Shikra: Unleashing Multimodal LLM's Referential Dialogue Magic.

[6] Taizhou Chen, Yi-Shiun Wu, and Zhu Kening. 2018. Investigating different modalities of directional cues for multi-task visual-searching scenario in virtual reality. 1–5. https://doi.org/10.1145/3281505.3281516

[7] Cathy Edwards. 2024. Circle (or highlight or scribble) to Search. Blog Post. https://blog.google/products/search/google-circle-to-search-android/ Accessed: 2024-05-19.

[8] BoYu Gao, Tong Shao, Huawei Tu, Qizi Ma, Zitao Liu, and Teng Han. 2024. Exploring Bimanual Haptic Feedback for Spatial Search in Virtual Reality. *IEEE transactions on visualization and computer graphics* PP (03 2024). https://doi.org/10.1109/TVCG.2024.3372045

[9] Valentina Gatteschi, Fabrizio Lamberti, Paolo Montuschi, and Andrea Sanna. 2016. Semantics-Based Intelligent Human-Computer Interaction. *IEEE Intelligent Systems* 31, 4 (2016), 11–21. https://doi.org/10.1109/MIS.2015.97

[10] Google. 2024. The Circle of Life: Bringing Google Search to Android. https://blog.google/products/search/google-circle-to-search-android/. Accessed: 2024-08-13.

[11] Kristen Grinyer and Robert J. Teather. 2022. Effects of Field of View on Dynamic Out-of-View Target Search in Virtual Reality. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. 139–148. https://doi.org/10.1109/VR51125.2022.00032

[12] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 2023. 3D-LLM: Injecting the 3D World into Large Language Models. arXiv:2307.12981 [cs.CV] https://arxiv.org/abs/2307.12981

[13] Sathaporn Hu, Joseph Malloch, and Derek Reilly. 2020. A Comparative Evaluation of Techniques for Locating Out of View Targets in Virtual Reality. In *Graphics Interface 2021*. https://doi.org/10.20380/GI2021.32

[14] Ananya Ipsita, Levi Erickson, Yangzi Dong, Joey Huang, Alexa K Bushinski, Sraven Saradhi, Ana M Villanueva, Kylie A Peppler, Thomas S Redick, and Karthik Ramani. 2022. Towards Modeling of Virtual Reality Welding Simulators to Promote Accessible and Scalable Training. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 566, 21 pages. https://doi.org/10.1145/3491102.3517696

[15] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* 55, 12, Article 248 (mar 2023), 38 pages. https://doi.org/10.1145/3571730

[16] Salomon Kabongo, Jennifer D'Souza, and Sören Auer. 2024. Effective Context Selection in LLM-based Leaderboard Generation: An Empirical Study. *arXiv preprint arXiv:2407.02409* (2024). https://doi.org/10.48550/arXiv.2407.02409

[17] CHI 2017nic Kao, Alejandra J. Magana, and Christos Mousas. 2021. Evaluating Tutorial-Based Instructions for Controllers in Virtual Reality Games. *Proc. ACM Hum.-Comput. Interact.* 5, CHI PLAY, Article 234 (oct 2021), 28 pages. https://doi.org/10.1145/3474661

[18] Oliver Beren Kaul and Michael Rohs. 2017. HapticHead: A Spherical Vibrotactile Grid around the Head for 3D Guidance in Virtual and Augmented Reality. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 3729–3740. https://doi.org/10.1145/3025453.3025684

[19] Jin Young Kim, Mark Cramer, Jaime Teevan, and Dmitry Lagun. 2013. Understanding how people interact with web search results that change in real-time using implicit feedback. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management* (San Francisco, California, USA) *(CIKM '13)*. Association for Computing Machinery, New York, NY, USA, 2321–2326. https://doi.org/10.1145/2505515.2505663

[20] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. 2017. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. arXiv:1609.04802 [cs.CV] https://arxiv.org/abs/1609.04802

[21] Chang Liu, Felicia Fang-Yi Tan, Shengdong Zhao, Abhiram Kanneganti, Gosavi Arundhati Tushar, and Eng Tat Khoo. 2024. Facilitating Virtual Reality Integration in Medical Education: A Case Study of Acceptability and Learning Impact in Childbirth Delivery Training. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 458, 14 pages. https://doi.org/10.1145/3613904.3642100

[22] Lijia Ma, Xingchen Xu, and Yong Tan. 2024. Crafting Knowledge: Exploring the Creative Mechanisms of Chat-Based Search Engines. *arXiv preprint arXiv:2402.19421* (2024). https://doi.org/10.48550/arXiv.2402.1942

[23] Mark Mcdaniel, Gilles Einstein, Thomas Graham, and Erica Rall. 2004. Delaying execution of intentions: Overcoming the costs of interruptions. *Applied Cognitive Psychology* 18 (07 2004), 533 – 547. https://doi.org/10.1002/acp.1002

[24] Ann McNamara, Katherine Boyd, Joanne George, Weston Jones, Somyung Oh, and Annie Suther. 2019. Information Placement in Virtual Reality. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. 1765–1769. https://doi.org/10.1109/VR.2019.8797891

[25] Victor Adriel Oliveira, Luca Brayda, Luciana Nedel, and Anderson Maciel. 2017. Designing a Vibrotactile Head-Mounted Display for Spatial Awareness in 3D Spaces. *IEEE Transactions on Visualization and Computer Graphics* PP (01 2017), 1–1. https://doi.org/10.1109/TVCG.2017.2657238

[26] Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying Large Language Models and Knowledge Graphs: A Roadmap. *IEEE Transactions on Knowledge and Data Engineering* 36, 7 (2024), 3580–3599. https://doi.org/10.1109/TKDE.2024.3352100

[27] Kyeong-Beom Park and Jae Yeol Lee. 2016. Comparative Study on the Interface and Interaction for Manipulating 3D Virtual Objects in a Virtual Reality Environment. *Transactions of the Society of CAD/CAM Engineers* 21 (03 2016), 20–30. https://doi.org/10.7315/CADCAM.2016.020

[28] Daniele Regazzoni, Caterina Rizzi, and Andrea Vitali. 2018. Virtual reality applications: guidelines to design natural user interface. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, Vol. 51739. V01BT02A029. https://doi.org/10.1115/DETC2018-85867

[29] Maximilian Rettinger, Niklas Müller, Christopher Holzmann-Littig, Marjo Wijnen-Meijer, Gerhard Rigoll, and Christoph Schmaderer. 2021. VR-based Equipment Training for Health Professionals. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, Article 252, 6 pages. https://doi.org/10.1145/3411763.3451766

[30] Rufat Rzayev, Polina Ugnivenko, Sarah Graf, Valentin Schwind, and Niels Henze. 2021. Reading in VR: The Effect of Text Presentation Type and Location. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 531, 10 pages. https://doi.org/10.1145/3411764.3445606

[31] Jitao Sang, Tao Mei, Ying-Qing Xu, Chen Zhao, Changsheng Xu, and Shipeng Li. 2013. Interaction Design for Mobile Visual Search. *IEEE Transactions on Multimedia* 15, 7 (2013), 1665–1676. https://doi.org/10.1109/TMM.2013.2268052

[32] Maurice Schleußinger. 2021. Information retrieval interfaces in virtual reality—A scoping review focused on current generation technology. *Plos one* 16, 2 (2021), e0246398.

[33] Maurice Schleußinger. 2021. Information retrieval interfaces in virtual reality-A scoping review focused on current generation technology. *PloS one* 16 (02 2021), e0246398. https://doi.org/10.1371/journal.pone.0246398

[34] Nikhil Sharma, Q. Vera Liao, and Ziang Xiao. 2024. Generative Echo Chamber? Effect of LLM-Powered Search Systems on Diverse Information Seeking. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 1033, 17 pages. https://doi.org/10.1145/3613904.3642459

[35] Xuehua Shen, Bin Tan, and ChengXiang Zhai. 2005. Context-sensitive information retrieval using implicit feedback. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Salvador, Brazil) *(SIGIR '05)*. Association for Computing Machinery, New York, NY, USA, 43–50. https://doi.org/10.1145/1076034.1076045

[36] Shashi Kant Singh, Shubham Kumar, and Pawan Singh Mehra. 2023. Chat GPT and Google Bard AI: A Review. In *2023 International Conference on IoT, Communication and Automation Technology (ICICAT)*. 1–6. https://doi.org/10.1109/ICICAT57735.2023.10263706

[37] Miriam Sturdee and Joseph Lindley. 2019. Sketching & drawing as future inquiry in HCI. In *Proceedings of the Halfway to the Future Symposium 2019*. 1–10.

[38] Ashok Veilumuthu and Parthasarathy Ramachandran. 2007. Discovering Implicit Feedbacks from Search Engine Log Files. (2007), 231–242. https://doi.org/10.1007/978-3-540-75488-6_22

[39] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Large Search Model: Redefining Search Stack in the Era of LLMs. *SIGIR Forum* 57, 2, Article 23 (jan 2024), 16 pages. https://doi.org/10.1145/3642979.3643006

[40] Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. 2024. Visionllm: Large language

model is also an open-ended decoder for vision-centric tasks. *Advances in Neural Information Processing Systems* 36 (2024).

[41] Austin Ward, Sandeep Avula, Hao-Fei Cheng, Sheikh Muhammad Sarwar, Vanessa Murdock, and Eugene Agichtein. 2023. Searching for Products in Virtual Reality: Understanding the Impact of Context and Result Presentation on User Experience. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Taipei, Taiwan) *(SIGIR '23)*. Association for Computing Machinery, New York, NY, USA, 2359–2363. https://doi.org/10.1145/3539618.3592057

[42] Austin Ward and Rob Capra. 2020. Immersive Search: Using Virtual Reality to Examine How a Third Dimension Impacts the Searching Process. 1621–1624. https://doi.org/10.1145/3397271.3401303

[43] Austin Ward, Yiyin Gu, Sandeep Avula, and Praneeth Chakravarthy. 2021. Interacting with Information in Immersive Virtual Environments. 2600–2604. https://doi.org/10.1145/3404835.3462787

[44] Frederik Winther, Linoj Ravindran, Kasper Paabøl Svendsen, and Tiare Feuchtner. 2020. Design and Evaluation of a VR Training Simulation for Pump Maintenance Based on a Use Case at Grundfos. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. 738–746. https://doi.org/10.1109/VR46266.2020.00097

[45] Penghao Wu and Saining Xie. 2023. V*: Guided Visual Search as a Core Mechanism in Multimodal LLMs. arXiv:2312.14135 [cs.CV] https://arxiv.org/abs/2312.14135

[46] Junjielong Xu, Ziang Cui, Yuan Zhao, Xu Zhang, Shilin He, Pinjia He, Liqun Li, Yu Kang, Qingwei Lin, Yingnong Dang, Saravan Rajmohan, and Dongmei Zhang. 2024. UniLog: Automatic Logging via LLM and In-Context Learning. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering* (Lisbon, Portugal) *(ICSE '24)*. Association for Computing Machinery, New York, NY, USA, Article 14, 12 pages. https://doi.org/10.1145/3597503.3623326

[47] Xuhai Xu, Anna Yu, Tanya R. Jonker, Kashyap Todi, Feiyu Lu, Xun Qian, João Marcelo Evangelista Belo, Tianyi Wang, Michelle Li, Aran Mun, Te-Yen Wu, Junxiao Shen, Ting Zhang, Narine Kokhlikyan, Fulton Wang, Paul Sorenson, Sophie Kim, and Hrvoje Benko. 2023. XAIR: A Framework of Explainable AI in Augmented Reality. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 202, 30 pages. https://doi.org/10.1145/3544548.3581500

[48] Soorim Yang, Hyeong jun Joo, and Jaeho Kim. 2024. Metaverse search system: Architecture, challenges, and potential applications. *ICT Express* 10, 2 (2024), 431–441. https://doi.org/10.1016/j.icte.2023.12.006

[49] Zhe-Xin Zhang. 2024. A Design of Interface for Visual-Impaired People to Access Visual Information from Images Featuring Large Language Models and Visual Language Models. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 390, 4 pages. https://doi.org/10.1145/3613905.3648648

[50] Andrew Zhou and Grace Yang. 2018. Minority Report by Lemur: Supporting Search Engine with Virtual Reality. 1329–1332. https://doi.org/10.1145/3209978.3210179

[51] Paul Zikas, Manos Kamarianakis, Ioanna Kartsonaki, Nick Lydatakis, Steve Kateros, Mike Kentros, Efstratios Geronikolakis, Giannis Evangelou, Achilles Apostolou, Paolo Alejandro Alejandro Catilo, and George Papagiannakis. 2021. Covid-19 - VR Strikes Back: innovative medical VR training. In *ACM SIGGRAPH 2021 Immersive Pavilion* (Virtual Event, USA) *(SIGGRAPH '21)*. Association for Computing Machinery, New York, NY, USA, Article 11, 2 pages. https://doi.org/10.1145/3450615.3464546

A APPENDIX: FORMATIVE STUDY INFORMATION

A.1 Demographic information table

Table 1: Demographic of formative study participants (N=10)

ID	Age	Experience (years)	Frequency	Purpose
P1	23	8	3-4 times per week	Development, game
P2	22	2	4-5 times per year	Game
P3	24	1	2-3 times per month	Development
P4	24	2	2-3 times per year	VR research participant
P5	18	3	Everyday	VR Chat
P6	25	4	2-3 times per week	Development
P7	40	8	3 hours per month	Game, new applications
P8	42	5	20 hours per month	Work, social VR
P9	25	4	10 hours per month	Game, VR Chat, teaching
P10	24	6	5 times per week	Development

A.2 Semi-structured interview questions

(1) What do you usually use VR devices for?
(2) Do you have any search needs in VR (similar to your search needs on PC and mobile phones)? Have you ever conducted a search within standalone VR applications, and if so, for what purposes?
(3) When you encounter a search need in VR, what is your current solution? Why did you choose this method?
(4) Does the current solution you use affect your immersive experience? Why or why not?
(5) In VR, how would you prefer to search for a target object, image or text? Why?
(6) Suppose you want to learn about the green object in the VR scene without exiting the current application, as shown in Fig. 8. How would you prefer to search for it? (any interaction modality or method is acceptable)

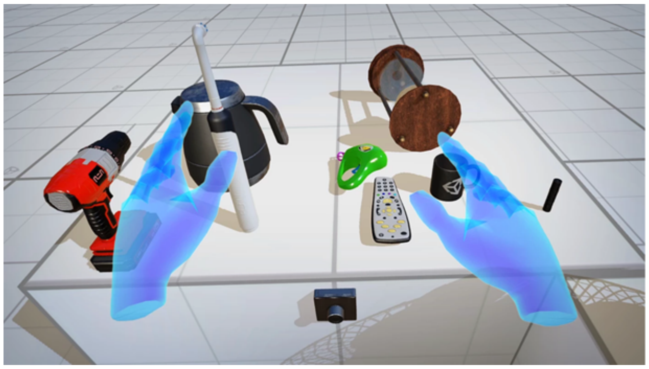


Figure 8: Semi-structured interview sixth question scene

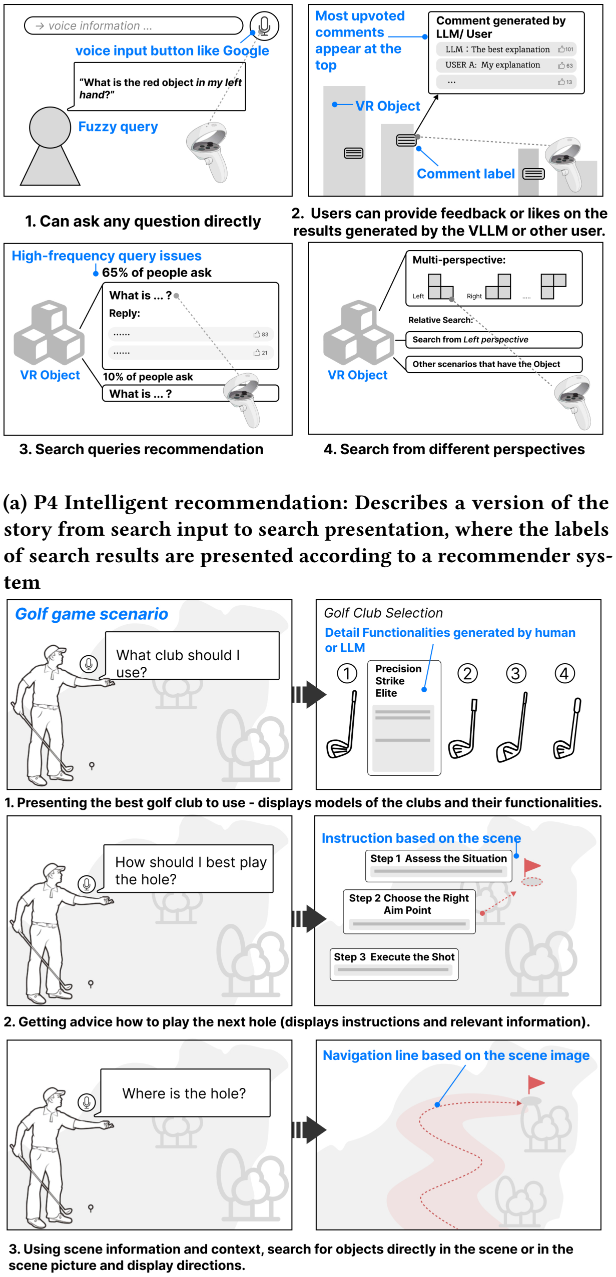
B APPENDIX: WORKSHOP INFORMATION

B.1 Demographic information table

Table 2: Demographic of workshop participants (N=1)

ID	Group	Age Range	Experience (years)	Occupation
P1	1	41-45	5	Game Developer
P2	2	18-25	5	Postgraduate Student
P3	3	36-40	4	Product Owner
P4	4	18-25	1	Postgraduate Student
P5	5	18-25	0 year, only 1-2 times	Postgraduate Student
P6	5	18-25	0 year, only 1-2 times	Postgraduate Student
P7	6	18-25	1	College graduate
P8	6	26-30	4	Creative technologist
P9	7	18-25	5	VR game player
P10	7	18-25	7	VR researcher
P11	8	18-25	6	VR researcher; VR Game developer

B.2 Co-design sketches


→ voice information ...
voice input button like Google
"What is the red object *in my left hand?*"
Fuzzy query
1. Can ask any question directly
Most upvoted comments appear at the top
Comment generated by LLM/ User
LLM : The best explanation 101
USER A: My explanation 63
... 13
VR Object
Comment label
2. Users can provide feedback or likes on the results generated by the VLLM or other user.
High-frequency query issues
65% of people ask
What is ... ?
Reply:
...... 83
...... 21
VR Object
10% of people ask
What is ... ?
3. Search queries recommendation
Multi-perspective:
Left
Right
.....
Relative Search:
Search from *Left perspective*
Other scenarios that have the Object
VR Object
4. Search from different perspectives

(a) P4 Intelligent recommendation: Describes a version of the story from search input to search presentation, where the labels of search results are presented according to a recommender system

Golf game scenario
What club should I use?
Golf Club Selection
Detail Functionalities generated by human or LLM
①
Precision Strike Elite
②
③
④
1. Presenting the best golf club to use - displays models of the clubs and their functionalities.
How should I best play the hole?
Instruction based on the scene
Step 1 Assess the Situation
Step 2 Choose the Right Aim Point
Step 3 Execute the Shot
2. Getting advice how to play the next hole (displays instructions and relevant information).
Where is the hole?
Navigation line based on the scene image
3. Using scene information and context, search for objects directly in the scene or in the scene picture and display directions.

(b) P3 Context sensitive: Describes a golfing scenario for a search task

Figure 9: Redrawn co-design workshop sketches

C APPENDIX: AN EXAMPLE OF VR SEARCH APPLICATION SCENARIO

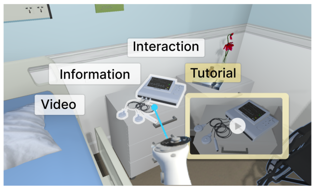
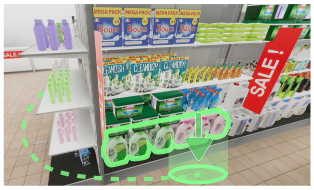
	 **Case 1: Know Objects Behaviors**	 **Case 2: Find Objects Behaviors**	**Universal Workflow**
Phase 1 Scene Knowledge Graph Construction	Developers upload the tutorial video into the scene's prior knowlege.	Developers upload the position information of the objects into the scene's prior knowlege.	3D Object/ Scene Multimedia Material Knowledge Knowledge Graph
Phase 2 Acquire Context Sensitive Information	When the scene is running, system will acquire user's state information and environment information in real time as context information.	When the scene is running, system will acquire user's state information and environment information in real time as context information.	**Player State** Activity, Attention, Potential Intent... **Contextual Information** Coordinate Information, Environment Object, Time...
Phase 3 User Input	User select the equipment and audio ask "How to use this equipment?"	User ask "Where is the detergent in the supermarket?"	Picture + Prompts Selection + Prompts Only Prompts
Phase 4 Context-sensitive results generation	Based on the information given from phase 1-3, system generates results and chooses how to display the results. Under scenario in this case, system chooses to display video result because video is easier for user to understand how to use this equipment than text.	Based on the information given from phase 1-3, system generates results and chooses which information to display. Under finding object scenario in this case, system chooses navigation display to help users find objects.	Use previous 3 phases' knowledge to generate results Choose the best display method according to the result type
Phase 5 Display and Action	In this case, system displays video result given in th prior information.	In this case, system acquires the position information from the piror information and displays highlight navigation to guide users to the object they want to find.	**Display information** Label Display Animation Display Video Display etc.

Figure 10: An example of VR search application scenarios