# A Comparative Study of Speech-and-Pointing and Disocclusion Mini-Map Techniques for Object Selection in Virtual Reality

Junlong Chen
University of Cambridge
Cambridge, United Kingdom
jc2375@cam.ac.uk

Jens Grubert
Coburg University of Applied Sciences and Arts
Coburg, Germany
jens.grubert@gmail.com

Per Ola Kristensson
University of Cambridge
Cambridge, United Kingdom
pok21@cam.ac.uk

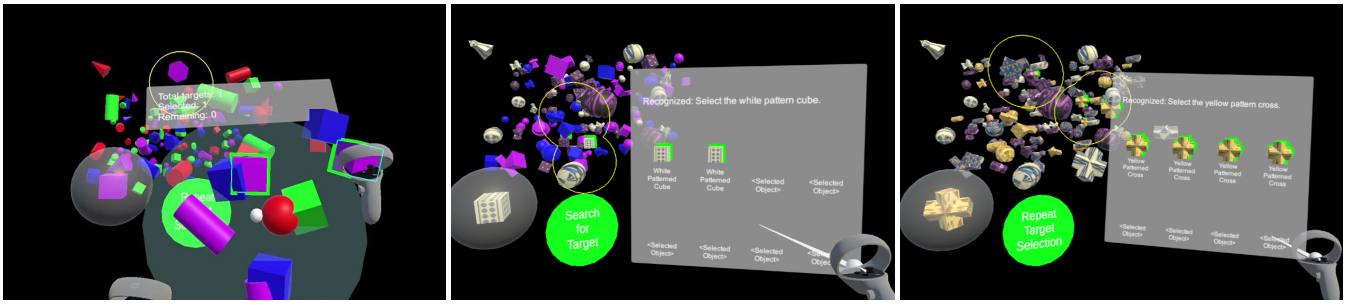


**Figure 1: First-person views from the object selection user study with DiscPIM [29] and AssistVR [10] under different scene perplexity conditions. DiscPIM displays objects hit by the raycast on a circular mini-map. Cluttered objects in the mini-map can be further expanded along the circumference to facilitate selection. AssistVR supports single or multi-object selection via speech or raycast. Recognized speech and selected objects are displayed on a draggable panel. Left: Low perplexity, DiscPIM; Middle: Medium perplexity, AssistVR; Right: High perplexity, AssistVR.**

## Abstract

Object selection in virtual reality (VR) becomes increasingly challenging when targets are occluded or when multiple objects must be selected in complex 3D scenes. In this paper, we present a comparative study of two interaction techniques: a speech-and-pointing dual-modality approach (AssistVR) which adopts natural language input and raycasting, and a disocclusion mini-map (DiscPIM) which displays dynamically-updated minimized objects in the selection cone while preserving the object relative position in two dimensions. We conducted a within-subjects study (n=24) that varied the number of targets to be selected and the perplexity of the scene to create a grid of conditions designed to reveal performance crossover points. The study aims to explore *modality strengths and tradeoffs*, not superiority. Quantitative and qualitative results show that each technique exhibits strengths and weaknesses depending on the task context: for example, mini-maps facilitate efficient selection in high-occlusion single-target scenarios, while speech-and-pointing is better suited for multi-target tasks. Our findings offer practical guidance for VR interface designers by outlining the conditions under which each technique is most effective. We conclude with a summary table of design recommendations grounded in study results to inform the choice of object selection techniques in future immersive applications.


*SUI '25, Montreal, QC, Canada*

ACM ISBN 979-8-4007-1259-3/25/11
https://doi.org/10.1145/3694907.3765913

## CCS Concepts

• **Human-centered computing** → **Virtual reality**; **Usability testing**.

## Keywords

Human-computer interaction (HCI), virtual reality, speech interfaces, large language models.

**ACM Reference Format:**
Junlong Chen, Jens Grubert, and Per Ola Kristensson. 2025. A Comparative Study of Speech-and-Pointing and Disocclusion Mini-Map Techniques for Object Selection in Virtual Reality. In *ACM Symposium on Spatial User Interaction (SUI '25), November 10–11, 2025, Montreal, QC, Canada.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3694907.3765913

## 1 Introduction

Object selection in virtual reality is an important and widely researched task. However, this task becomes increasingly challenging when objects are occluded, densely packed, or when multiple targets must be selected simultaneously. For occluded object selection alone, a myriad of tools [29, 45, 63, 64] have been developed to facilitate the task and improve user experience. In such cases, designers must choose from a variety of interaction techniques, each with distinct strengths and limitations. Understanding the tradeoffs between these techniques is essential for creating effective and context-appropriate VR interfaces.

This paper presents a comparative study of two object selection techniques in VR: *speech-and-pointing* (AssistVR [10]) which adopts natural language input and raycasting, and the *disocclusion*

*mini-map* (DiscPIM [29]), which reveals occluded objects in the selection cone on a mini-map attached to the controller.

Our focus on DiscPIM is motivated by its representative role among spatial assistance techniques that use visual feedback to overcome occlusion. While a wide range of object selection techniques exist in VR through raycasting, cones, and hand avatars [1], few are designed explicitly for occluded object selection in cluttered 3D environments. We chose DiscPIM over other approaches because it outperforms existing occluded object selection techniques in VR [29]. For the speech-and-pointing technique, we build upon the existing work of Chen et al. [10] to leverage language models to complement traditional object selection techniques by adapting the conversation-based AssistVR technique [10] for the task of object selection in VR. Compared with Chen et al. [10] which focused on interaction strategies without a comparative evaluation and Maslych et al. [29] which focused on selecting a single occluded object with a distinct appearance, our approach provides a systematic user study to investigate speech-and-pointing versus disocclusion minimap techniques in occluded single and multi-object selection tasks. Through this comparison, we provide empirical evidence of the strengths and weaknesses of each approach and provide design recommendations for future systems. In our adaptation, AssistVR combines speech and raycast as an interoperable and complementary selection technique. In the remainder of this paper, AssistVR will refer to our adapted technique for object selection unless otherwise specified. As such, we are able to directly compare DiscPIM [29] with AssistVR. This comparison enables us to explore tradeoffs between two fundamentally different approaches: one that relies on explicit visual remapping, and another that leverages cognitive offloading through language-based intent expression.

Rather than proposing a new interaction technique or evaluating one as superior to the other, our goal is to systematically investigate the conditions under which each technique is most effective. To this end, we conduct a within-subjects study that varies the number of targets to be selected and the visual and referencing complexity of the scene (referred to as *perplexity*). Here, we hypothesize an ideal experimental environment to remain consistent with the original study setup of the baseline technique [29], where the target object is known to the user, the region of interest is clearly indicated, and target objects in this region possess a distinct name and appearance compared with distractor objects. To facilitate comparison with the DiscPIM baseline [29], we recruited the same number of 24 participants and invited them to complete a *search* task and a *repeat* task [29, 37, 64] for each combination of independent variables. *Search* and *repeat* trial completion time serves as performance metrics, while ratings from questionnaires provide an indicator of system usability, user experience, and task load. By following this established protocol, we ensure comparability of our results to prior benchmarks. This design enables us to explore potential performance crossover points and identify scenarios where one technique outperforms the other. Our findings suggest that each technique has situational advantages. For example, the mini-map supports faster selection with few targets, while speech-and-pointing excels when selecting two or more objects, even when objects were difficult to reference verbally. Both techniques were able to attain similar user experience ratings.

These results are intended to inform design decisions rather than validate a single best approach. We conclude with a set of practical design recommendations in a summary table. Through this comparative exploration, we aim to contribute actionable insights for VR designers seeking to match interaction techniques with the demands of specific selection tasks.

## 2 Related Work

Our work is embedded in the areas of occluded object selection in VR, speech-and-pointing interaction in immersive environments, as well as work on language models, which we contextualize next.

### 2.1 Occluded Object Selection in VR

Object selection and manipulation are fundamental interactions in VR [3, 31, 40, 54]. They are often evaluated in testbed experiments [5] and as the interface evolves, are also widely applicable in the consumer application space [30]. As such, many works have been dedicated to improve object selection [26, 46] and manipulation [66] by proposing novel interaction techniques. Besides the virtual hand and raycast techniques, which are commonly used as interaction metaphors for selection and manipulation [48], gestures [26] and eye gaze [46, 55] are also prominent for object selection. These basic selection and manipulation tasks are widely studied in both VR and AR [27, 38]. Selection can be considered as the first step in the sequential process of referencing [8]. Following Weiß et al. [57] and Schüssel et al. [44], selection tasks provide a fundamental prerequisite for subsequent manipulation tasks, and findings on selection tasks will provide important implications for the design of interaction techniques when they are applied in other scenarios for other tasks.

We focus on challenging scenarios like occluded object selection. Back in 2007, Vanacken et al. [53] highlighted the research gap in dense and occluded object selection in 3D virtual environments and proposed the depth ray and the 3D bubble cursor to address this gap. Yu et al. [62] studied the performance of different techniques in object selection under dense and occluded environments. Baloup et al. [2] proposed RayCursor to use a controllable cursor on the ray to select 3D objects. Later in 2020, Yu et al. [64] developed a set of seven techniques (Alpha Cursor, Flower Cone, Gravity Zone, Grid Wall, Lasso Grid, Magic Ball, and Smash Probe) for fully-occluded target selection in VR. In the same year, Sidenmark et al. proposed Outline Pursuits [46]. Subsequent works adopted ray-based metaphors such as IntenSelect [14] and IntenSelect+ [25], LenSelect [58], TouchRay [32], ClockRay [60], redirected rays [17], and freehand pointing selection techniques [45]. Additionally, ray metaphors can also be augmented by scene context [56], minimap grabbing selections [29], and eye gaze [9] for occluded object selection tasks.

One prominent approach among these techniques is the Disc Projective Interactive Map (DiscPIM) [29], which dynamically projects occluded targets on a disc-shaped mini-map. This preserves the original spatial relationship and remaps dense occluded objects onto the mini-map to make them sparse. According to Maslych et al. [29], DiscPIM outperformed the previously best-performing technique GravityZone+ and results in lower average trial time compared with CylinderPIM proposed in the same paper. Based on

its state-of-the-art performance and representative minimap-based disocclusion strategy which reduces user effort and offers direct manipulation in complex scenes, we have selected DiscPIM to contrast meaningfully with language-based methods in our evaluation of different modalities in VR object selection tasks.

## 2.2 Multimodal Interaction Techniques

Back in 1980, Richard Bolt proposed "Put-That-There" [4], a voice and gesture multimodal interface for placing elements on a graphics display. The work became an example of how the speech modality could be applied in conjunction with other referencing modalities and inspired subsequent works. In 1999, Sharon Oviatt summarized ten myths in multimodal interaction [35]. In 2004, Reeves et al. proposed a set of guidelines for the design of multimodal interfaces [42]. Olwal et al. [33] proposed how prosodic features of speech and audio localization could be applied in interactive applications. During this period, Oviatt et al. [36] suggested that users spontaneously shift to multimodal communication when task load increases, ultimately reaching a mix of unimodal and multimodal communication patterns, which is also observed in our study when participants switched from single to multi-object selection tasks. Later in 2013, Schüssel et al. [44] studied influencing factors of multimodal interaction and found a strong predominance in touch single modality input, and a rare occurrence in multimodal inputs for easy selection tasks. Turk [52] outlined challenges in sensing, recognition, usability, interaction, and privacy for multimodal HCI interface design. Recent reviews [23, 39, 41] summarize the growing role of multimodal interfaces in virtual and augmented reality.

For the speech interaction modality specifically, Clark et al. [13] conducted a thematic review of speech interfaces in HCI. Weiß et al. [57] compared 2D/3D/speech interfaces in VR, finding speech prone to usage, parsing, and recognition errors, but easy to learn for text-heavy tasks. More recent work focuses on design for accessibility [28], natural conversation [24], and testing tools [19].

Dual-modality input, especially combining speech and pointing, offers a flexible alternative to purely spatial techniques by enabling users to leverage natural language for object referencing [4, 51]. Speech input can provide semantic descriptions, disambiguate targets, or express complex selection intents that may be difficult to convey visually. When combined with pointing (e.g., raycasting), this approach supports precise spatial selection of targets which can't be verbally referenced easily. This paper focuses on the interaction aspect of speech and raycast dual-modality systems to study the performance and experience of users in VR object selection tasks, which complements previous works on speech and multimodal interaction.

## 2.3 Language Models and Applications in 3DUI

Transformer-based language models enable intent/entity recognition from natural language input. BERT [16] achieved state-of-the-art performance and has been used in joint intent-slot models [12], speech-based classification [22], and named entity recognition [7, 47]. BERT offers customizability and interpretability via intermediate outputs.

In contrast to transformer-based NLP models like BERT, general-purpose large language models (LLMs) like GPT-4 [34] can generate human-like responses. Classic multimodal exemplars such as "Put-That-There" [4] rely on speech but were restricted by technological limitations at the time. More recent works explored the integration of natural language interfaces and spatial interaction within immersive environments. For instance, De La Torre et al. [15] introduced LLMR, a system enabling real-time prompting of interactive worlds through large language models, while Giunchi et al. [18] proposed DreamCodeVR, which empowers users to design object behaviors in VR using voice-driven programming. Yin et al. [61] developed a text-to-scene generation system for VR, highlighting the growing relevance of semantic input in immersive creation workflows. Tang et al. [50] provided a comprehensive review of LLMs in extended reality (XR), underscoring the increasing maturity of such techniques for XR applications. These systems show that LLM-driven interfaces can augment or replace traditional controller input in immersive environments, especially for high-level interaction tasks.

However, as LLM-based systems face issues such as low transparency [65] and limited customizability [11], this paper does not adopt the state-of-the-art general-purpose LLMs to process user speech input. Instead, we follow Chen et al. [10] and adopt an off-the-shelf Azure Conversational Language Understanding (CLU) model[1], a customizable language model to process natural language input from the user, as the model is easy to train and evaluate, while providing high accuracy and greater control in intent and entity recognition tasks. The structured output from the model also allows us to seamlessly integrate the CLU model with the interactive system in the 3D scene.

In this paper, we demonstrate how lightweight NLP tools can enhance occluded object selection in VR and compare it against a mini-map approach to tease out the strengths and weaknesses of both techniques to propose design recommendations, thereby contributing to the research community's understanding of object selection tasks in VR.

# 3 Interaction Techniques

## 3.1 Disocclusion Mini-Map (DiscPIM)

DiscPIM is designed to address the common problem of selecting occluded or hidden objects in virtual reality environments. Instead of relying solely on the main first-person view, DiscPIM displays objects within a cone-shaped highlighter on a flat mini-map. The location of objects on the mini-map is updated in real-time by projecting the original object position on the coordinate system of the hand which holds the cone-shaped highlighter. This allows users to see objects further behind in the cone-shaped highlighter which may be occluded by objects in front.

To select an object, users can press and hold the left controller trigger button. This allows the mini-map to be updated with objects currently within the cone-shape highlighter attached to the left controller. Upon releasing the trigger button, the mini-map "freezes" and objects can be selected by hovering the right controller over objects on the mini-map and by pressing the right controller grabbing button to select. If multiple objects overlap on the mini-map, the user can hover the right controller over these overlapping objects

[1]Source: https://azure.microsoft.com/en-us/products/ai-services/conversational-language-understanding.

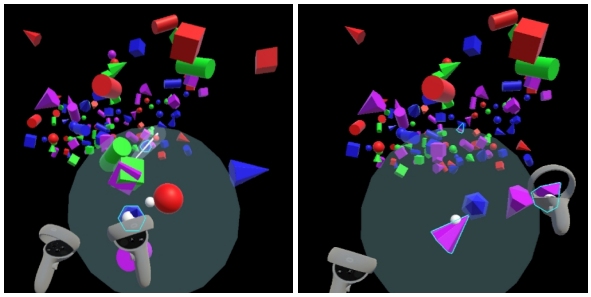

**Figure 2: Selecting objects with DiscPIM directly from the mini-map (left) and the mini-map circumference (right).**


**Figure 3: Draggable panel of AssistVR which displays the recognized speech and a list of names and 3D previews of all selected objects. Here, the user states 'Select all purple spheres', and all four purple spheres in the scene are selected.**

and press the right controller trigger button to "freeze" the overlapping objects in slots on the circumference of the mini-map. In this case, the objects can be selected by hovering the right controller on objects on the circumference and by pressing the right grabbing button to select. Figure 2 provides an illustration of how objects can be selected with DiscPIM directly from the mini-map (left) and from the mini-map circumference (right).

## 3.2 Speech-and-Pointing (AssistVR)

AssistVR is a dual-modality selection technique combining a speech-based interface powered by Microsoft Azure CLU with a raycast-based selection method. We detail the implementation below.

*3.2.1 Speech-based Selection.* We adopted the AssistVR workflow proposed by Chen et al. [10], except for the following differences: (1) Only Azure Conversational Language Understanding (CLU) is involved, and GPT-4o is not included in the workflow; (2) The predefined intents include 'Select', 'CancelAll', and 'None', while the key entities include 'Original Color' and 'Original Shape'.

In the training phase, the first step is to select data and define *intents* and *entities*. As the task focuses on object selection, we focused on the following three intents: 'Select', 'CancelAll', and 'None'. The second step is to create the training set by suggesting typical utterances for each intent and label all entities that appear in each utterance. For example, an utterance such as: 'Select the purple cube.' would belong to the 'Select' intent, with 'purple' labeled as 'Original Color' and 'cube' labeled as 'Original Shape'. The entire dataset consists of 51 utterances for the 'Select' intent, and 6 utterances for the 'CancelAll' intent. The complete set of utterances together with the labeled intents and entities for the training and test set can be found as a JSON file[2] in the 'Supplemental Materials' folder on OSF. Subsequently, the model is trained, and several rounds of improvement are made. Finally, the model is deployed on the Azure cloud server.

In the deployment phase, the user presses a button on the controller and speaks to the system via the VR headset. The Azure text-to-speech service then transcribes audio input into text. Requests are posted by Unity via the Azure API to upload the recognized user speech input to the model. The model then outputs the prediction in JSON format, which contains information on the most likely intent, recognized entities, and their confidence scores. Based on this model output data, a Unity C# post-processing script parses all selectable objects and retrieves the object material name. For commands with intent 'Select', If the user command contains only the 'Original Color' or the 'Original Shape' entity, all objects whose material contains the color or shape property are selected. If the user command contains both the 'Original Color' and 'Original Shape' entity, all objects whose material contains the color and shape property are selected. If the intent is 'CancelAll', all objects are deselected. If the intent is 'None', no action is performed.

Throughout the object selection process, all selected objects are displayed on a draggable panel (Figure 3) and are highlighted in green, both on the panel and in the 3D scene, to provide visual feedback to the user. When objects are deselected, they are removed from the panel. In addition to selected objects, the draggable panel also displays the recognized speech of the user. If the speech is not recognized, this is also conveyed to the user via the panel.

*3.2.2 Raycast-based Selection.* As speech-based interfaces have the limitation of being prone to errors [49, 57] and usability tends to improve if speech-based interfaces are used in conjunction with traditional interfaces [4, 49], we complement speech-based selection with a raycast technique to enable users to make fine-grained and precise selections/deselections. Here, the speech and raycast modalities are used separately without information exchange between them, hence the term *dual*-modality. Users are able to select/deselect objects hit by the ray which is cast from the right controller by pressing the trigger button. Selected objects are highlighted in green, while deselected objects are highlighted in red. When the ray moves away from a deselected object, the red highlight is removed. For selected objects, the green highlight is preserved even when the ray is directed away. As with objects selected using speech, objects selected using raycast also appear on the draggable panel with their names and green outlines.

[2]The JSON file can be imported as a Conversational Language Understanding project to Azure to recover the project.

# 4 Study Design

The user study aims to evaluate and compare user performance, user experience, and usability of the speech and raycast dual-modality 3D object selection technique (AssistVR) and the disocclusion mini-map technique (DiscPIM) under different scene perplexity conditions and different numbers of target objects to identify performance crossover points and propose design recommendations. This study design follows established works [21, 29] to study both subjective workload and measured performance under scene setups with different visual complexity conditions which can yield different results. In our study, "scene perplexity" refers to whether or not users are familiar with the the object category and object property. A detailed explanation is provided in Section 4.1. This section introduces the design of the user study, which follows recommendations and guidelines proposed by Bergström et al. [3]. For example, we choose selection speed as the main dependent variable, recruit 24 participants, use a fixed starting position, and use three independent variables (technique, number of target objects, and scene perplexity condition). To facilitate comparison with the baseline technique DiscPIM [29], several elements of the experiment setup (such as the red/green button in front of the user, the yellow search region, the black background, and the search space dimensions) follow the setup proposed by Maslych et al. [29]. As during all study trials for DiscPIM [29] the target object was a magenta sphere and none of the distractor objects shared the same appearance as the target, we followed this setup to make the target object visually distinguishable from other distractor objects and make its name known to the user prior to the study.

## 4.1 Design

We adopt a within-subjects design to evaluate the performance of AssistVR [10] and DiscPIM [29] in three levels of scene perplexity conditions (Low, Medium, and High), as well as three levels of target objects (1Target, 2Target, and 4Target). The scene perplexity is reflected by the number of object categories and object properties which can easily be referenced verbally. Examples of the Low, Medium, and High scene perplexity conditions can be found in Figure 1. In the user study, there are eight different object categories (of which four were easily recognizable by users and four were more difficult) and eight different textures to reflect different object properties (four were easily recognizable and four were more difficult). There are considerable differences between the known and unknown object categories and textures—a study with 21 participants prior to the study provides statistical evidence for the categorization of known and unknown objects' shapes and colors. Details of the survey findings can be found in the online appendix. Before the study, the names of all objects and textures are briefed to the participant to simulate real-world use cases where users have access to object names when using a speech-based system.

The Low perplexity condition consists of four easily-identifiable object shapes (cube, sphere, cylinder, and pyramid) and four easily-identifiable colors (purple, blue, green, and red). The Medium perplexity condition consists of two easily-identifiable object shapes (cube and sphere) and two object shapes which are more difficult to identify (barrel and pyramid cuboid), as well as two easily-identifiable colors (purple and blue) and two colors that are more difficult to identify (purple pattern and white pattern). The High perplexity condition consisted of four object shapes which are difficult to identify (barrel, cross, pyramid cuboid, and truncated cylinder) and four colors which are difficult to identify (purple pattern, white pattern, yellow pattern, and blue pattern). The order of the two techniques and the order of the perplexity conditions are both counterbalanced across all 24 participants.

The object density of distractor objects remains the same throughout all experiments. In each scene perplexity level, 120 distractor objects are scattered within the environment of 20 meters in depth, 10 meters in width, and 5 meters in height in front of the user. Including the varying number of 1, 2, or 4 target objects, the total number of objects is 121, 122, or 124. The object shape and color of all objects depends on the scene perplexity level. Within each perplexity level, the shape, color, position, and orientation of distractor objects remains the same for different target objects.

## 4.2 Task

The task involves using the speech and raycast dual-modality technique (AssistVR), or the baseline mini-map occluded object selection technique DiscPIM [29] to select different numbers of target objects (1, 2, or 4 targets) among a set of selectable objects. For AssistVR, participants had the freedom to use only the speech modality, or only the raycast modality, or a combination of speech and raycast in each trial. First-person views of both techniques in the task are shown in Figure 1. At the beginning of each trial, the user directs both the left and right raycast at a red button in front of the user[3]. Upon pressing both triggers simultaneously to start the trial, the red button becomes green, the timer starts, and a home object[4] encapsulated within a semi-transparent sphere appears to the left of the user. At the same time, the target object(s), a yellow search region, together with 120 distractor objects appear within the 10m×5m×20m search space in front of the user. The shape and color of the target object(s) are randomly drawn from the set of 4 object shapes and 4 colors determined by the scene perplexity condition. Next, the participant uses the object selection technique to select a certain number of targets specified by the NumTargets condition. If an error occurs, participants are allowed to deselect and select again. Finally, they press Button "B" on the right controller to confirm the selection. This stops the timer for trial completion time, and participants move on to the next trial if the selection is correct, that is, if and only if all target objects are selected and all selected objects are targets.

The time difference between starting the trial and users making the confirmation is recorded as the trial completion time. Following prior work[29, 64], users are asked to repeat the same selection under the exactly same conditions after completion of the first selection trial. In this paper, we refer to these two tasks as the *search* trial and the *repeat* trial.

In summary, the study for each participant consists of 18 combinations of independent variables (2 Techniques × 3 Perplexities × 3 NumTargets). Participants complete the first half of the study with one technique, then complete the first half of the questionnaire,

[3]This step ensures that the user's raycast direction remains the same at the beginning of all trials [29]

[4]The home object is identical to the target object, which also serves to inform the user which object to select.

before moving on to the second technique and the second half of the questionnaire. Within each technique, participants complete trials under all three scene perplexities. Within each Perplexity condition, participants complete all three NumTargets conditions. Within each combination of Technique, Perplexity, and NumTargets, the participant completes three sets of one *search* trial followed by one *repeat* trial.

### 4.3 Participants and Apparatus

An a priori power analysis using G*Power 3.1 with an effect size of 0.25, $\alpha = 0.05$, a desired power of 0.8, a correlation of 0.5 among repeated measures, and nonsphericity correction $\varepsilon = 1$ revealed a sample size of 34 for 2 measurements to explain the Technique main effect and a sample size of 19 for 6 measurements to explain the Technique×NumTargets and Technique×Perplexity interaction effects, and a sample size of 10 for 18 measurements to explain the Technique×NumTargets×Perplexity interaction effect. The final sample size follows object selection study guidelines [3] and is a compromise between feasibility, statistical power, and counterbalancing requirements.

We recruited 24 participants (16 males and 8 females) aged between 18 and 33 ($M = 24.3 \pm 4.46$). All participants were right-handed and had normal or corrected-to-normal vision. Around 50% of participants were familiar with head-mounted VR. All participants understood and spoke English, with around 58% native English speakers. None of the participants reported any known visual, auditory, or physical disability. During the experiment, participants wore an Oculus Quest 2 headset and held the left and right controllers. The headset was connected to a Windows 10 laptop PC (Intel i5-9300H CPU, 16GB memory, and GTX 1050 graphics card) via cable. Virtual scenes were implemented with Unity 3D (Version 2020.3.47f1) and publicly available online asset resources.

### 4.4 Procedure

Before the study, participants were asked to review an information sheet and sign a consent form. After the collection of basic demographic information, participants were then asked to familiarize themselves with the Oculus Quest 2 headset and controllers. Next, we provided an overview of the study procedure and the tasks they were asked to complete using images and verbal introduction, which included explaining that the experiment will consist of using two techniques to perform object selection, and each technique consisted of three perplexity conditions, where each combination of Technique and Perplexity consisted of three sets of a *search* trial and a *repeat* trial. We then demonstrated the usage of the two techniques to each participant. Participants had the opportunity to ask questions. Subsequently, participants had the opportunity to practice completing the trials with the two techniques under the Medium Perplexity condition.

After familiarization, participants completed three sets of *search* and *repeat* trials for each perplexity level and each number of targets using one of the two techniques, before having a five-minute break and completing another three sets for each perplexity level and each number of targets with the other technique. Throughout the experiment, the order of techniques was alternated across participants. For each technique, each participant completed three blocks each with a different perplexity level. Within each block, participants completed three NumTargets conditions. Within each combination of Technique×NumTargets×Perplexity, participants completed three sets of *search-repeat* trials. Across all 24 participants, the sequence of NumTargets was rotated within each block and the sequence of Perplexity was rotated across different blocks. The complete study sequence is provided in the online appendix. In each trial, participants completed a *search* task followed by a *repeat* task. Specifically, they pointed the left and right raycast at the 'Start' button and pressed the left and right trigger simultaneously to start a 3-second countdown to start the *search* task. Users use AssistVR or DiscPIM [29] to select the target object(s). If the selection is correct, the timer will stop and the scene will reset. Users will then need to trigger the 'Start' button again to begin the *repeat* task. If the selection is incorrect, the system will play a tone to prompt the user to try again, and the total number of attempts will be recorded. In the subsequent *repeat* task, the procedure is the same, except that the object positions are exactly the same as in the *search* task, which are no longer randomly generated. After the *search* and *repeat* tasks, a new target object is drawn and placed at a new position, but the distractor objects remain at the same position.

After completing all trials for each technique, the participants were asked to complete SUS [6] and NASA-TLX [20] surveys. Following prior work [64], perceived user experience was measured using the short version of User Experience Questionnaire (UEQ-S) [43] on a 7-point scale. The surveys were followed by a five-minute break before progressing to the next technique. After completing trials for both techniques, participants were asked to rank the techniques based on their overall preference and provide feedback and comments on the features they preferred and disliked. The study took around 1.5 hours in total, and participants were compensated with a £15 Amazon voucher for their time. The study is approved by the ethics committee in the Department of Engineering at the University of Cambridge.

## 5 Results

Statistical significance tests on trial completion time were carried out using a repeated measures analysis of variance (RM-ANOVA) with Holm-Bonferroni adjustments for the post-hoc tests. Task load, system usability, and user experience ratings were analyzed with non-parametric Wilcoxon signed-rank tests.

### 5.1 Trial Completion Time

During the study, we recorded the trial completion time as an indicator for object selection performance for both *search* and *repeat* tasks under all combinations of Technique, Perplexity, and NumTargets conditions as a quantitative measure of user performance in the object selection task. Across all object selection trials, participants spent an average of 12.8 minutes ($SD = 3.44$) with AssistVR and an average of 17.5 minutes with DiscPIM ($SD = 4.72$). In total, 2592 data points were collected (24 participants × 2 Techniques × 3 Perplexities × 3 NumTargets × 2 tasks × 3 repetitions). In line with prior work [29], we removed 32 outlier data points (1.23%) where the trial completion time was more than 4 standard deviations away from the mean in each condition. We did not discard trials which took participants more than one attempt to complete.

**Table 1: RM-ANOVA results for the trial completion time of both the *search* and *repeat* task after applying the Aligned Rank Transform. Gray rows show significant findings. T = Interaction Technique, N = Number of Targets, P = Scene Perplexity.**

| | Search Task | | | | | Repeat Task | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $df_1$ | $df_2$ | F | p | $\eta_p^2$ | $df_1$ | $df_2$ | F | p | $\eta_p^2$ |
| T | **1** | **391** | **48.012** | **< .001** | **.11** | **1** | **391** | **133.628** | **< .001** | **0.25** |
| N | **2** | **391** | **144.421** | **< .001** | **.42** | **2** | **391** | **282.899** | **< .001** | **0.59** |
| P | **2** | **391** | **21.596** | **< .001** | **.10** | **2** | **391** | **31.541** | **< .001** | **0.14** |
| T × N | **2** | **391** | **67.158** | **< .001** | **.26** | **2** | **391** | **85.130** | **< .001** | **.30** |
| T × P | 2 | 391 | .193 | .825 | < .001 | 2 | 391 | .548 | .579 | .003 |
| T × N × P | 4 | 391 | 1.363 | .246 | .01 | 4 | 391 | .790 | .532 | .008 |

As each participant is exposed to all conditions, a repeated-measures ANOVA (RM-ANOVA) test was conducted on both the *search* and *repeat* trial completion time data to determine whether significant differences existed in trial completion time across different conditions. The trial completion time was not normally distributed (Shapiro-Wilk $p < .001$), so Aligned Rank Transform [59] was applied before conducting RM-ANOVA. Search and repeat trial times were analyzed separately, and ART-C [59] was applied before running pairwise comparisons with Bonferroni adjustment.

Table 1 presents the RM-ANOVA results on *search* and *repeat* trial completion time for the independent variables Technique, NumTargets, and Perplexity, together with interaction terms. Degrees of freedom and effect sizes are derived and reported based on results from the ART-C procedure [59] and may differ from classical RM-ANOVA values.

*5.1.1 Main Effect of Technique.* Results revealed a significant main effect of Technique on the *search* ($F_{1,391} = 48.012$, $\eta_p^2 = .11$, $p < .001$) and *repeat* ($F_{1,391} = 133.628$, $\eta_p^2 = .25$, $p < .001$) trial completion time. Post-hoc tests with Bonferroni adjustment suggested that participants took significantly less time ($p < .001$) to complete the *search* task using AssistVR ($M = 16.9$, $SD = 9.79$) as opposed to using DiscPIM ($M = 22.1$, $SD = 16.3$). For the *repeat* task, participants also took significantly less time ($p < .001$) with AssistVR ($M = 10.1$, $SD = 5.73$) compared with using DiscPIM ($M = 14.3$, $SD = 9.33$). Here, results for the main effect of Technique are averaged over NumTargets and Perplexity.

*5.1.2 Main Effect of NumTargets.* Results revealed a significant main effect of NumTargets on the *search* ($F_{2,391} = 144.421$, $\eta_p^2 = .42$, $p < .001$) and *repeat* ($F_{2,391} = 282.899$, $\eta_p^2 = .59$, $p < .001$) trial completion time. Post-hoc tests with Bonferroni adjustment revealed that for the *search* task, significant differences ($p < .001$) existed between all pairwise comparisons of the 1Target ($M = 11.9$, $SD = 7.02$), 2Targets ($M = 18.7$, $SD = 11.8$), and 4Targets ($M = 27.9$, $SD = 15.7$) conditions. For the *repeat* task, significant differences ($p < .001$) also existed between all pairwise comparisons of the 1Target ($M = 7.09$, $SD = 3.32$), 2Targets ($M = 11.5$, $SD = 5.29$), and 4Targets ($M = 17.9$, $SD = 9.72$) conditions.

*5.1.3 Main Effect of Perplexity.* Results revealed a significant main effect of Perplexity on the *search* ($F_{2,391} = 21.596$, $\eta_p^2 = .10$, $p < .001$) and *repeat* ($F_{2,391} = 31.541$, $\eta_p^2 = .14$, $p < .001$) trial completion time. Post-hoc tests with Bonferroni adjustment also revealed that *search* trial completion time was significantly different between the High perplexity condition ($M = 23.4$, $SD = 16.1$) and the Low perplexity condition ($M = 16.5$, $SD = 11.2$) ($p < .001$), between the High and Medium perplexity condition ($M = 18.6$, $SD = 12.3$) ($p < .05$), but not between the Low and Medium perplexity condition ($p = .060$). *Repeat* trial completion time was significantly different between the High perplexity condition ($M = 14.4$, $SD = 9.58$) and the Low perplexity condition ($M = 10.4$, $SD = 6.75$) ($p < .001$), between the High and Medium perplexity condition ($M = 11.7$, $SD = 6.90$) ($p < .001$), as well as between the Low and Medium perplexity condition ($p < .05$).

*5.1.4 Interaction Effect of Technique × NumTargets.* Figure 4 (left) shows the *search* and *repeat* trial completion time of different Technique and NumTargets combinations. RM-ANOVA tests revealed a significant interaction effect of Technique × NumTargets on the *search* ($F_{2,391} = 67.158$, $\eta_p^2 = .26$, $p < .001$) and *repeat* ($F_{2,391} = 85.130$, $\eta_p^2 = .30$, $p < .001$) trial completion time. For the AssistVR technique, post-hoc tests with Bonferroni adjustment did not reveal significant differences in *search* trial completion time between the 1Target and 2Targets condition ($p = 1.0$) or the 2Targets and 4Targets condition ($p = .467$), but did reveal significant differences between the 1Target and 4Targets condition ($p < .05$). Significant differences were also found in *repeat* trial completion time between the 1Target and 2Targets ($p < .05$) and 1Target and 4Targets ($p < .001$) conditions, as well as between the 2Targets and 4Targets ($p < .001$) condition.

For the DiscPIM technique, post-hoc tests revealed a significant difference in *search* and *repeat* trial completion time when the number of targets varied ($p < .001$ for all pairwise comparisons of NumTargets conditions in both *search* and *repeat* tasks).

Comparing the trial completion time between AssistVR and DiscPIM under different NumTargets conditions, post-hoc comparisons revealed that in the *search* task, DiscPIM ($M = 9.34$, $SD = 4.62$) was significantly faster than AssistVR ($M = 14.5$, $SD = 8.03$) under the 1Target condition ($p < .001$), but AssistVR ($M = 19.8$, $SD = 11.6$) was significantly faster than DiscPIM ($M = 35.9$, $SD = 15.1$) under the 4Targets condition ($p < .001$). The difference in *search* trial completion time between both techniques is not significant under the 2Target condition ($p = .077$). In the *repeat* task, AssistVR ($M = 9.96$, $SD = 5.02$) was significantly faster than DiscPIM ($M = 13.0$, $SD = 5.13$) under the 2Targets condition ($p < .001$). AssistVR ($M = 12.5$, $SD = 6.81$) was also significantly faster than DiscPIM ($M = 23.3$, $SD = 9.23$) under the 4Targets condition ($p < .001$). However, AssistVR ($M = 7.74$, $SD = 4.01$) was significantly slower than DiscPIM ($M = 6.43$, $SD = 2.30$) in the 1Target condition ($p < .05$).

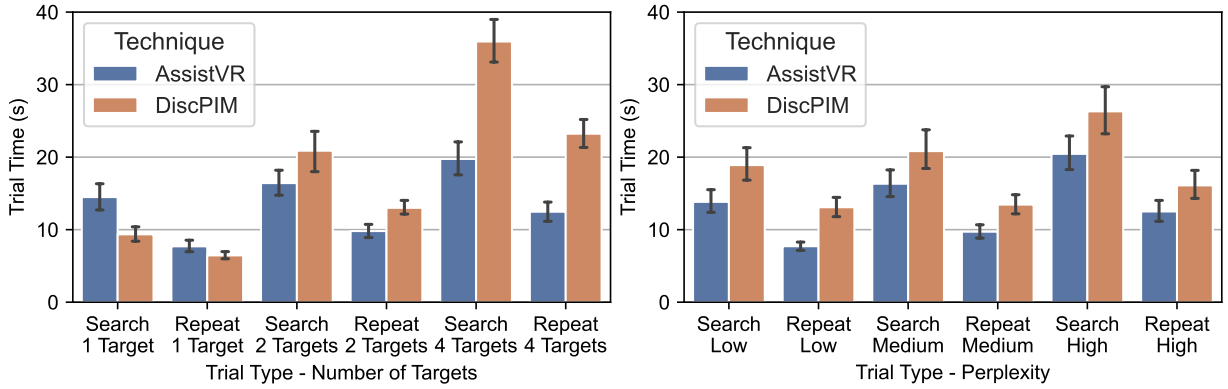


**Figure 4: Trial completion time (seconds) for each technique across *search* and *repeat* trial types for each NumTargets condition (left) and each Perplexity condition (right), with 95% confidence intervals of the mean estimate.**

**Table 2: Wilcoxon signed-rank test pairwise comparison results of NASA-TLX scores, SUS scores, and UEQ-S scores and their subcategories (if any) between both Techniques. Gray rows show significant findings. The Wilcoxon Statistic $W$, statistical significance $p$, and effect size $r$ are reported, where $r = z/\sqrt{N}$. A negative effect size indicates that the mean rating of AssistVR is higher than that of DiscPIM.**

| NASA-TLX Scores | | | | SUS Scores | | | | UEQ-S Scores | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | $W$ | $p$ | $r$ | | $W$ | $p$ | $r$ | | $W$ | $p$ | $r$ |
| Overall | **59.5** | **< .05** | **.461** | Overall | 157 | .853 | -0.038 | Overall | 106.5 | .346 | .197 |
| Mental | 76.0 | .103 | .348 | | | | | Pragmatic | 176.5 | .457 | -.152 |
| Physical | **55.5** | **< .05** | **.491** | | | | | Hedonic | **64.5** | **< .05** | **.496** |
| Temporal | **2.0** | **< .05** | **.828** | | | | | | | | |
| Performance | 97.5 | .935 | -.019 | | | | | | | | |
| Effort | **37.5** | **< .05** | **.593** | | | | | | | | |
| Frustration | 119.5 | .903 | -.027 | | | | | | | | |

*5.1.5 Interaction Effect of Technique × Perplexity.* Figure 4 (right) presents bar plots of the *search* and *repeat* trial completion time for different combinations of Technique and Perplexity conditions. RM-ANOVA tests did not reveal a significant interaction effect of Technique × Perplexity on either the *search* ($F_{2,391} = .193$, $\eta_p^2 < .001$, $p = .825$) or *repeat* ($F_{2,391} = .548$, $\eta_p^2 = .003$, $p = .579$) trial completion time.

*5.1.6 Interaction Effect of Technique × NumTargets × Perplexity.* The interaction effect of Technique × NumTargets × Perplexity was not significant on either the *search* ($F_{4,391} = 1.363$, $\eta_p^2 = .01$, $p = .246$) or *repeat* ($F_{4,391} = .790$, $\eta_p^2 = .008$, $p = .532$) trial completion time.

## 5.2 Task Load

A Wilcoxon signed rank test on the NASA-TLX ratings (unweighted version) [20] revealed that the overall task load rating of AssistVR ($M = 4.06$, $SD = 1.71$) was significantly lower ($W = 59.5$, $p < .05$, $r = .461$) than that of DiscPIM ($M = 4.94$, $SD = 1.72$). Results are summarized in Table 2.

## 5.3 System Usability

A Wilcoxon signed rank test on the system usability scale [6] did not reveal any significant differences ($W = 157$, $p = .853$, $r = -.038$) between AssistVR ($M = 71.0$, $SD = 14.4$) and DiscPIM ($M = 68.0$, $SD = 23.0$). Results are summarized in Table 2.

## 5.4 User Experience

Wilcoxon signed rank tests on the short version User Experience Questionnaire (UEQ-S) ratings did not reveal a significant difference ($W = 106.5$, $p = .346$, $r = .197$) in the overall UEQ-S score between AssistVR ($M = .547$, $SD = 1.23$) and DiscPIM ($M = .880$, $SD = 1.07$). For the subcategories of the UEQ-S ratings, significant differences ($W = 64.5$, $p < .05$, $r = .496$) were found in the Hedonic quality between AssistVR ($M = -0.021$, $SD = 1.61$) and DiscPIM ($M = .958$, $SD = 1.16$), but not in the Pragmatic quality ($W = 176.5$, $p = .457$, $r = -.152$) between AssistVR ($M = 1.11$, $SD = 1.21$) and DiscPIM ($M = .802$, $SD = 1.37$). Results are summarized in Table 2.

## 5.5 Overall Preference and Open Comments

In the post-experience questionnaire, we asked participants about their overall preference among the two techniques and invited them to leave comments about features they liked/disliked. Among all 24 participants, 13 preferred AssistVR, while 11 preferred DiscPIM.

For the AssistVR technique, participants liked the fact that it was easy to use (P7), efficient (P10, P20, P21, P22), and allowed 'selecting multiple objects in one go' (P2, P14), and participants could select objects without knowing where the object is (P13, P14,

**Table 3: Empirically derived tradeoffs between AssistVR and DiscPIM.**

| Aspect | Speech-and-Pointing (AssistVR) | Disocclusion Mini-map (DiscPIM) |
|---|---|---|
| Task Completion Time | Significantly faster with 2 targets ($p < .05$) and 4 targets ($p < .001$), also demonstrated in trial completion time of individuals (P10, P20, P21, P22) who reported AssistVR being more efficient. | Faster with single target selection ($p < .05$). Speech recognition errors (P1, P2, P7, P11, P13, P21, P24) and delay in speech processing (P20) were reported for AssistVR. |
| Workload (NASA-TLX) | Significantly lower ($p < .05$). P13, P14, P15, P23 reported AssistVR allows selection without having to know object location. | Significantly higher. Participants found DiscPIM tedious (P1), annoying (P2), and a bit tricky and boring for multi-object selection (P5). |
| System Usability (SUS) | Moderately higher but not significantly different. P7 reported 'easy to use'. | Moderately lower. DiscPIM can be time-consuming (P4), and slow when occlusion is not present (P20, P21). |
| User Experience (UEQ-S) | Moderately worse but not significantly different. AssistVR has a limited number of supported speech commands (P5), requires remembering object names (P12, P13, P14, P19), and is sensitive to command structure (P5, P23). | Moderately better. Participants found DiscPIM fun (P14), intuitive (P7, P12), and more engaging (P4), providing a better sense of control and direct manipulation (P1, P2, P5, P19). |

P15, P23), or moving their hands to execute any action (P12). P19 also found that 'the combination of speech and raycast stroke a nice balance', as raycast was more efficient for selecting one or two visible objects and voice selection helped to select multiple objects.

Participants disliked the fact that AssistVR 'did not support many commands' (P5). Further, speech recognition sometimes failed and the command was not executed correctly (P1, P2, P7, P11, P13, P21, P24). Consequently, the system 'either doesn't select anything or selects wrong objects' (P19), which led to frustration (P7) and loss of trust (P19). Specifically, P8, P12, P21, P22 commented that speech recognition was sensitive to accent, without the capability to auto-correct recognized speech based on the context (P22), which repeatedly led to errors. For example, P12 (non-native English speaker) noted that 'sphere' was often misrecognized as 'Sofia', while P21 (native, Scottish accent) noted that 'cuboid' was often misrecognized as 'keyboard'. Some participants found it somewhat difficult to remember object names (P12, P13, P14, P19). Sometimes participants had to repeat several times before getting the speech command right (P5, P23). P20 also commented that the time it takes to speak and the slight delay in speech processing makes it sometimes faster to engage in manual selection as opposed to using speech. P24 also suggested that the Deselect function could deselect a specific object, rather than deselecting all objects.

For DiscPIM, participants liked the fact that the design of the torch and mini-map felt 'fun' (P14) and 'intuitive' (P7, P12) and the mini-map freezes to provide direct visual feedback (P15, P16), which gives users a better sense of control and direct manipulation (P1, P2, P5, P19), and improves selection accuracy (P18). It 'makes objects far away easier to see' (P23), allowing users to focus on a complicated region with many occluded objects (P10), without the need to know the object's name (P13). P4 also found DiscPIM to be 'more engaging and less repetitive than speech recognition'.

Regarding limitations, participants found DiscPIM to be 'tedious' (P1), 'annoying' (P2), 'tricky and boring' (P5), and 'slower' (P14) when there are many targets. P4 found the technique 'tiring for the eyes' and 'time consuming'. P19 suggested that the mini-map object expansion list could appear closer to the left hand or allow users to customize the position. Otherwise, users would need to raise the head to look at it, making it inconvenient. P20 and P21 commented that the benefits brought by DiscPIM is situational and it could slow down the search process when there is not much occlusion.

## 6 Discussion

Based on existing techniques using ray-based metaphors [14, 17, 25, 29, 32, 58, 60, 64], gestures [45], and eye gaze [9, 46] for object selection in VR, together with works studying speech interaction [24, 28, 57], multimodal interaction [23, 39, 41] in immersive technologies, as well as customizable purpose-built LLMs [11], we have advanced the research community's understanding of the design of systems for object selection in VR by providing a comparative analysis of Speech-and-Pointing (AssistVR) and Disocclusion Mini-Map (DiscPIM) techniques for object selection in virtual reality, revealing important tradeoffs that inform design decisions especially for *occluded* objects. The results highlight that neither technique is universally superior. Instead, their effectiveness depends on task complexity, target quantity, and user cognitive load.

Table 3 summarizes the tradeoffs of both techniques based on quantitative and qualitative data from our empirical user study. The mini-map technique (DiscPIM) enabled faster single-target selection with better user experience, which confirms its suitability for simple tasks where users perform direct manipulation efficiently. Due to lack of timestamped audio onset and action feedback data as well as FPS data, we are unable to quantify processing delays in Speech-and-Pointing (AssistVR) and associate the longer single-target selection times here to a specific cause. As scene perplexity and target counts increased, the cognitive load associated with selecting multiple targets became significant, possibly leading to increased task completion time and perceived workload. Participants were allowed short breaks between different Technique and Perplexity conditions to minimize discomfort. While full session durations were not recorded, selection trials lasted 25-35 minutes in total, which is in line with typical VR object selection studies [3]. However, fatigue may have affected user performance and perceived load, particularly in 4-object selection tasks with DiscPIM, as demonstrated by the high trial completion time in Figure 4 (left). Such fatigue effects should be considered in future studies.

Speech-and-Pointing (AssistVR) maintained lower overall perceived load, supporting users in occluded object selection tasks by leveraging natural language and raycast selections efficiently.

**Table 4: Design recommendations grounded in study results (in square brackets) for occluded object selection techniques in VR.**

| Design Factor | Recommendation |
|---|---|
| Scene Perplexity | Providing easy access to the names of objects can allow Speech-and-Pointing to outperform mini-map techniques even when object names are long and difficult to pronounce [Section 5.1.1; Figure 4 (right)] |
| Number of Targets | Favor DiscPIM for single-target selection and speech-and-pointing for multi-targets [Section 5.1.4] |
| User Cognitive Load | Consider user mental workload; Speech-and-pointing reduces overall load in multi-object selection tasks [Section 5.2] |
| Interaction Training | Provide training on speech commands and pointing; Familiarize users with mini-map selection [Section 5.5] |
| Speech Recognition | Ensure high-quality speech recognition (for example by leveraging contextual information in past conversations and supporting recognition of different accents) to minimize user frustration in Speech-and-Pointing [Section 5.5] |
| Speech Processing | Post-processing of recognized intents and entities should be robust to all types of user inputs [Section 5.5] |
| User Agency | Speech-based interactive systems could ensure visibility of the mapping between speech commands and subsequent actions of the system to allow users to have agency over the system [Section 5.5] |
| User Experience | User experience could be improved via direct manipulation with rich visual feedback such as mini-maps and by making speech interactions more engaging [Section 5.5] |

While it supports significantly faster multi-selection of two or more targets with significantly lower perceived workload and moderately higher usability, users reported moderately worse user experience compared to mini-maps which can be more fun and offer better direct manipulation. Errors with speech recognition and processing can also restrict the performance of speech-based techniques. For example, non-native English speakers (P8, P12, P22) and native speakers with regional accents (P21) highlighted recognition errors, suggesting the need for more inclusive systems to support contextual speech recognition for different accents. Based on these findings, we suggest practical guidelines for VR designers in Table 4.

These guidelines are valuable for real use cases such as selecting occluded anatomical structures in medical training or collaborative 3D architectural design. These scenarios pose challenges due to scene perplexity and number of targets, and require minimal user load, low system latency, and robust speech recognition. We acknowledge that the controlled study conditions such as known object names and same type of targets in multi-selection trials assume an ideal experimental environment but limit the generalizability of findings in complex real-world VR applications, which may involve complicated target object sets and dynamic environments. In such cases, hybrid approaches which adopt speech-and-pointing for rapid, coarse selection of multiple objects, complemented with detailed precise adjustments based on minimap techniques hold the potential for complex selection tasks. Through the preserved spatial relation in minimap-based selection or semantic labels (e.g., 'Select the leftmost mug') in speech and pointing, hybrid systems can generalize to more complex environments and select a subset of identical objects. Such hybrid systems could be helpful in applications such as selecting dynamic moving people or vehicles in emergency response training, or selecting all walls in an indoor scene in creative design. In data visualization, such a hybrid system could also support quick category-based selections (e.g., select all nodes with more than 5 edges). Future research could explore adaptive multimodal systems that combine speech, pointing, and visual aids like DiscPIM based on task context and user preferences. Longitudinal studies may also reveal learning effects and evolving user strategies in a complicated dynamic object selection task.

## 7 Conclusion

In this paper, we present a comparative evaluation of Speech-and-Pointing (AssistVR) and Disocclusion Mini-Map (DiscPIM) techniques for occluded object selection in virtual reality across varying scene perplexities and number of targets. Our findings reveal that each technique offers distinct advantages and challenges depending on task demands. Disocclusion Mini-Maps excel in selection tasks involving fewer targets, providing better user experience and greater sense of control through direct manipulation via the mini-map. In contrast, Speech-and-Pointing supports users better in multi-object selection tasks by leveraging natural language and direct raycast selection for efficient interaction.

Rather than advocating a single best solution, our results emphasize the importance of understanding these tradeoffs to propose recommendations which inform design decisions for occluded object selection in different contexts. We encourage VR practitioners to consider task perplexity, user workload, and target object characteristics (including heterogeneity of target objects and number of target objects) when selecting or combining interaction modalities. Future work could explore adaptive and hybrid systems that dynamically leverage the complementary strengths of these techniques to enhance user performance and experience.

## Supplemental Materials

Supplemental materials can be found at https://osf.io/8g234/. These include: (1) The pre-experiment survey and anonymized results, (2) Questionnaire for the occluded object selection user study, (3) Training and testing data for the Azure CLU system, (4) The online appendix, and (5) The Unity project and analysis code.

## Acknowledgments

This work is supported by Cambridge Trust and the China Scholarship Council. The authors would like to thank the authors of DiscPIM [29] for open-sourcing their code to facilitate the study design and comparison against the baseline technique in this paper.

## References


[1] Ferran Argelaguet and Carlos Andujar. 2013. A survey of 3D object selection techniques for virtual environments. *Computers & Graphics* 37, 3 (2013), 121–136.

[2] Marc Baloup, Thomas Pietrzak, and Géry Casiez. 2019. RayCursor: A 3D Pointing Facilitation Technique based on Raycasting. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3290605.3300331

[3] Joanna Bergström, Tor-Salve Dalsgaard, Jason Alexander, and Kasper Hornbæk. 2021. How to Evaluate Object Selection and Manipulation in VR? Guidelines from 20 Years of Studies. 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 533, 20 pages. doi:10.1145/3411764.3445193

[4] Richard A. Bolt. 1980. "Put-that-there": Voice and gesture at the graphics interface. *SIGGRAPH Comput. Graph.* 14, 3 (July 1980), 262–270. doi:10.1145/965105.807503

[5] Doug A Bowman and Larry F Hodges. 1999. Formalizing the design, evaluation, and application of interaction techniques for immersive virtual environments. *Journal of Visual Languages & Computing* 10, 1 (1999), 37–53.

[6] John Brooke et al. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.

[7] Yuan Chang, Lei Kong, Kejia Jia, and Qinglei Meng. 2021. Chinese named entity recognition method based on BERT. In *2021 IEEE International Conference on Data Science and Computer Application (ICDSCA)*. 294–299. doi:10.1109/ICDSCA53499.2021.9650256

[8] Jeffrey W Chastine, Ying Zhu, and Jon A Preston. 2006. A framework for inter-referential awareness in collaborative environments. In *2006 International Conference on Collaborative Computing: Networking, Applications and Worksharing*. IEEE, 1–5.

[9] Di Laura Chen, Marcello Giordano, Hrvoje Benko, Tovi Grossman, and Stephanie Santosa. 2023. GazeRayCursor: Facilitating Virtual Reality Target Selection by Blending Gaze and Controller Raycasting. In *Proceedings of the 29th ACM Symposium on Virtual Reality Software and Technology* (Christchurch, New Zealand) *(VRST '23)*. Association for Computing Machinery, New York, NY, USA, Article 19, 11 pages. doi:10.1145/3611659.3615693

[10] Junlong Chen, Jens Grubert, and Per Ola Kristensson. 2025. Analyzing Multimodal Interaction Strategies for LLM-Assisted Manipulation of 3D Scenes . In *2025 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*. IEEE Computer Society, Los Alamitos, CA, USA, 206–216. doi:10.1109/VR59515.2025.00045

[11] Jin Chen, Zheng Liu, Xu Huang, Chenwang Wu, Qi Liu, Gangwei Jiang, Yuanhao Pu, Yuxuan Lei, Xiaolong Chen, Xingmei Wang, et al. 2024. When large language models meet personalization: Perspectives of challenges and opportunities. *World Wide Web* 27, 4 (2024), 42.

[12] Qian Chen, Zhu Zhuo, and Wen Wang. 2019. BERT for Joint Intent Classification and Slot Filling. arXiv:1902.10909 [cs.CL] https://arxiv.org/abs/1902.10909

[13] Leigh Clark, Philip Doyle, Diego Garaialde, Emer Gilmartin, Stephan Schlögl, Jens Edlund, Matthew Aylett, João Cabral, Cosmin Munteanu, Justin Edwards, et al. 2019. The state of speech in HCI: Trends, themes and challenges. *Interacting with computers* 31, 4 (2019), 349–371.

[14] Gerwin de Haan, Michal Koutek, and Frits H. Post. 2005. IntenSelect: using dynamic object rating for assisting 3D object selection. In *Proceedings of the 11th Eurographics Conference on Virtual Environments* (Aalborg, Denmark) *(EGVE'05)*. Eurographics Association, Goslar, DEU, 201–209.

[15] Fernanda De La Torre, Cathy Mengying Fang, Han Huang, Andrzej Banburski-Fahey, Judith Amores Fernandez, and Jaron Lanier. 2024. LLMR: Real-time Prompting of Interactive Worlds using Large Language Models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 600, 22 pages. doi:10.1145/3613904.3642579

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805* (2018).

[17] Jenny Gabel, Susanne Schmidt, Oscar Ariza, and Frank Steinicke. 2023. Redirecting Rays: Evaluation of Assistive Raycasting Techniques in Virtual Reality. In *Proceedings of the 29th ACM Symposium on Virtual Reality Software and Technology* (Christchurch, New Zealand) *(VRST '23)*. Association for Computing Machinery, New York, NY, USA, Article 38, 11 pages. doi:10.1145/3611659.3615716

[18] Daniele Giunchi, Nels Numan, Elia Gatti, and Anthony Steed. 2024. DreamCodeVR: Towards Democratizing Behavior Design in Virtual Reality with Speech-Driven Programming . In *2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*. IEEE Computer Society, Los Alamitos, CA, USA, 579–589. doi:10.1109/VR58804.2024.00078

[19] Emanuela Guglielmi, Giovanni Rosa, Simone Scalabrino, Gabriele Bavota, and Rocco Oliveto. 2024. Help Them Understand: Testing and Improving Voice User Interfaces. *ACM Transactions on Software Engineering and Methodology* 33, 6 (2024), 1–33.

[20] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[21] Yahya Hmaiti, Mykola Maslych, Amirpouya Ghasemaghaei, Ryan K Ghamandi, and Joseph J. LaViola. 2024. Visual Perceptual Confidence: Exploring Discrepancies Between Self-reported and Actual Distance Perception In Virtual Reality. *IEEE Transactions on Visualization and Computer Graphics* 30, 11 (Nov. 2024), 7245–7254. doi:10.1109/TVCG.2024.3456165

[22] Yidi Jiang, Bidisha Sharma, Maulik Madhavi, and Haizhou Li. 2021. Knowledge Distillation from BERT Transformer to Speech Transformer for Intent Classification.

[23] Joo Chan Kim, Teemu H Laine, and Christer Åhlund. 2021. Multimodal interaction systems based on internet of things and augmented reality: A systematic literature review. *Applied Sciences* 11, 4 (2021), 1738.

[24] Yelim Kim, Mohi Reza, Joanna McGrenere, and Dongwook Yoon. 2021. Designers Characterize Naturalness in Voice User Interfaces: Their Goals, Practices, and Challenges. 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 242, 13 pages. doi:10.1145/3411764.3445579

[25] Marcel Kruger, Tim Gerrits, Timon Romer, Torsten Kuhlen, and Tim Weissker. 2024. IntenSelect+: Enhancing Score-Based Selection in Virtual Reality . *IEEE Transactions on Visualization & Computer Graphics* 30, 05 (May 2024), 2829–2838. doi:10.1109/TVCG.2024.3372077

[26] Nianlong Li, Teng Han, Feng Tian, Jin Huang, Minghui Sun, Pourang Irani, and Jason Alexander. 2020. Get a Grip: Evaluating Grip Gestures for VR Input using a Lightweight Pen. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3313831.3376698

[27] Yuan Li, Ibrahim A Tahmid, Feiyu Lu, and Doug A Bowman. 2022. Evaluation of pointing ray techniques for distant object referencing in model-free outdoor collaborative augmented reality. *IEEE Transactions on Visualization and Computer Graphics* 28, 11 (2022), 3896–3906.

[28] Kate Lister, Tim Coughlan, Francisco Iniesto, Nick Freear, and Peter Devine. 2020. Accessible conversational user interfaces: Considerations for design. In *Proceedings of the 17th International Web for All Conference* (Taipei, Taiwan) *(W4A '20)*. Association for Computing Machinery, New York, NY, USA, Article 5, 11 pages. doi:10.1145/3371300.3383343

[29] Mykola Maslych, Yahya Hmaiti, Ryan Ghamandi, Paige Leber, Ravi Kiran Kattoju, Jacob Belga, and Joseph J. LaViola. 2023. Toward Intuitive Acquisition of Occluded VR Objects Through an Interactive Disocclusion Mini-map . In *2023 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*. IEEE Computer Society, Los Alamitos, CA, USA, 460–470. doi:10.1109/VR55154.2023.00061

[30] Mykola Maslych, Difeng Yu, Amirpouya Ghasemaghaei, Yahya Hmaiti, Esteban Segarra Martinez, Dominic Simon, Eugene M. Taranta, Joanna Bergstrom, and Joseph J. LaViola. 2024. From Research to Practice: Survey and Taxonomy of Object Selection in Consumer VR Applications . In *2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE Computer Society, Los Alamitos, CA, USA, 990–999. doi:10.1109/ISMAR62088.2024.00115

[31] Mark R Mine. 1995. Virtual environment interaction techniques. *UNC Chapel Hill CS Dept* (1995).

[32] Joao Monteiro, Daniel Mendes, and Rui Rodrigues. 2023. TouchRay: Towards Low-effort Object Selection at Any Distance in DeskVR . In *2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE Computer Society, Los Alamitos, CA, USA, 999–1005. doi:10.1109/ISMAR59233.2023.00116

[33] Alex Olwal and Steven Feiner. 2005. Interaction techniques using prosodic features of speech and audio localization. In *Proceedings of the 10th International Conference on Intelligent User Interfaces* (San Diego, California, USA) *(IUI '05)*. Association for Computing Machinery, New York, NY, USA, 284–286. doi:10.1145/1040830.1040900

[34] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2024. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL] https://arxiv.org/abs/2303.08774

[35] Sharon Oviatt. 1999. Ten myths of multimodal interaction. *Commun. ACM* 42, 11 (1999), 74–81.

[36] Sharon Oviatt, Rachel Coulston, and Rebecca Lunsford. 2004. When do we interact multimodally? Cognitive load and multimodal communication patterns. In *Proceedings of the 6th International Conference on Multimodal Interfaces* (State College, PA, USA) *(ICMI '04)*. Association for Computing Machinery, New York, NY, USA, 129–136. doi:10.1145/1027933.1027957

[37] Julian Petford, Miguel A. Nacenta, and Carl Gutwin. 2018. Pointing All Around You: Selection Performance of Mouse and Ray-Cast Pointing in Full-Coverage Displays. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3173574.3174107

[38] Duc-Minh Pham and Wolfgang Stuerzlinger. 2019. Is the Pen Mightier than the Controller? A Comparison of Input Devices for Selection in Virtual and Augmented Reality. In *Proceedings of the 25th ACM Symposium on Virtual Reality Software and Technology* (Parramatta, NSW, Australia) *(VRST '19)*. Association for Computing Machinery, New York, NY, USA, Article 35, 11 pages. doi:10.1145/

3359996.3364264

[39] Stéphanie Philippe, Alexis D Souchet, Petros Lameras, Panagiotis Petridis, Julien Caporal, Gildas Coldeboeuf, and Hadrien Duzan. 2020. Multimodal teaching, learning and training in virtual reality: a review and case study. *Virtual Reality & Intelligent Hardware* 2, 5 (2020), 421–442.

[40] Ivan Poupyrev and Tadao Ichikawa. 1999. Manipulating objects in virtual worlds: Categorization and empirical evaluation of interaction techniques. *Journal of Visual Languages & Computing* 10, 1 (1999), 19–35.

[41] Ismo Rakkolainen, Ahmed Farooq, Jari Kangas, Jaakko Hakulinen, Jussi Rantala, Markku Turunen, and Roope Raisamo. 2021. Technologies for multimodal interaction in extended reality—a scoping review. *Multimodal Technologies and Interaction* 5, 12 (2021), 81.

[42] Leah M Reeves, Jennifer Lai, James A Larson, Sharon Oviatt, TS Balaji, Stéphanie Buisine, Penny Collings, Phil Cohen, Ben Kraal, Jean-Claude Martin, et al. 2004. Guidelines for multimodal user interface design. *Commun. ACM* 47, 1 (2004), 57–59.

[43] Martin Schrepp, Andreas Hinderks, and Jörg Thomaschewski. 2017. Design and evaluation of a short version of the user experience questionnaire (UEQ-S). *International Journal of Interactive Multimedia and Artificial Intelligence, 4 (6), 103-108.* (2017).

[44] Felix Schüssel, Frank Honold, and Michael Weber. 2013. Influencing factors on multimodal interaction during selection tasks. *Journal on Multimodal User Interfaces* 7 (2013), 299–310.

[45] Rongkai Shi, Jialin Zhang, Yong Yue, Lingyun Yu, and Hai-Ning Liang. 2023. Exploration of Bare-Hand Mid-Air Pointing Selection Techniques for Dense Virtual Reality Environments. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 109, 7 pages. doi:10.1145/3544549.3585615

[46] Ludwig Sidenmark, Christopher Clarke, Xuesong Zhang, Jenny Phu, and Hans Gellersen. 2020. Outline Pursuits: Gaze-assisted Selection of Occluded Objects in Virtual Reality. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3313831.3376438

[47] Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2019. Portuguese named entity recognition using BERT-CRF. *arXiv preprint arXiv:1909.10649* (2019).

[48] Anthony Steed and Chris Parker. 2005. Evaluating effectiveness of interaction techniques across immersive virtual environmental systems. *Presence* 14, 5 (2005), 511–527.

[49] Bernhard Suhm, Brad Myers, and Alex Waibel. 2001. Multimodal error correction for speech user interfaces. *ACM transactions on computer-human interaction (TOCHI)* 8, 1 (2001), 60–98.

[50] Yiliu Tang, Jason Situ, Andrea Yaoyun Cui, Mengke Wu, and Yun Huang. 2025. LLM Integration in Extended Reality: A Comprehensive Review of Current Trends, Challenges, and Future Perspectives. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 1054, 24 pages. doi:10.1145/3706598.3714224

[51] Laura A Thompson and Dominic W Massaro. 1986. Evaluation and integration of speech and pointing gestures during referential understanding. *Journal of experimental child psychology* 42, 1 (1986), 144–168.

[52] Matthew Turk. 2014. Multimodal interaction: A review. *Pattern recognition letters* 36 (2014), 189–195.

[53] L. Vanacken, K. Coninx, and T. Grossman. 2007. Exploring the Effects of Environment Density and Target Visibility on Object Selection in 3D Virtual Environments . In *2007 IEEE Symposium on 3D User Interfaces*. IEEE Computer Society, Los Alamitos, CA, USA, null. doi:10.1109/3DUI.2007.340783

[54] Lode Vanacken, Tovi Grossman, and Karin Coninx. 2009. Multimodal selection techniques for dense and occluded 3D virtual environments. *International Journal of Human-Computer Studies* 67, 3 (2009), 237–255.

[55] Uta Wagner, Matthias Albrecht, Andreas Asferg Jacobsen, Haopeng Wang, Hans Gellersen, and Ken Pfeuffer. 2024. Gaze, Wall, and Racket: Combining Gaze and Hand-Controlled Plane for 3D Selection in Virtual Reality. *Proceedings of the ACM on Human-Computer Interaction* 8, ISS (2024), 189–213.

[56] Miao Wang, Zi-Ming Ye, Jin-Chuan Shi, and Yang-Liang Yang. 2021. Scene-Context-Aware Indoor Object Selection and Movement in VR . In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*. IEEE Computer Society, Los Alamitos, CA, USA, 235–244. doi:10.1109/VR50410.2021.00045

[57] Yannick Weiß, Daniel Hepperle, Andreas SieB, and Matthias Wolfel. 2018. What User Interface to Use for Virtual Reality? 2D, 3D or Speech–A User Study. In *2018 International Conference on Cyberworlds (CW)*. IEEE Computer Society, Los Alamitos, CA, USA, 50–57. doi:10.1109/CW.2018.00021

[58] René Weller, Waldemar Wegele, Christoph Schröder, and Gabriel Zachmann. 2021. LenSelect: Object selection in virtual environments by dynamic object scaling. *Frontiers in Virtual Reality* 2 (2021), 684677.

[59] Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. 2011. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 143–146. doi:10.1145/1978942.1978963

[60] Huiyue Wu, Xiaoxuan Sun, Huawei Tu, and Xiaolong Zhang. 2024. ClockRay: A Wrist-Rotation Based Technique for Occluded-Target Selection in Virtual Reality. *IEEE Transactions on Visualization and Computer Graphics* 30, 7 (July 2024), 3767–3778. doi:10.1109/TVCG.2023.3239951

[61] Zhizhuo Yin, Yuyang Wang, Theodoros Papatheodorou, and Pan Hui. 2024. Text2VRScene: Exploring the Framework of Automated Text-driven Generation System for VR Experience . In *2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*. IEEE Computer Society, Los Alamitos, CA, USA, 701–711. doi:10.1109/VR58804.2024.00090

[62] Difeng Yu, Hai-Ning Liang, Feiyu Lu, Vijayakumar Nanjappan, Konstantinos Papangelis, Wei Wang, et al. 2018. Target Selection in Head-Mounted Display Virtual Reality Environments. *J. Univers. Comput. Sci.* 24, 9 (2018), 1217–1243.

[63] Difeng Yu, Xueshi Lu, Rongkai Shi, Hai-Ning Liang, Tilman Dingler, Eduardo Velloso, and Jorge Goncalves. 2021. Gaze-Supported 3D Object Manipulation in Virtual Reality. 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 734, 13 pages. doi:10.1145/3411764.3445343

[64] Difeng Yu, Qiushi Zhou, Joshua Newn, Tilman Dingler, Eduardo Velloso, and Jorge Goncalves. 2020. Fully-occluded target selection in virtual reality. *IEEE transactions on visualization and computer graphics* 26, 12 (2020), 3402–3413.

[65] Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology* 15, 2 (2024), 1–38.

[66] Tim Zindulka, Myroslav Bachynskyi, and Jörg Müller. 2020. Performance and Experience of Throwing in Virtual Reality. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–8. doi:10.1145/3313831.3376639