Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

innovation center virtual reality

Impact of Visual Elements in Video See-Through Systems

Matan Davidi

Bachelor Thesis
August 2024

Long Cheng

Prof. Dr. habil. Andreas Kunz
Prof. Dr. Marc Pollefeys

Abstract

The presence of visual elements, whether physical or virtual, is pervasive in contemporary environments. Research indicates that these elements can exert a considerable influence on user performance and perceived mental load. However, the relationship between the two is not straightforward and varies depending on the task, the nature of the visual elements, and the quantity thereof. Despite the significance of elucidating the influence of visual elements, existing research suffers from shortcomings, including an emphasis on physical settings and a lack of attention to Diminished Reality (DR) as a modality for regulating visual elements in Mixed Reality (XR) environments. This thesis presents a review of the current state of research on visual elements and on the topic of DR, with a particular focus on the need for further study on the impact of visual elements in XR systems.

This study made use of state-of-the-art hardware and software tools, including the Meta Quest 3 Head-Mounted Display (HMD) for the creation of video see-through mixed reality environments. The study employed DR techniques, specifically the see-through method, and sampled visual elements from the standardised Chieti Affective Action Videos (CAAV) database to ensure a comprehensive range of emotional responses. The primary objective was to conduct a button-clicking exercise that would be cognitively demanding and adaptable in difficulty. This was achieved through a user study with a 2x2x2 within-subject design, which examined the effects of physical/XR environments, the presence/absence of additional visual elements, and single/double button activation on a clicking task.

The results of the study, based on data from 22 participants, indicated that XR conditions resulted in a significantly higher incidence of misclicks compared to physical conditions, with the average number of misclicks in XR conditions found to be 10.55% higher. The presence of visual elements and the complexity of the tasks undertaken significantly affected the performance metrics, namely completion time and the number of misclicks. Although physical environments were perceived as more usable than XR environments, participants reported higher cognitive absorption under diminished conditions. The study found that the removal of visual elements in DR is not equivalent to their removal in the physical world, affecting user performance in different ways. These findings highlight the importance of thoughtful design in DR applications and suggest that current XR technologies still face challenges in replicating the usability of physical environments.

Zusammenfassung

Die Präsenz visueller Elemente, ob physisch oder virtuell, ist in heutigen Umgebungen omnipräsent. Die Forschungsliteratur belegt, dass diese Elemente einen beträchtlichen Einfluss auf die Benutzerleistung und das wahrgenommene mentale Belastung ausüben. Jedoch ist die Beziehung zwischen beiden nicht linear und variiert je nach Aufgabe, Natur der visuellen Elemente und deren Menge. Trotz der Bedeutung der Klärung des Einflusses visueller Elemente ist die bestehende Forschung von Einschränkungen geprägt, darunter eine Konzentration auf physische Umgebungen und die Vernachlässigung von Diminished Reality (DR) als Modus für die Regulierung visueller Elemente in Mixed Reality (XR)-Umgebungen. Dieses Diplomarbeit präsentiert eine Übersicht über den aktuellen Forschungsstand bezüglich visueller Elemente und DR.

Die Methodik dieser Studie setzte auf state-of-the-art-Hardware- und Software-Tools ein, einschliesslich des Meta Quest 3-Head-Mounted-Displays (HMD) für die Erstellung von Videosee-through-Mixed-Reality-Umgebungen. Die Studie verwendete DR-Techniken, speziell den See-Through-Methoden, und sampelte visuelle Elemente aus dem Chieti Affective Action Videos (CAAV)-Datenbank, um eine umfassende Vielfalt von emotionalen Antworten sicherzustellen. Das primäre Ziel war es, ein Button-Clicking-Exercise durchzuführen, das kognitiv anfordernd und anpassbar in Schwierigkeit wäre. Dies wurde durch eine Nutzerstudie mit einem 2x2x2-Design innerhalb von Probanden erreicht, in der die Auswirkungen von physischen/XR-Umgebungen, das Vorhandensein/Abwesenheit zusätzlicher visueller Elemente und die Aktivierung von Einzel-/Doppeltasten auf eine Klickaufgabe untersucht wurden.

Die Ergebnisse der Studie, basierend auf Daten von 22 Teilnehmern, zeigten, dass XR-Bedingungen zu einer höheren Anzahl von Fehlklicks im Vergleich zu physischen Bedingungen führten, mit einem Durchschnitt von 10.55% Fehlklicks in XR-Umgebungen. Die Präsenz von visuellen Elementen und die Komplexität der Aufgabe beeinflussten die Leistungsmetriken, nämlich die Durchführungszeit und die Anzahl von Fehlklicks. Obwohl physische Umgebungen als benutzbarer empfunden wurden als XR-Umgebungen, berichteten Teilnehmer über eine höhere kognitive Absorption unter diminished Bedingungen. Die Studie fand heraus, dass die Entfernung von visuellen Elementen in DR nicht dem gleichen ist wie ihrer physischen Entfernung, was die Benutzerleistung beeinflusst. Diese Ergebnisse betonen die Bedeutung sorgfältiger Gestaltung in DR-Anwendungen und deuten an, dass aktuelle XR-Technologien noch Herausforderungen bei der Reproduktion der Benutzbarkeit physischer Umgebungen haben.

IWF
Institut für Werkzeugmaschinen und Fertigung
Institute of Machine Tools and Manufacturing

Impact of Visual Elements in Video See-Through Systems

Keywords: Mixed Reality, Diminishing Reality, Digital Twin

Overview

Diminishing reality refers to the ability to remove objects from a scene seamlessly, rendering them undetectable. Traditionally, this capability has relied on photo or video editing, with the drawback of being two-dimensional and lacking resilience to changes in perspective. The head-mounted-display (HMD) holds the promise of delivering an immersive and convincing 'diminishing reality' experience by manipulating visual content directly in front of the user's eyes. However, contemporary augmented reality systems often suffer from noticeable visual artifacts, such as dimmer views, unavoidable transparency when overlaying virtual elements, and limited field-of-view. In contrast, video see-through headsets introduce a new dimension to reality manipulation, with their full capability to render opaque objects across the entire field of view displayed by the camera.

To attain a convincing 'diminishing reality' experience, recent approaches have delved into the study of 'diminishing reality' through Head-Mounted Displays (HMDs). Nevertheless, the dynamic process of diminishing moving objects continues to pose a significant challenge.

Reality

Diminishing reality

Goal of the Thesis

The goal of this thesis is to achieve a believable diminishing reality experience using the video see-through camera of a modern HMD. Scene lighting, preferably real-time global illumination, as well as tracking capability using the scene mesh are required.

Tasks

Your work will start with a literature research on latest research on scene responsiveness and diminishing reality. Next, you will become acquainted with the MR platform “Unity”, photogrammetry, and the alignment of the virtual scene with the real world. Then, you will combine those skills for object tracking with the real-time scene mesh. Then you will define a suitable application. You will conduct a pilot user study for testing the system. Finally, you summarize your findings in a written report, and present them in an intermediate and final presentation.

Workpackages

- Literature on diminishing reality
- Become acquainted with Unity, photogrammetry, and scene alignment
- Develop and implement the diminishing reality system
- Application design and pilot user study
- Intermediate and final presentation
- Written report

Skills

- Programming Skills in C#
- Unity / Visual Computing skill is a plus
- Strong communication and interpersonal skills

Results

The results of this thesis need to be summarised in a written report and will be presented to the ICVR in a 20min talk.

Contact

Long Cheng, CLT D13 long.cheng@iwf.mavt.ethz.ch
Andreas Kunz, CLT E13 kunz@iwf.mavt.ethz.ch

icvr
innovation center virtual reality

Acknowledgment

I would like to express my most sincere and deepest gratitude and appreciation to the following individuals for their invaluable support and contributions to the completion of this bachelor's thesis.

- Long Cheng, for his continuous availability and invaluable feedback throughout the entire period of working on this thesis.
- Prof. Dr. habil. Andreas Kunz, for his crucial feedback that redirected my user study design onto the right path, and likely saved me from a lower-quality study implementation and plenty of hours of frustration and debugging.
- Joy Gisler, Valentina Gorobets, and Mathieu Lutfallah, for their feedback towards the midway point of the work period, which allowed me to identify some issues with the design of my planned study.
- Laura Zeller, it is in no small part thanks to her early guidance that I was able to find a topic to research and a group to research it in, and her continued moral and intellectual support proved vital in maintaining motivation high, and allowed me to survive until the end.
- Nicole Mottale, for providing precious feedback on my study design, guidance on the psychological aspects of the work, and invaluable directions in identifying the dataset pivotal to this study.
- Aaron Zeller, for allowing the exchange of ideas about the Meta Scene platform, and for contributing with thoughts and ideas on the study design.
- Paolo Camplani, for his continued patience and understanding throughout the entire thesis process, which allowed me to prioritize it and deliver the best possible results.
- Flaviana and Hagai Davidi, for providing me with a space in which to relax and sleep, and thus catch up on all the hours of sleep I had left behind.

Contents

List of Figures

List of Tables

List of Acronyms

3D Three Dimensional.

ADB Android Debugging Bridge.

ADHD Attention Deficit Hyperactivity Disorder.

AI Artificial Intelligence.

API Application programming interface.

AR Augmented reality.

ASD Autism Spectrum Disorder.

CAAV Chieti Affective Action Videos.

CAQ Cognitive Absorption Questionnaire.

DR Diminished reality.

EEG Electroencephalography.

FDR False Discovery Rate.

HMD Head-mounted display.

LiDAR Light Detection and Ranging.

NASA-TLX NASA Task Load Index.

OSPAN Operation Span.

OST Optical See-Through.

RSPM Raven's Standard Progressive Matrices.

SSQ Simulator Sickness Questionnaire.

SUS System Usability Scale.

UI User Interface.

USB Universal Serial Bus.

VR Virtual reality.

VST Video See-Through.

XR Mixed Reality.

1

Introduction

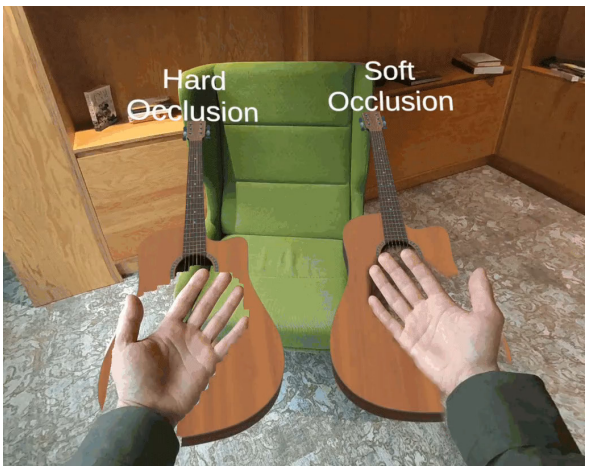
In modern-day environments, whether we are considering physical ones in which people work, meet up with others, relax alone, etc., or virtual ones, visual elements, meant as anything that is visible, but not central to what a person is doing, are ubiquitous. It suffices to think about the smartphone that most people reading this work own [1]. Regardless of model, production year, or price, any notification, icon, widget, or text shown to the user of the device is a visual element, and most of them tend to passively exist on the screen, there for the user to interact with, whereas others attempt to grab the user's attention, most notably notifications or in-app contents. In doing so, however, they risk drawing the user's focus away from some other task that they were previously performing, with potential negative effects in terms of performance in the previous task and cognitive function or concentration [2]. Research indicates that reducing interruptions by notification in the main task can be beneficial for both performance and perceived strain [3], and for this reason we attempted to study a generalisation of these circumstances by abstracting smartphone notifications to simply some visual elements, and researching which possibilities there are to implement their removal from a user's perception and the impact that this removal can have on performance in a central task, along with the mental load that the user perceives because of it.

1.1 Motivation

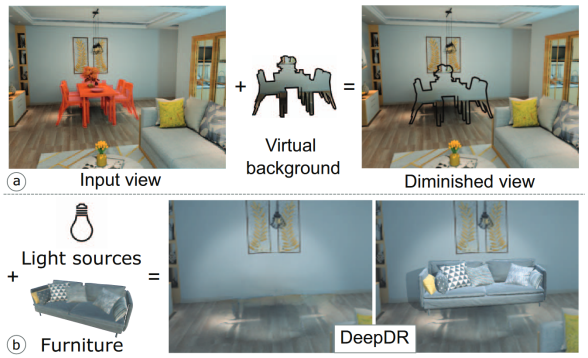
1.1.1 Visual Elements

As previously mentioned, researching visual elements can provide insight into the impact that they can have on a user, along with possible reasons, and design principles that could be followed to ease a system's use on its users. To be more accurate and allow scientific research, however, it is necessary to define what a visual element is; thus, for the purposes of this research, we defined a visual element as a *fundamental component of the visual design of an environment that contributes to the overall perception and interpretation of visual information in a given context*. To provide an intuitive example, any object on a desk where a user is working can be considered a visual element, either because it is relevant to the user's task (and so they will use it to some extent), or because it is irrelevant and cluttering (so they will

be somewhat "distracted" by it).
As additional instances of visual elements, we provide the following figures. In Figure 1.1(a), both guitars and the chair behind them are examples of visual elements. Specifically, virtual visual elements and physical visual elements, respectively. In Figure 1.1(b), both the table in subfigure (a) and the couch in subfigure (b) are examples of visual elements. Specifically, physical visual elements and virtual visual elements, respectively.



(a) Example of visual elements as defined for the purposes of this research. Taken from uploadvr.com [4]



(b) Example of visual elements as defined for the purpose of this research. Taken from Gsaxner et al. [5]

Figure 1.1: *Examples of visual elements relevant to the context of the presented work.*

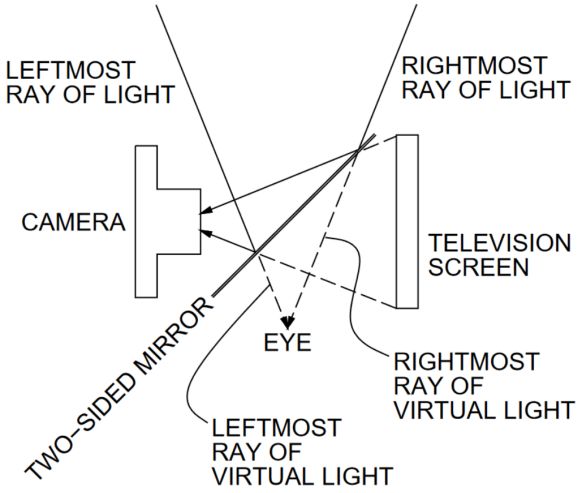
The study of visual elements in *Virtual Reality* (VR) and *Mixed Reality* (XR) can be a crucial aspect of understanding how users perceive and interact with virtual environments. The visual cues presented in VR and XR systems can significantly impact the user's experience, influencing their ability to complete tasks efficiently and effectively [6, 7]. By examining the effects of visual elements on user performance and perceived mental load, we can inform the design of more intuitive and user-friendly VR or XR systems. This, in turn, can lead to improved learning outcomes [8], enhanced training experiences [9], and increased productivity in various fields, such as healthcare [10], and industry [11]. Furthermore, a deeper understanding of how visual elements affect cognitive processes can also provide valuable insights into human perception, attention, and decision-making, ultimately contributing to the development of more sophisticated and human-centered VR applications.

1.1.2 Diminished Reality

As we have seen, the effect of removing visual elements from the user's perception can be significant. Furthermore, there may be additional advantages for individuals with conditions such as *Autistic Spectrum Disorder* (ASD) or *Attention Deficit Hyperactivity Disorder* (ADHD), some of whom may display heightened sensitivity to visual information [12, 13] or be prone to overstimulation [14, 15]. Regrettably, physical removal is not always a viable solution. For example, consider the case of an open-plan office in which colleagues are watching a highly dynamic, colourful video, or a lecture in which the door is positioned in front of the viewer while the corridor behind it is busy. In such cases, virtual removal can be an interesting alternative, as the visual elements can simply be hidden rather than actually removed.
Diminished Reality (DR) is a family of techniques, first introduced by Steve Mann and James Fung [16], that allows the visual perception of the user of a *Head-Mounted Display* (HMD) to be manipulated by

filtering out visual clutter to replace it with relevant subject matter, and has since developed into a technology that selectively subtracts or attenuates unwanted visual information from the user's field of view. In fact, 16 years later, a literature review published in 2017 by Mori et al. shows that four trends have since emerged for potential implementations of DR: *Diminish*, *See-Through*, *In-Paint*, and *Replace* [17].

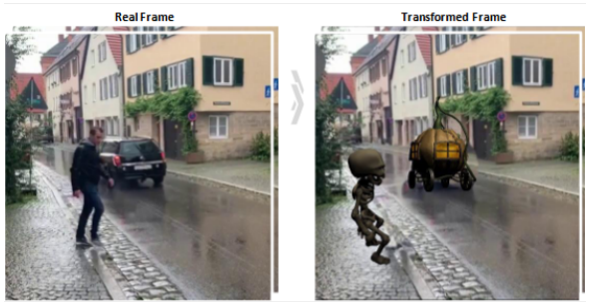
The first, *Diminish*, involves degrading visual functions, such as distorting a field of view by weakening, desaturating, or distorting color information, for a specific purpose. This technique can be achieved through the acquisition, editing, and presentation of light rays that enter the eyes via a see-through Head-Mounted Display (HMD) as a "reality interface". An example of this can be seen in Mann's paper [18], where he developed a head-mounted reality mediation device that would show the wearer a distorted visual feed of their surrounding environment to compensate for the mismatch between the camera's and eye's viewpoints, thereby creating a more natural and immersive augmented reality experience, see Figure 1.2(a). In contrast, the *See-Through* approach involves covering real objects with images of the background they are occluding to make them virtually invisible in our vision. This process substitutes the light rays from actual objects with those from the background, allowing us to recreate a scene where real objects are absent from our visual field using HMDs. A similar approach is used in augmented and mixed reality, where reconstructed backgrounds are superimposed onto a perceived visual field. Such an implementation can be seen in the work by Kawai et al. [19], where furniture in a virtual environment was hidden by superimposing its background and keeping the room's geometry into consideration, see Figure 1.2(b). The *In-Paint* approach involves generating plausible background images based on the surroundings, which differs from the See-Through approach in that it requires real-time computation of a plausible background for the visual element to be diminished, rather than using a pre-computed background. This approach then superimposes the computed background to give the illusion of removal. This can be seen, for instance, in Figure 1.2(c). Finally, the *Replace* implementation method takes this concept a step further by overlapping a real object with a virtual object, making the real item appear to be replaced by the virtual one. This requires blocking the light rays from the real visual element with the overlaid virtual visual element, which can be achieved by using a virtual object of a similar or larger size to the physical one, or by performing See-Through or In-Paint processes first and applying the overlay later. This last concept is implemented in [20], where a man and a car can be seen in-painted, and replaced by halloween-themed virtual counterparts.



(a) *Illustration for a Diminish implementation of DR, taken from Mann [18], and showcasing the design for the orthoscopic reality–mediator, used to distort the visual feed of an* optical see-through *(OST) device: A double-sided mirror device diverts light to a camera and displays a mirrored TV screen, creating a transparency illusion with quantigraphic[2] image processing.*

(b) *See-Through implementation of DR, adapted from Kawai et al. [19]. The images on the left represent the original inputs, while those on the right showcase their outcomes after superimposing the mesh of the floor and walls over the bookcase, the speaker, and the ball that rests atop it, effectively hiding them from the user's view.*

(c) *In-Paint implementation of DR, adapted from img.ly [21], where an in-painting algorithm was applied to the left image to remove the car by computing a plausible background with which to replace it, resulting in the right image.*



(d) *Replace implementation of DR, adapted from Kari et al. [20], showing a scene on the left where an in-painting algorithm was applied to remove the person and the car, and then halloween-themed virtual visual elements were added to create the scene on the right.*

Figure 1.2: *Comparison of the four approaches to DR implementation as described by Mori et al.*

Given the scarcity of research in the field of Diminished Reality (DR), particularly its ability to 'diminish' or hide elements of the physical environment from the user's view, we decided to investigate the potential benefits of applying such techniques. A *Video See-Through* (VST) system, which overlays digital information and objects onto the real world by capturing and processing video footage of the user's surroundings, is one technology that enables this capability by allowing a researcher to manipulate the physical environment displayed through the HMD by running specific algorithms, or by introducing virtual elements. If the believable illusion of object removal is possible, we hypothesise that it could be used to simplify or remove clutter from the user's environment [22], to the benefit of some central task that the user is performing.

[2]Technique for capturing and manipulating light fields to create realistic 3D images or holograms.

1.2 Related Work

1.2.1 Impact of visual elements

In their study, Li et al. [23] examine the impact of VR on cognitive load and learning outcomes. The study examines the impact of cognitive load and task performance in both real-world and VR environments. A total of 46 participants completed tasks with and without a secondary task in both environments. The secondary task, a detection response task (DRT), was employed to gauge cognitive load based on response time and omission rate. The results demonstrated no statistically significant differences in cognitive load and task performance between the VR and non-VR environments. Nevertheless, the study identified a potential trend indicating that VR environments may confer benefits, such as heightened immersion, which could enhance concentration and task performance.

Furthermore, the study identified discrepancies in performance between male and female participants, with females exhibiting diminished performance in VR environments and experiencing marginally elevated cognitive load. Overall, the study suggests that VR environments may have an influence on performance and/or task load that is absent from physical environments. Despite these results, this study does not analyse the impact of visual elements alone, but always in conjunction with a separate, secondary task, and mainly works in physical and VR environments, and no XR.

Additionally, Groening et al. [24] examine the influence of varying the number of game design elements on motivation and performance in cognitive tasks. The study employed three online experiments, involving a total of 440 participants, to manipulate the number of game design elements. These experiments employed both between-subjects and within-subjects designs. The findings indicate that an augmented number of game design elements typically fosters enhanced motivation and performance. However, the study also identified a curvilinear relationship, indicating the potential for a threshold beyond which the inclusion of additional elements may not result in further improvements, and may even have a detrimental effect on effectiveness. It is noteworthy that the removal of game design elements from an ongoing task did not result in a decline in performance.

The study emphasises the pivotal role of competence and relatedness in driving the positive effects of gamification. The findings indicate that a substantial number of game design elements can markedly enhance motivation and performance, but it is crucial to strike a balance to prevent overloading users. Consequently, this research suggests a non-linear relationship between the incorporation of additional visual elements, task performance and mental load. However, this paper only concerns itself with game design elements, which are only relevant for a specific type of application, namely video games, and does not take into consideration XR environments, where the impact may differ.

In regard to this topic, research indicates that the impact of supplementary visual elements on performance in a given task and the user's perceived mental load is not straightforward and varies considerably based on the task, the nature of the visual elements, and their quantity. The following table presents a summary of these findings. N.b., in the following table, the acronym *Electroencephalography* (EEG) is used, a method used to record electrical activity of the brain.

Paper	Task context	Independent variables	Dependent variables	Effect on performance	Effect on mental load
Tsurukawa et al. [25]	Visual search tasks on computer screen, filtered amount of visual information	Filtered vs Unfiltered amount of visual information	Cognitive load, task accuracy and response time	N/A	Fewer elements $\Longrightarrow$ lower mental load
Vasiljevic et al. [26]	Neurofeedback-based game while varying graphical elements presented on screen	Presence and type of graphical elements	EEG signals (user attention), in-game success metrics	N/A	Fewer elements $\Longrightarrow$ lower mental load
Johnson [27]	Time-sensitive, deep focus task while visual distractions were introduced	Presence vs absence of visual elements	Task accuracy and speed, galvanic skin Response (stress levels)	Fewer elements $\Longrightarrow$ improved performance	Fewer elements $\Longrightarrow$ lower mental load
Redlinger et al. [28]	Visual working memory task in VR with various game-like visual features	Presence vs absence of game-like visual features	Task accuracy and reaction time, EEG power changes	None	None
Kim et al. [29]	Video-based learning activities while visual cues were introduced	Presence vs absence of visual cues, high vs low level of self-regulation	Pupil dilation, behavior logs (cognitive load), learning performance (performance)	N/A	None
Workman [30]	Interpreting complex graphical linguistic representations	Complexity of graphical linguistic representations, presence vs absence of visual elements	Subjective ratings and physiological indicators (cognitive load), comprehension tests (accuracy)	N/A	Fewer elements $\Longrightarrow$ higher mental load, and vice versa
Yu et al. [31]	Visual discrimination task with varying levels of visual degradation	Level of visual degradation	EEG signals (cognitive load), accuracy and reaction time (task performance)	Fewer elements $\Longrightarrow$ higher mental load	N/A
Hoesl et al. [32]	Camera-based cinematographic *user interface* (UI) where visual elements were progressively reduced	Level of visual element reduction	task completion time and accuracy (performance), subjective ratings and physiological indicators (cognitive load)	More visual elements $\Longrightarrow$ improved performance up to threshold, then worse	N/A
Dahlstrom-Hakki et al. [33]	Educational activities in VR while being exposed to video/audio distractions	Visual vs auditory distractions, level of distraction (none, moderate, high)	Task completion time and accuracy (performance), subjective ratings and physiological indicators (cognitive load)	More visual elements $\Longrightarrow$ higher mental load	More visual elements $\Longrightarrow$ worse performance
Deng et al. [34]	Visual tasks on a simulated maritime operation platform	Combinations of visual components (charts, maps, indicators), scenario complexity (low, medium, high)	Response times, error rates (performance), subjective ratings (cognitive load)	N/A	More visual elements $\Longrightarrow$ higher mental load
Chung et al. [35]	Human-computation game with varying amounts of visual information	Amount of visual information (low, medium, high)	Task completion time, accuracy (performance), subjective ratings (cognitive load)	More visual elements $\Longrightarrow$ improved performance	N/A

Table 1.1: *Summary table of previous research showing the study procedure for the effect of visual elements on perceived mental load and task performance.*

	Lower mental load	Same mental load	Higher mental load	Improved performance	Worse performance
Fewer visual elements	[25–27]	[28, 29]	[30, 31]	[27, 32][3]	[32][3]
More visual elements	[30]	[28]	[33, 34]	[35]	[33]

Table 1.2: *Overview table of previous research showing the seemingly contradictory results on the effect of visual elements on perceived mental load and task performance.*

As can be seen from the above tables, most of these studies do not consider XR environments and DR as a technique for diminishing visual elements, and most do not even employ VR. From all of these limitations of previous research, we can clearly see that XR seems to be relatively ignored as a field for the study of visual elements, and thus the impact of visual elements when filtered through the HMD's cameras is still mostly unknown. Furthermore, DR seems to be an unexplored venue for controlling visual elements in an XR environment as a consequence, despite showing considerable promise.

1.2.2 DR to control visual elements

Research in DR [16, 17] shows that physical elements can be visually removed from the user's field of view with a believable level of illusion through VST systems, for instance employing a See-Through implementation (see Section 1.1.2) using a mesh generated using photogrammetry [36], which consists of capturing digital images and using them to extrapolate a digital mesh of the physical environment.

Regarding the relationship between DR and visual elements, Cheng et al. [22] shows that although their DR implementation was conducted in VR as a proof of concept and only diminished static visual elements, their participants' attitudes towards DR were generally positive, and provides some guidelines to consider when designing DR implementations or applications, including but not limited to:

- Users should be made aware of the manipulation of their visual perception and should remain conscious of what is being removed from their perception (e.g. by drawing the outline of the removed visual element).
- Only irrelevant information should be removed.
- Users expressed apprehension for accidentally colliding with diminished physical elements that they are no longer able to see.
- Social awareness should be guaranteed at all times.
  - If a colleague, friend, etc. comes into view, the user should be able to perceive them normally. The DR implementation should also allow the user to avoid inadvertently invading other people's personal space.

Finally, Lee et al. [37] studied the cognitive challenges posed by smartphones in modern life, since they are a fundamental part of our daily routines, even though they can have negative cognitive effects. They investigated the possibility of using DR "by selectively removing or occluding distracting objects from the user's field of view" through the use of a Microsoft HoloLens 2 OST HMD. This approach demonstrates potential, as evidenced by the findings that reducing the participant's smartphone usage, as measured by both the OSPAN [38] and RSPM [39] metrics, effectively mitigates the cognitive distractions associated with it, comparable to the effect of analogous physical removal. Given that both are widely used and validated cognitive capacity and sustained attention measurement scales, this result is especially significant. Furthermore, they found that although the hologram that was used to diminish

sight of the smartphone was often noticeable, its impact on the primary task exhibited variability and often diminished over time, indicating that See-Through and Replace DR implementations (see Section 1.1.2) can be used with varying impact on the main task. This paper's primary limitation is its emphasis on smartphones as visual elements to be diminished, coupled with an exclusive focus on OST devices, which lack the capacity to fully conceal a physical visual element due to the transparency of the screen on which holograms are projected and the restricted field of view, which allows users to perceive diminished objects out of the corner of their eye.

1.3 Task Formulation

A review of the literature reveals inconsistencies in the research on the impact of visual elements. There is a tendency for the focus to shift away from the impact of visual elements themselves and towards the utilisation of visual elements as a means of achieving other objectives. For instance, visual elements may be utilised as a means of incorporating game design elements [24] or of implementing a secondary task [23]. This may be attributed to the vagueness surrounding the definition of a visual element. It is therefore essential to examine the introduction of visual elements in isolation. Nevertheless, the majority of literature agrees that the introduction of supplementary visual elements can have a considerable effect on a user's performance and the perceived mental load associated with a mentally demanding task. Furthermore, the majority of existing research focuses on physical environments, with only a limited number of studies exploring avenues in XR. This underscores a notable deficiency in research examining the influence of visual elements in XR systems. Such research would facilitate the utilisation of DR techniques to regulate the quantity of visual elements perceived by the user. This is an area that has been identified as being somewhat under-researched, particularly given that the existing studies that do investigate this topic tend to use OST HMDs, which are not as effective as VST HMDs in properly concealing visual elements.

This is the primary rationale behind our decision to conduct a similar study, employing explicit DR techniques and VST HMDs. The former are supported by the literature review conducted by Mori et al. [17], while the latter afford greater flexibility in controlling the user's environment. In particular, the objective is to ascertain whether the discrepancy between objective performance and subjective perception is analogous between a physical environment and a VST environment when comparing metrics between conditions with and without the additional visual elements. However, given the contradictory nature of previous research on the effects of visual elements, which suggests a careful study design if the aim is to limit the scope to only a positive or only a negative effect, and which may require background knowledge of attention and visual cognition, we will focus not on the specific influence these elements have, but rather on whether there is any influence at all when the visual elements are physically removed, and whether an analogous influence can be detected with DR removal.

2

Methodology

This chapter sets forth the methodological framework that serves as the foundation for this research project. It provides a comprehensive overview of the hardware and software tools employed, the implementation of DR, the selection and implementation of additional visual elements, and the design of the main task, including the considerations that informed its development and the final implementation. Finally, the chapter outlines the study design, including the conditions under which the task was performed, the metrics collected, the counterbalancing measures employed, and the setup of the conditions.

2.1 Hardware and Software Utilization

This subsection outlines the hardware and software components employed during the entire work on the thesis.

2.1.1 Hardware

The primary hardware used in this study included:

- **Meta Quest 3 Head-Mounted Display:** The Meta Quest 3 was utilized to create VST environments for the *Mixed Reality* (XR) conditions. The device's capabilities were tested extensively, particularly focusing on the Unity real-time engine integration. The headset was chosen for the following reasons:
  - *Standalone capabilities:* The Meta Quest 3 operates in standalone mode, meaning it doesn't require a connection to a PC or any other external hardware to function. This feature offers a significant advantage in terms of flexibility and mobility. Users are not constrained to a specific physical location near a PC, making the approach and the implementation more adaptable for potential day-to-day use. This standalone capability allows for a broader range of applications and use-cases, making it an ideal choice for this research.

  - *Full-colour VST capabilities:* Thanks to its high-resolution external cameras, the Meta Quest 3 boasts full-colour, high-quality, stereoscopic video see-through capabilities. This feature allows for the creation of rich and immersive VST environments, using state-of-the-art quality video for our visual illusion implementation. The high-quality video feed enhances the believability of the visual illusions created by the DR techniques, contributing to a more convincing and effective user experience, which makes this capability crucial for the success of the research. Compared to other headsets, it offers stand-alone, high quality XR experiences, with a less limited resolution, at a more affordable price.

- **Computing System:** A high-performance desktop computer with the following specifications:
  - *Processor:* Intel® Core™ i9-10900K CPU (Comet Lake architecture, Socket 1200) @ 3.70GHz, 3696 Mhz, 10 Cores, 20 Logical Processors
  - *Graphics Card:* NVIDIA GeForce RTX 3090, with 24GB GDDR6X VRAM
  - *RAM:* 32GB DDR4
  - *Operating System:* Microsoft Windows 10 Education, build 19045.4529
- **Samsung Galaxy M33 5G mobile phone:** To perform LiDAR-based 3D scanning of the physical desk and wall.
- **iPhone 14 Plus**: To perform Gaussian splatting-based 3D scanning of the physical desk and wall.

2.1.2 Software

The software tools and platforms used during the study included:

- **Meta Horizon OS**: v63, operating system present on the Meta Quest 3 HMD.
- **Unity Real-Time Development Platform:** Version 2022.3.22f, was used to develop the user study application. This included implementing the button-clicking task and attempting integration with the Meta Scene *Application programming interface* (API) (which comprises the Depth API and Mesh API) to attempt to achieve DR, although these attempts were ultimately unsuccessful, and thus simpler approaches and photogrammetry were preferred.
- ***Chieti Affective Action Videos* (CAAV) database** [40]: This dataset provided the video contents used in the study, featuring clips with a wide range of standardized stimuli with varying levels of emotional valence and arousal to test their impact on task performance.
- **Polycam and RealityScan:** Version 1.3.6 and Version 1.4.1, respectively. These applications facilitated the creation of accurate LiDAR-based 3D scans for the DR implementation.
- **Scaniverse:** Version 3.0.1. This application facilitated the creation of accurate Gaussian splatting-based 3D scans for the DR implementation.
- **Blender:** Version 4.1, was used for merging and refining the 3D scans to hide physical elements effectively on the real-life desk.

The combination of these hardware and software tools enabled the creation of a comprehensive research and study environment.

2.2 DR Implementation

A large part of the first few months has been spent researching and familiarising with the Meta APIs for Virtual- and Mixed Reality, as well as with the capabilities of the Meta Quest 3 in terms of performance, latency, visual acuity and fidelity, as well as the code that the headset made available to the Unity engine through its APIs. The results of this experimentation can be found in the following sections.

2.2.1 Meta Depth API

The Meta Depth API is a tool for XR developers, offering real-time depth estimates through the headset's external depth sensor that enhance XR experiences. It allows virtual objects to interact with the physical environment in a believable way through dynamic occlusions[1]. The API supports both hard (more noticeable) and soft occlusion methods and is compatible with Unity's rendering pipelines, including the Built-in and *Universal Rendering Pipeline* (URP).

Developers can leverage the Meta Depth API in various applications, from immersive gaming to VR simulations, where the interaction between virtual and physical content is crucial. However, it does come with some limitations. It requires specific software versions, e.g. is not immediately compatible with the latest Unity versions, or it may require a minimum headset operating system version, and can be resource-intensive, particularly when using soft occlusion, which demands more GPU power.

On the upside, the Meta Depth API significantly improves the realism of XR applications by accurately blending virtual and real worlds. It offers flexibility with different occlusion modes and shader options, allowing developers to choose the best fit for their project. On the downside, the increased GPU requirements for soft occlusion could limit its use in less powerful systems.

Limitations

Unfortunately, the raw data obtained from the external depth camera of the HMD cannot be accessed, which poses a significant limitation to the usability of this API. Indeed, at the time of my experimentation, it is only through Meta's pre-made default shaders that one can enable precise occlusion computation between virtual objects and physical ones, as described in the Meta specification. This precludes the opportunity of performing object tracking through this functionality. While searching for Meta's official documentation, it becomes apparent that the Meta Depth API is primarily designed for occlusion-based effects, and is thus very limited outside of that scope.

2.2.2 Meta Mesh API

The Meta Mesh API is a tool for XR developers, designed to enhance XR experiences on Meta Quest devices. It works in conjunction with the Meta Depth API to provide a more immersive and interactive environment for users. The Mesh API allows developers to access and utilize real-time 3D mesh data of the user's physical environment by performing an early scan of it, and later having the possibility of applying collisions, or even replacing certain surfaces or furniture with virtual prefabs[2], enabling more sophisticated interactions between virtual and real-world elements.

[1] Real-time calculation and rendering of which objects are visible and which are hidden behind others as the viewer moves through a virtual environment

[2] Reusable asset that acts as a fully configured `GameObject` template, complete with all its components, property values, and

Developers can leverage the Meta Mesh API for various applications, from advanced gaming experiences to practical XR simulations. The API supports the creation of believable XR experiences by allowing virtual objects to interact with the physical environment in more realistic ways. This includes accurate physics simulations, dynamic occlusions, and enhanced spatial awareness for virtual elements.

The Meta Mesh API is part of a broader ecosystem of tools provided by Meta for XR development. It integrates with Unity's AR Foundation framework, which is designed for multiplatform AR development, and the XR Interaction Toolkit, which provides a high-level, component-based interaction system for creating XR experiences.

On the positive side, the Meta Mesh API significantly improves the realism and interactivity of XR applications. It offers developers the ability to create more immersive experiences by allowing virtual content to interact more naturally with the real world. The API's integration with Unity's existing XR tools also provides a familiar development environment for many developers.

Limitations

Similarly to the Meta Depth API, access to the raw data underlying the functioning of the API is limited. First of all, once the room scan is performed, it is impossible to edit it again without performing the scan anew, thus no longer being able to remove or relocate furniture, or correct walls placements or inclinations. This becomes problematic not only in dynamic contexts such as a user study, where if there is a technical issue it either needs to be corrected straight away, or taken into consideration when adapting further runs of the study, but also when simply developing consumer-level XR applications, where the user experience needs to be streamlined as much as possible to facilitate their use. If this were the only issue, it would not consists of a significant limitation, as it can simply be taken into consideration when designing the application and paying extra attention when performing the room scan. However, inconsistencies were found using the Meta Mesh API in the registration[3] of the virtual room with the physical room at the start of an application, which resulted in inconsistencies in height, position and rotation of certain added furniture, or even entire walls, which, once again, could only be remedied by re-scanning the physical room. Although these are mostly minor inconsistencies, and thus negligible on the scale of the entire room, this would make this API difficult to use in smaller scale applications, which would only make use of, e.g., collisions of virtual objects with a physical table, or positioning of virtual visual elements on top of physical furniture, etc.

2.2.3 Computer vision

Computer Vision can play a crucial role in some implementations of DR. Recent advancements in this field have leveraged deep learning techniques and large-scale internet photo collections to enhance the realism and efficiency of DR systems. Many different implementations have risen since the early days of the DR field, employing many different computer vision techniques [5, 41–43]. The Meta Quest 3 HMD features an outward-facing camera used for its VST functionality.
Unfortunately, however, due to privacy and security concerns, Meta does not provide access to its real-time footage out of the box, and the only possibility to allow this is not officially supported, and involves working around the limitations of the headset abusing the Android Debugging Bridge (ADB)[4] to stream the headset's camera footage to an external USB-connected computer. However, this approach only

child `GameObjects`

[3] Process of alignment of a virtual object to its physical counterpart to ensure proper correspondence

[4] Command-line tool that enables communication between a computer and Android devices

allows access to the footage after being processed by the internal operating system of the device (thus including all UI and virtual elements of the running application), including the latency that comes with that processing. This, coupled with the fact that the contents are streamed to an external computer that would need to perform the computer vision itself, encode the results and send them back to the headset, either via WiFi or using the headset's file system, which would add on to the already existing latency, makes computer vision too cumbersome for being utilised in DR applications.

2.2.4 Conclusions

Due to the aforementioned limitations, the choice of DR implementation fell onto the See-Through technique, as the Diminish approach was deemed unfit for this study due to the distortion of visual information which does not allow removal of visual elements, the In-Paint method was impossible due to a lack of computer vision capabilities (see Section 2.2.3) and the Replace procedure would not not allow to remove a visual element but simply to replace it (which has also been done [37] and is thus possible, see Section 1.2.2), but more importantly, due once again to the lack of computer vision capabilities (see Section 2.2.3) and of depth perception-based object tracking (see Section 2.2.1), it would make simply replacing a visual element with a virtual counterpart only possible statically, without adapting to the user's movements. The only possibility for Replace DR would be to perform a one-time registration of the virtual visual element onto its physical counterpart (in a Replace or See-Through manner, as detailed in Section 1.1.2), during an early setup phase. However, following some experimentation, it became evident that this approach was not viable. It became clear that the accuracy of the registration was insufficient to create a convincing illusion for the purposes of this research, due to the fact that the virtual element did not move at the same rate as the physical one when the user moved backwards, forwards or sideways. This, unfortunately, posed too big of a limitation on the believability of the illusion to not jeopardise the results of the study.

2.3 Visual Elements Implementation

Since, as mentioned in Section 1.3, the scope of the work is not to restrict our research to visual elements with a positive or negative impact on the user and the task, we simply require some sort of effect from the added visual elements. Thus, to ensure that this effect can be seen in the greatest number of participants, a set of visual elements with a wide range of levels of salience is required; while to avoid bias, an even distribution of visual elements that trigger positive and negative feelings is imperative. Additionally, research shows that dynamic visual elements (such as videos compared to pictures) can provide additional benefits compared to static ones [44], e.g. they can be highly effective in terms of viewer engagement and attention [45], as movement onset has been found to be related with increased attention attraction [46]. However, if video is picked as a visual element medium, attention is required in ensuring that even when removing its audio, the influence on the user is guaranteed.

2.3.1 Chieti Affective Action Videos database

The *Chieti Affective Action Videos* (CAAV) database [40] is a standardised and validated [40,47] database consisting of 360 short mute video clips depicting one or two actors performing actions belonging to a wide range of emotions based on the Circumplex model of human emotion [48], which defines two dimensions: **valence** and **arousal**.

- **Valence** refers to the attractiveness (if positive) or averseness (if negative) of the provided stimulus. Emotions that are pleasant or desirable will be defined by high valence, whereas ones that are unpleasant or undesirable will be defined by low negative valence.
- **Arousal** is defined as the physiological and psychological state of being awoken or stimulated. Emotions that are intense and activating will be related to high levels of arousal, while calm and deactivating ones will be tied to low arousal.

The CAAV database was not only designed to provide standardised visual stimuli with the aim of eliciting emotional reactions from the observer, but was also created as a tool to assist in psychology- and cognitive research.

Out of the 360 different video clips, 90 of them portray different actions, and for each of them there are 4 possible points of view:

- Female actor, 1st-person
- Female actor, 3rd-person
- Male actor, 1st-person
- Male actor, 3rd-person

The emotional response itself is elicited through the contents of the video, as they portray actions of differing values in both valence and arousal, and thus causing an emotional response that was then validated through 444 participants' self reports. The clips are meticulously designed to portray everyday scenarios that are likely to evoke a specific emotional response, such as happiness, sadness, anger, or fear. A video might depict a person receiving positive news, experiencing a loss, or confronting a threat. The objective is to evoke an emotional response in the viewer, creating the impression that they are directly experiencing the situation by employing relatable scenarios and subtle emotional cues. Four examples of still frames from four different clips can be seen in Figure 2.1.

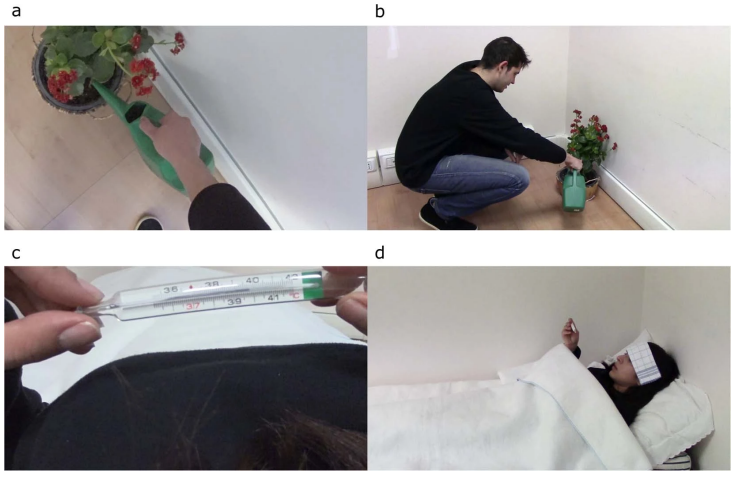


Figure 2.1: *Still frames extracted from four different videos taken from the CAAV database, taken from [40]. Frames from CAAV videos: (a) "Watering a plant", first-person male; (b) "Watering a plant", third-person male; (c) "Measuring fever", first-person female; (d) "Measuring fever", third-person female.*

2.3.2 Motivation

Firstly, video clips were chosen as visual elements instead of static elements like pictures, non-moving 3D objects, etc. due to the higher potential for advantages, such as more complex narratives and less need for interpretation, and the ability to show the passing of time, movement, actions, consequences, reactions and emotions in a more convincing and easier to understand manner [44]. This has the potential to grab the attention of the user more than simple static elements [45], which would likely only take a short amount of time for user to zone out. Then, the choice of which video to show exactly befell the CAAV database because not only is it a standardised set of short, mute clips, thus not risking the interference of other extraneous variables such as auditory distractions in the environment, but also because it was validated on all adult age ranges [40, 47], and because it was designed specifically for research on visual elements and "perception, visual attention, and emotional memory" [40], being based on a widely accepted model for human emotion, namely the Circumplex model [48]. An alternative could have been the use of videos containing flashing lights or colours. However, this was ultimately rejected on the grounds that it could introduce bias based on the experimenter's subjective perception of what constitutes a distraction. The use of a standardised database ensures consistency in the presentation of visual elements across all conditions, thereby reducing variability and facilitating the isolation of the effect of visual elements on task performance. Furthermore, the utilisation of a standardised database facilitates the replication of the study, as identically-validated visual elements can be employed across multiple participants and conditions. Additionally, the potential for causing mild to moderate reactions in participants if the intensity of the flashing were to become excessive presented a significant risk for our participants, as well as an additional extraneous variable.

2.4 DR Visual Elements Implementation

Once the supplementary visual elements had been chosen, the subsequent phase was to implement the visual elements that would facilitate the application of DR. Since the focus of the study is on a central task (to be introduced in Section 2.6), a regular office desk was chosen for the overall environment because of its versatility and suitability for a large number of possible mentally demanding tasks. This choice works in two ways, as such a desk could be used as a tool to hide the visual elements using the See-Through technique (see Section 2.2.4), by simply superimposing a mesh of the empty desk over the "cluttered" physical one.
Two possible methodologies were identified for the modelling of the desk: the first was to undertake the modelling process manually using software solutions such as Blender or Maya; the second was to employ photogrammetry. Photogrammetry was chosen as a more versatile, flexible, scalable method that can be used to capture a wide range of environments and objects, from small artifacts to large landscapes with a higher level of detail accuracy and realism, especially in the mesh's material and texture, which may enhance the believability of the DR experience. While all of these benefits are technically possible when modelling the mesh by hand, they would likely have required a significant time investment. Furthermore, photogrammetry is an actively researched and constantly improving field, whereas hand modelling has a hard upper limit given by the combination of time available for modelling and the familiarity and skill of the modeller. The advantage of photogrammetry, at least in terms of convenience, is likely to increase with time, and we have therefore chosen it as a stepping stone on which to base future work.

2.4.1 Photogrammetry

As mentioned in Section 2.1.2, three applications were used. Their differences are briefly mentioned in Table 2.1:

Application name	Main working principle	Features	Pros	Cons
RealityScan	Not publicly disclosed (likely traditional techniques)	Based on RealityCapture for desktop; Designed for integration with Unreal Engine	High polygon count in models; High visual fidelity	File sizes can be very large and require additional cleanup
PolyCam	LiDAR (if available), traditional photogrammetry	AI-powered 3D scanning; Supports 360-degree photo capture	Web interface; Higher acceptance for lower quality pictures/adverse lighting conditions	Lower fidelity; Texture size is limited
Scaniverse	Gaussian Splatting (LiDAR can be used if available) and traditional photogrammetry	Visualise scans in *Augmented Reality* (AR); Convert scans to videos and share them in-app	High variety in export file formats; High precision and accuracy	Accuracy drops when scanning large objects

Table 2.1: *Summary table of the differences between the three tested photogrammetry mobile applications.*

LiDAR (Light Detection and Ranging) is a remote sensing technology that uses laser light to measure distances and create high-resolution 3D models of environments. By emitting laser pulses and measuring the time-of-flight, LiDAR systems can accurately map the geometry of a scene, often with centimeter-level precision. In contrast, Gaussian splatting is a computer graphics technique used for texture mapping and rendering, where a 2D texture is projected onto a 3D surface using a Gaussian distribution to create a smooth, continuous appearance. While LiDAR and Gaussian splatting are both used in various applications, they differ significantly from traditional photogrammetry techniques, such as stereo photogrammetry, multi-view stereo, and structure from motion, which rely on 2D images and feature matching to reconstruct 3D models.
Multiple scans were taken with all applications, and the results were later merged using Blender to make up for the small imperfections of single scanned meshes. Blender was picked for its versatility, fairly reasonable learning curve, robust set of features and tools designed for 3D modeling, active community, extensive documentation, and the fact that it is free and open-source. As previously mentioned, the study would take place on a regular office desk, and thus an empty desk against a wall, on which the study would later have to be conducted, was scanned, along with part of the wall, to allow to diminish not only short visual elements on its surface, but higher visual elements towards the wall, as well. A picture of the result of the scans and the merging process can be found in Figure 2.2.

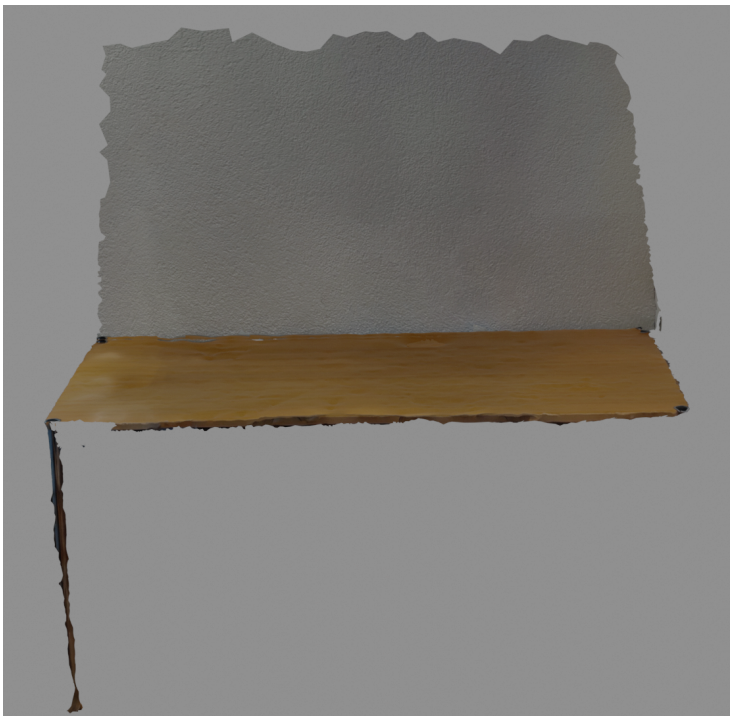


Figure 2.2: *Showcase of the resulting desk mesh.*

2.5 Hypotheses

In this section, before going further with the design of the user study, we briefly discuss the null hypotheses that we would like to test with it. For convenience and ease of reading, we denote H_i as the null hypothesis H_0^i for some $i \in \mathbb{N}$. This means that, although we will simply call them "Hypothesis H_i", they are actually null hypotheses.
For convenience, the context of all of these hypotheses will always be that of the data extracted from the central task (see Section 2.6).

- Hypothesis H_1: Additional visual elements have no effect on performance.
- Hypothesis H_2: Additional visual elements have no effect on mental load.
- Hypothesis H_3: Higher task difficulty have no effect on performance.
- Hypothesis H_4: Higher task difficulty have no effect on mental load.
- Hypothesis H_5: Physical environments are not more "immersive" than XR environments.
- Hypothesis H_6: Physical environments are not more "usable" than XR environments.
- Hypothesis H_7: A higher task load has no effect on performance.
- Hypothesis H_8: No difference under any of the metrics can be found between physical removal and removal using DR.

2.6 Main task

The design of the central task in our study was informed by a number of key considerations. The primary objective was to examine the influence of visual elements on performance and cognitive load in a mentally challenging task. In order to achieve this, it was necessary to devise a task that would be cognitively demanding and could be performed within a video see-through environment. The task was designed to require rapid and accurate responses to changing visual stimuli, allowing for the accurate measurement of the effects of visual elements on user performance and perceived mental load in the process. Further, to study the impact of visual elements, it was necessary to introduce additional visual stimulation that could potentially affect the user simultaneously, along with the main task. This resulted in the design of an exercise that could be confined to a predetermined area, such as a desk or screen, in order to prevent the aforementioned additional elements from being pushed outside of the user's field of view due to lack of room. To further challenge the participants and measure their performance under increasing cognitive load, a task was designed whose difficulty could be readily and rapidly increased without requiring a termination or pause of the task, thus avoiding the prolongation of an already monotonous user study. This approach would permit more significant metrics for error rates, allowing a study based on different difficulty conditions, and facilitate a more nuanced comprehension of the influence of visual elements on task performance. In addition, in order to accurately assess the impact of visual elements on task performance, it was necessary to devise a task that could be evaluated precisely and objectively. Finally, the study needed to be conducted in a controlled environment, where most extraneous variables could be controlled, and which was somewhat familiar for the user, otherwise no see-through illusion would have been possible. Therefore, participants were seated at a standard office desk, and the task design was tailored to fit this setting.

For the aforementioned reasons, a clicking task over buttons on a screen was selected, as it is simple and requires little explanation for the participants to understand, represents well a generic activity on a desktop screen such as navigating through a folder structure, allows easy extrapolation of data from it for a performance evaluation which can be readily analysed, as well as providing the possibility of using standard metrics for evaluation, such as Fitts' Law. Its level of difficulty and, thus, mental load, could be easily adjusted as needed by activating multiple buttons at once, forcing the participant to perform a visual search for the button that should be pressed instead of simply relying on their attention being caught by the change in appearance, and permits showing additional visual elements in the user's peripheral vision, beside the area dedicated to the task, which is still clear perceived as the focus of the experiment.

2.6.1 Implementation

Thanks to the previously-mentioned considerations, the final implementation of the central task displayed a 10x10 matrix of buttons, each labeled with a two-digit integer $n \in [10, 99]$ and layed-out randomly. The beginning of the task was signaled by the participant clicking a designated 'Start' button on their screen.

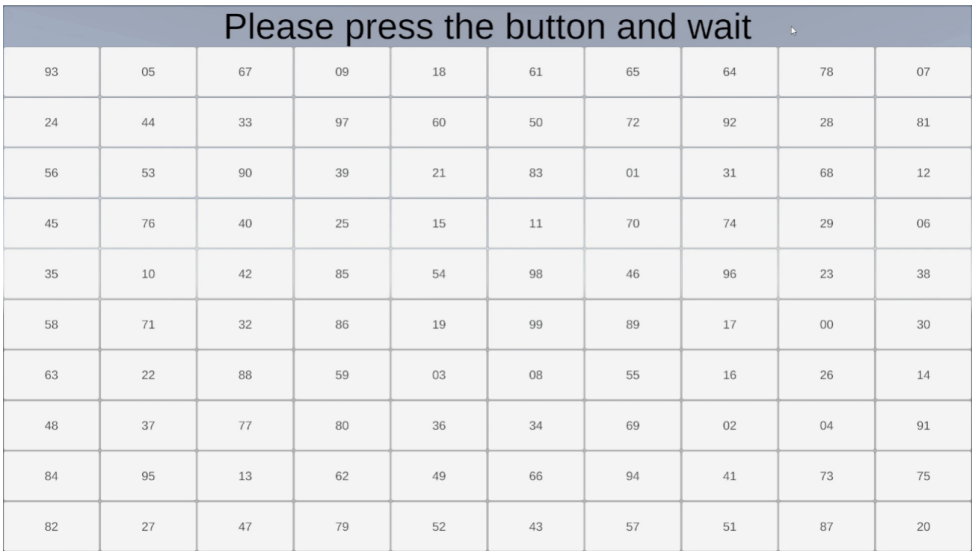


Figure 2.3: *Visual representation of the clicking task before the first button had activated.*

Upon initiation of the task, which displayed a grid similar to that in Figure 2.3, a five-second interval elapsed before a single button on the matrix, uniformly at random, underwent a color transformation, indicating its 'activation', as visible in Figure 2.4. Participants were required to respond by clicking the activated button within a three-second timeframe.

Successful interaction with the activated button, or the expiration of the three-second limit, resulted in the deactivation of the current button (signified by a reversion to its original color scheme) and the subsequent activation of a new random button. This cycle of activation and deactivation was repeated, with participants tasked with interacting with each activated button. Notably, there was no penalty imposed for incorrect clicks, although such instances were recorded as errors.

Figure 2.4: *Visual representation of the clicking task when one button activated at once.*

Following the 50th cycle of activation and deactivation, the task parameters were modified. Instead of a

single button activation, two buttons were simultaneously activated. A directive, displayed at the top of the screen, instructed participants to click a specific one, identified by its label (n), as shown in Figure 2.5. It is important to note that only the correct interaction resulted in the deactivation of the two buttons, while an incorrect click, even on an active button, was registered as a misclick. This modified procedure was repeated for a further 25 cycles.

Figure 2.5: *Visual representation of the clicking task when two buttons activated at once.*

Upon completion of a total of 75 activation cycles, the task under the specific condition was concluded.

It is crucial to emphasise that no button could be activated on two occasions throughout the entirety of a single run of the task, not even between different button conditions (e.g. once for the single-button phase and once for the double-button phase). This was done to prevent any potential for a favourable performance due to a specific button which happens to be in a convenient location and activates two or more times.

2.7 Study design

In this chapter, we delve into the design of the user study conducted for this research.

2.7.1 Conditions

A $2 \times 2 \times 2$ within-subject design was implemented for the following independent variables, each with 2 possible levels:

- Physical/XR: whether the environment is visualised through the Meta Quest 3 HMD (XR) or not (Physical).
- Visual elements/Diminished visual elements: whether the environment is presented with additional visual elements (Visual elements) or they are removed/hidden (Diminished visual elements).
- 1-button/2-button activation: whether in the clicking task there is a single button activating at once

(1-button activation) or two (2-button activation).

These 3 independent variables lead to the 8 conditions as shown in the following table:

Platform condition	Visual elements 1-button activation	Visual elements 2-button activation	No visual elements 1-button activation	No visual elements 2-button activation
Physical	P1B	P2B	P1B-D	P2B-D
XR	XR1B	XR2B	XR1B-D	XR2B-D

Table 2.2: *Within-subject study design conditions table. The two rows define the level for the* Physical/XR *independent variable, whereas each column defines one combination of levels for the* Visual elements/Diminished visual elements *and* 1-button/2-button activation *independent variables. The intersection of a row and a column uniquely identifies a combination of the 3 independent variable, and thus a study condition under which the participant performed the task. Icons have been added for additional clarity, and for future reference.*

In Table 2.2, each possible level for each independent variable has been assigned a textual value, as follows:

- Physical/XR: *P* or *XR*, respectively.
- 1-button/2-button activation: *1B* or *2B*, respectively.
- Diminished visual elements: *-D*, otherwise the condition is assumed to contain visual elements.

To provide an example, “P1B” stands to identify the condition in a physical environment, with additional visual elements and 1-button activation, whereas “XR2B-D” represents the condition in an XR environment, where the additional visual elements have been diminished, and 2 buttons activate at once.

A within-subject design was chosen because of the large number of conditions, , which would have required a substantial number of participants, the availability of which could not be guaranteed a priori. Since learning effects for analogous task are unlikely [49], and made even less likely by the randomness of the buttons’ disposition and activation order, a within-subject design was deemed the most appropriate choice.

2.7.2 Metrics

The following metrics were collected as dependent variables, either by recording them in real-time as the participants performed the clicking task, or through participants’ self-reports:

- Objective data:
  - Number of timeouts: Amount of times that a participant took longer than 3 seconds to click on the active button.
  - Number of misclicks: Amount of times that a participant clicked on a button that was inactive, and they were thus not expected to press it.

  - Per-button time elapsed: Time elapsed from the moment of the target button's activation to the moment of its deactivation (either because of a participant's click or because of a timeout).
  - Per-button distance to target: Distance from the mouse pointer to the target button at the moment of its activation.
- Subjective data:
  - Simulator Sickness Questionnaire (SSQ) [50]
  - NASA Task Load Index (NASA-TLX) [51]
  - System Usability Scale (SUS) [52]
  - Cognitive Absorption Questionnaire (CAQ) [53]
  - Personal ranking of the conditions
  - Perceived level of distraction

The objective data is primarily recorded to examine the user's performance across the eight conditions. In particular, the number of timeouts and misclicks serve as an indicator of the error rate associated with the task. Additionally, the time elapsed and distance to the target can be used to assess overall performance, identify potential learning effects, analyse the participants' behaviour within the task, and a way to evaluate this task against Fitts' Law. In contrast, the subjective data were employed to gauge the impact of the visual elements on task load, along with the participants' preferences and their perceived experience. This enabled a comparison of the physical and XR conditions and an evaluation of the potential of VST and DR as technologies, particularly in relation to applications such as this one, with a view to their future use in everyday tasks. The reporting of subjective data was limited to a single instance following the completion of both the single-button and double-button activation phases of each condition of the clicking task. Each questionnaire was thus completed a total of four times, with the exception of the SSQ, which was requested before the study, as well. This was due to two key factors. Firstly, it mitigates the potential for participant fatigue by minimizing the number of times they are required to complete questionnaires, which can be both time-consuming and mentally exhausting, as well as risking jeopardising the quality of the results due to fatigue or annoyance. This approach ensures that participants remain engaged and provide high-quality responses. Secondly, it allows for a more holistic evaluation of their overall experience. By reflecting on both parts of the task together, participants can offer more comprehensive and integrated feedback on their cognitive load, usability perceptions, and overall engagement. This method enables them to compare and contrast their experiences across the two parts, leading to more meaningful and insightful responses. In contrast, the personal ranking of the conditions and the perceived level of distraction were only requested once the task had been completed under all conditions. Specifically, the same questions were posed twice: initially for single-button conditions and subsequently for double-button conditions.

2.7.3 User study

Given the within-subject design, meaning that all participants performed the task under all conditions, counterbalancing was necessary to avoid learning effects. Because of the large number of conditions, which would otherwise cause an explosion in the number of possible orderings ($8! = 40‘320$), we opted for incomplete counterbalancing by randomising the orders in the following way:

1. A fair coin was tossed at the beginning of the study, which would decide whether the participant would perform under the physical conditions first (🖵), or the XR conditions first (🥽). Let us call

the first condition resulting from the coin toss "Condition A", and the second one "Condition B".

2. Within condition A, a second fair coin toss decided whether the participant would begin by performing the task with the additional visual elements or the one where they were diminished. Let us call the first condition resulting from this second coin toss "Condition A_1", and the second one "Condition A_2".

3. Once both conditions A_1 and A_2 were terminated and the participant had filled out both conditions' relevant questionnaires, they would move onto condition B.

4. Within condition B, a third fair coin was tossed to decide whether the participant would begin by performing the task with the additional visual elements or the one where they are diminished. Let us call the first condition resulting from this last coin toss "Condition B_1", and the second one "Condition B_2".

5. Once condition B_2 was over and the participant had filled out the subjective questionnaires, they were then asked to rate the four conditions twice: once for 1-button activation conditions, and a second time for 2-button activation conditions. Then, finally, they would be asked how distracted they felt by the CAAV clips during the clicking task, once for 1-button activation conditions, and a second time for 2-button activation conditions.

For a flow chart representation of the counterbalancing process, see Appendix A.1

Finally, the single conditions were setup as follows. 3 regular PC monitors were placed on the desk that was scanned to produce the mesh used for the DR implementation (see Section 2.4). On the central one, an instance of the clicking task would be presented to the user, ready for their initial click. The two side screens would either show randomly-sampled clips from the CAAV database in case the condition included visual elements (🖼), or, in case the condition included the diminishing of visual elements (🖼), they would be physically hidden by placing them under the desk in physical conditions (🖵), or by superimposing the scanned mesh of the desk and the wall in XR conditions (🥽).

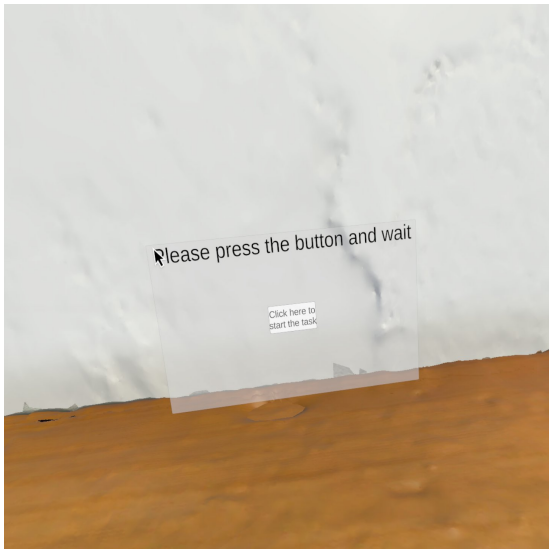


Figure 2.6: *View of the study environment through the Meta Quest 3 HMD from the point of view of a participant before beginning the clicking task in an XR condition without additional visual elements, which have been diminished by the aforementioned mesh of the desk and wall.*

Furthermore, in XR conditions (), the clicking task and the visual elements would not appear on the physical monitor as they would for physical ones (), but they would rather be presented on a virtual panel that was registered to the surface of the physical monitors by hand by the study supervisor, simply for a matter of comparability of visual acuity between physical and XR conditions. Other differences between the physical and XR conditions consist of the input device that was used, where a heavier bluetooth mouse was connected to the headset, but a USB mouse was attached to the desktop, the coordinate system used for the mouse and the button, where the physical conditions used absolute screen coordinates, whereas XR conditions used internal Unity coordinates, the visual acuity, where text at the same distance would normally be perceived as sharper and better defined on a desktop screen compared to a panel through the HMD. Furthermore, the HMD's weight was felt on one's head throughout all XR conditions, which was not the case in physical ones.

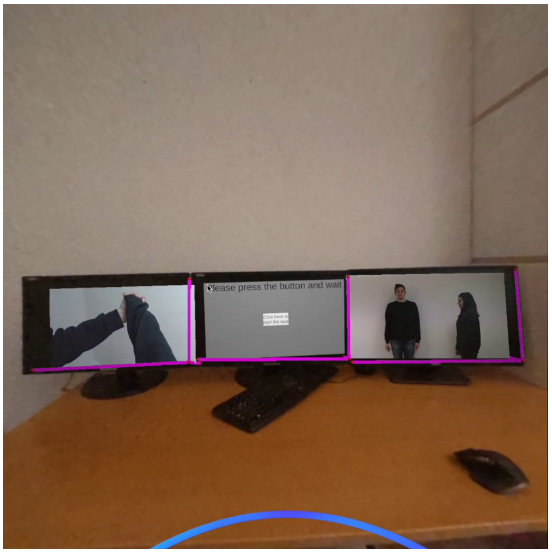


Figure 2.7: *View of the study environment through the Meta Quest 3 HMD partway through the setup for the XR condition with diminished visual elements.*

Under these conditions, the participants would perform a single run of the clicking task (comprising the 75 clicks as described in section Section 2.6.1) and then fill out the SSQ, NASA-TLX, SUS and CAQ.

3

Results and Discussion

In this chapter, we present the findings derived from the user study conducted throughout this research. We recruited 23 participants to perform our user study, one of which had to be excluded. It was found that the number of misclicks was significantly impacted by the XR conditions, with an average of 10.55% more misclicks on average compared to physical ones.

The counterbalancing was found to be imbalanced. As demonstrated in the table in Appendix A.2, some orderings had four or even five participants assigned to them, whereas others had as few as one.

Finally, in the context of the study that was run, where multiple dependent variables and conditions were had, and participants performed the same task under different conditions, Friedman's test was found to be well-suited because capable of comparing the differences between these conditions without assuming a normal distribution of the data, as it is a non-parametric statistical test that is ideal for comparing more than two groups that are related. For post-hoc analysis, a Benjamini-Hochberg FDR Corrected Conover's test was run to compute the post-hoc comparison matrix. Benjamini-Hochberg FDR Correction controls the expected proportion of false discoveries among the rejected hypotheses and is less conservative than other procedures (like Bonferroni correction), thus increasing the test's power, which makes it a good fit when conducting multiple simultaneous tests, as in the case of post-hoc analysis following a Friedman's test.
In instances where tests for differences between datasets encompassed multiple conditions (for example, a comparison of all physical data with all XR data), Bonferroni-corrected Wilcoxon signed-rank tests were instead performed. The correction method choice is due to the fact that Wilcoxon signed-rank tests are used to make multiple comparisons across different conditions where it is more important to control the risk of false positives than it is to control the risk of false negatives, making the Bonferroni correction a suitable choice.

3.1 Demographics

The data was collected from 23 participants with ages ranging between 20 and 62 years (median $\mu \approx 25.36$, standard deviation $\sigma \approx 8.83$). 18 were males, 4 were females and 1 was agender. The participants' experience with VR and XR equipment was somewhat skewed towards the extremes, with the highest and lowest experience categories accounting for the majority of participants (65.2%), while the middle and second highest/lowest categories were less represented (refer to Figure 3.1). It is important to note that one participant did not submit their data, which was therefore registered incompletely. Consequently, their input was not taken into consideration for any of the subsequent analyses.

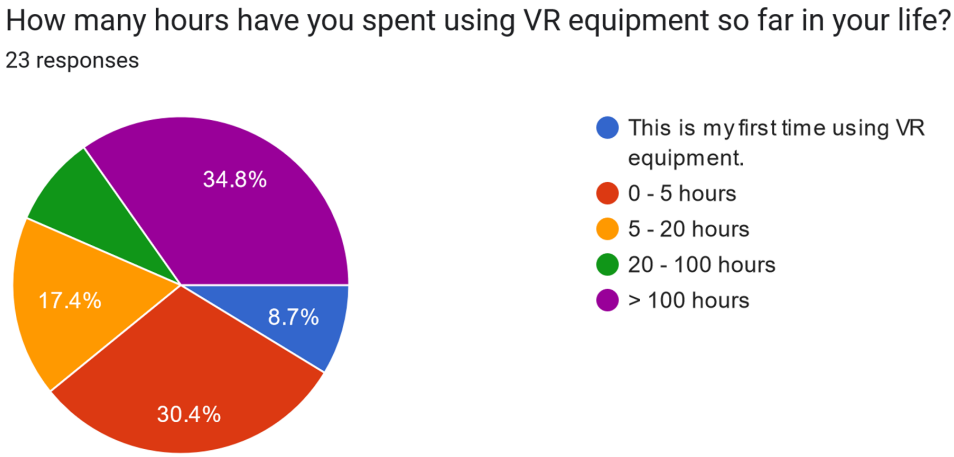


Figure 3.1: *Distribution of VR experience among participants.*

3.2 Objective Data

This section will deal with the results and the data analysis of the non-self-reported metrics: number of timeouts, number of misclicks, per-button time elapsed, and per-button distance to target (see Section 2.7.2).

3.2.1 Number of Timeouts

A Shapiro-Wilk test revealed that the timeouts data was not normally distributed ($p \in \{7.42 \cdot 10^{-10}, 3.32 \cdot 10^{-9}, 7.49 \cdot 10^{-8}, 1.20 \cdot 10^{-7}\}$ depending on the exact condition for the 1-button activation conditions' data (🍒); $p \in \{9.44 \cdot 10^{-8}, 2.66 \cdot 10^{-8}, 2.36 \cdot 10^{-8}, 2.15 \cdot 10^{-8}\}$ depending on the exact condition for the 2-button activation conditions' data (🍒🍒)). Levene's test showed homogeneity ($W \approx 0.09, p \approx 0.96$ for 1-button activation conditions' data (🍒), $W \approx 1.63, p \approx 0.19$ for 2-button activation conditions' data (🍒🍒)). This required the use of non-parametric statistical tests. Friedman's test showed no significant differences, either in the 1-button activation conditions' data (🍒, $p \approx 0.56$), or in the 2-button activation conditions' data (🍒🍒, $p \approx 0.36$).

For the complete data, see Figure A.2.

Explanation

Friedman's test did not find any significant differences in the number of timeouts across different conditions. This may be for a few possible reasons. For instance, the task of clicking on an active button within 3 seconds might not have been challenging enough to create significant differences between conditions, so if the task was relatively easy for all participants, they might have consistently performed well regardless of the condition. The data seems to accredit this theory, as the condition with the highest amount of timeouts, the XR condition without additional visual elements and with 2 buttons activating at once (), still saw only 5 participants out of 23 receiving timeouts, compared to 1 participant with timeouts under the condition with the smallest number of them (the physical condition with additional visual elements and with only 1 active button,).

3.2.2 Number of Misclicks

The first important result to remark is that for conditions with 2 active buttons () **no misclicks were recorded**, meaning that the additional button activation likely made people pay more attention to their clicks. A Shapiro-Wilk test revealed that the misclicks data was not normally distributed ($p \in \{1.01 \cdot 10^{-3}, 4.37 \cdot 10^{-3}, 3.77 \cdot 10^{-2}, 0.12\}$ depending on the exact condition for the 1-button activation conditions' data ()). Levene's test showed homogeneity ($W \approx 12.05, p \approx 1.00 \cdot 10^{-6}$ for 1-button activation conditions' data ()). This required the use of non-parametric statistical tests. Friedman's test showed significant differences (, $p \approx 2.53 \cdot 10^{-10}$). Post-hoc Benjamini-Hochberg Corrected Conover's tests found the following significant differences:

- When comparing physical () vs XR () conditions with visual elements () and 1-button activations (), a significant difference was found of 7.91 misclicks more under XR conditions, on average ($p \approx 9.51 \cdot 10^{-10}$)
- When comparing non-diminished () vs diminished () physical conditions () with 1-button activations (), a significant difference was found of 7.45 misclicks more under diminished conditions, on average ($p \approx 4.01 \cdot 10^{-9}$)

Post-hoc Bonferroni Corrected Wilcoxon signed-rank tests found the following significant differences:

- When comparing the average difference in misclicks per-partecipant between non-diminished () and diminished () conditions for physical () vs XR () conditions, thus studying whether there is a difference between physical removal and DR removal, a significant difference was found of 3.18 misclicks more on average for removal under XR conditions ($p \approx 0.013$).

For the complete data, see Figure A.3, and for the complete results of the pairwise post-hoc tests, see Appendix A.5.

Explanation

The task of clicking on an active button within a 3-second window might not have been overly complex for the participants. This could explain why there were no misclicks recorded for conditions with 2 active buttons. Interestingly, the additional cognitive load of having to identify and click on two active buttons may have increased participants' focus and precision, leading to fewer errors.

On the other hand, the significantly higher number of misclicks under XR conditions with visual elements and 1-button activations compared to physical conditions could be attributed to the immersive nature of

the XR environment. While XR environments can provide a high degree of immersion, they can also introduce additional challenges or distractions. For example, participants might have needed time to adjust to the XR environment or they might have been distracted by the novelty or complexity of the environment.

Furthermore, the significantly higher number of misclicks under diminished conditions in physical environments with 1-button activations compared to non-diminished conditions suggests that the removal or hiding of visual elements might have disrupted participants' spatial awareness or focus. Visual elements can provide important spatial cues and removing them might have made the task more challenging.

Lastly, the significant difference in misclicks between non-diminished and diminished conditions for physical vs XR conditions suggests that the process of removing or diminishing visual elements in an XR environment might be more disorienting or challenging for participants compared to doing so in a physical environment. This could be due to differences in how visual information is presented and processed in physical vs. XR environments.

3.2.3 Fitts' Law Analysis

Performing linear regression on the data using the index of difficulty (ID) , computed using the Shannon formulation of Fitts' Law [54]

$$\text{ID} = \log_2\left(\frac{D}{W} + 1\right),$$

for a distance from the target D and the target's width W, as a feature $x \in \mathbb{R}$, and the time elapsed to click on the active, target button (TT) as a label $y \in \mathbb{R}$, thus estimating parameters $(a, b) \in \mathbb{R}^2$ such that

$$\text{TT} \approx a \cdot \text{ID} + b = a \cdot \log_2\left(\frac{D}{W} + 1\right) + b,$$

yielded an R^2 accuracy score between 0% and 7% depending on how the data was split (with the highest score, 7% being obtained from the data corresponding to the XR condition with 1-button activation and diminished visual elements (🥽🖼️🍒)). This seemed to suggest that the relationship between the time required to click on a target button and the index of difficulty is unlikely to be linear.

For further information, please refer to Appendix A.6.

Explanation

The low R^2 values obtained from the linear regression analysis suggest that the relationship between the index of difficulty (ID) and the time to click (TT) in our dataset is not well-captured by Fitts' Law. This may be attributed to various factors, including noise in the data, which can arise from participant variability, experimental design, and measurement errors. This seems to suggest that the assumption of a linear relationship between ID and TT may not hold, and a non-linear model may be more suitable. Additionally, the range of IDs in our dataset may be insufficient to capture the underlying pattern, and task-specific factors such as target size from the user's point of view, which varied based on whether the task was performed through the HMD (🥽) or through the participant's own eyes (🖵), visual feedback due to the additional visual elements, and cognitive load may also influence the relationship between ID and TT. Moreover, participants may employ different strategies to complete the task, which can affect the time to click, and Fitts' Law may not capture the nuances of human behavior in our specific task.

3.2.4 Per-button time elapsed

A Shapiro-Wilk test revealed that the data on the time required to click on the target button was not normally distributed ($p \in \{2.75 \cdot 10^{-31}, 7.27 \cdot 10^{-29}, 2.38 \cdot 10^{-25}, 1.69 \cdot 10^{-19}, 7.92 \cdot 10^{-14}, 8.85 \cdot 10^{-13}, 1.33 \cdot 10^{-10}, 4.76 \cdot 10^{-05}\}$ depending on the exact condition). Levene's test showed inhomogeneity ($W \approx 28.71, p \approx 2.82 \cdot 10^{-39}$). This required the use of non-parametric statistical tests. Friedman's test showed significant differences ($p \approx 7.99 \cdot 10^{-25}$). Post-hoc Benjamini-Hochberg Corrected Conover's tests found the following significant differences:

- When comparing physical (🖵) vs XR (🥽) conditions with visual elements (🖼) and 1-button activations (🍒), a significant difference was found of 52.50 milliseconds more under XR conditions, on average ($p \approx 10.00 \cdot 10^{-4}$)
- Analogously, when comparing physical (🖵) vs XR (🥽) conditions without visual elements (🚫🖼) and 1-button activations (🍒), a significant difference was found of 193.47 milliseconds more under XR conditions, on average ($p \approx 3.86 \cdot 10^{-71}$)
- Analogously, when comparing physical (🖵) vs XR (🥽) conditions without visual elements (🚫🖼) and 2-button activations (🍒🍒), a significant difference was found of 168.29 milliseconds more under XR conditions, on average ($p \approx 8.23 \cdot 10^{-10}$)
- When comparing non-diminished (🖼) vs diminished (🚫🖼) conditions with physical (🖵) and 1-button activations (🍒), a significant difference was found of 17.58 milliseconds more under non-diminished conditions, on average ($p \approx 1.83 \cdot 10^{-2}$)
- However, when comparing non-diminished (🖼) vs diminished (🚫🖼) conditions with XR environments (🥽) and 1-button activations (🍒), a significant difference was found of 123.64 milliseconds more under diminished conditions, on average ($p \approx 1.84 \cdot 10^{-34}$)
- Similarly, when comparing non-diminished (🖼) vs diminished (🚫🖼) conditions with XR environments (🥽) and 2-button activations (🍒🍒), a significant difference was found of 121.21 milliseconds more under diminished conditions, on average ($p \approx 9.22 \cdot 10^{-6}$)

Post-hoc Bonferroni Corrected Wilcoxon signed-rank tests found the following significant differences:

- When comparing 1-button activation (🍒) and 2-button activation (🍒🍒) conditions, a significant difference was found of 412.87 milliseconds more under 2-button activations conditions, on average ($p \approx 8.93 \cdot 10^{-214}$).
- When comparing non-diminished (🖼) and diminished (🚫🖼) conditions, a significant difference was found of 53.4 milliseconds more under diminished conditions, on average ($p \approx 1.21 \cdot 10^{-12}$)

For the complete results of the pairwise post-hoc tests, see Figure A.3.

Explanation

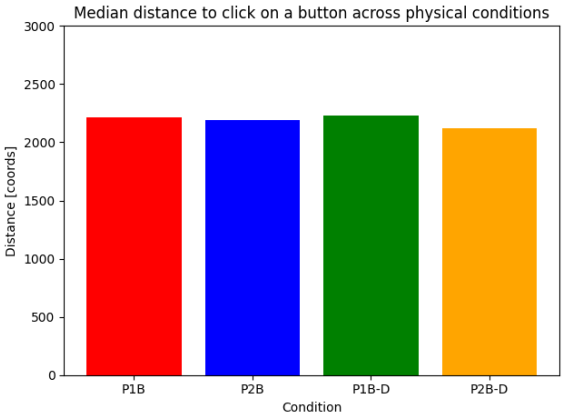
Participants took significantly longer to click on the target button under XR conditions with visual elements and 1-button activations compared to physical conditions. This could suggest that the immersive nature of the XR environment might have introduced additional challenges or distractions that increased the time it took participants to respond. Similarly, participants took significantly longer under diminished conditions in physical environments with 1-button activations compared to non-diminished conditions. This suggests that the removal or hiding of visual elements might have disrupted participants' spatial awareness or focus, leading to slower response times.

Interestingly, when comparing non-diminished and diminished conditions for physical vs XR conditions, participants took significantly longer under diminished conditions in the XR environment. This could indicate that the process of removing or diminishing visual elements in an XR environment using See-Through DR might be more disorienting or challenging for participants compared to doing so in a physical environment.

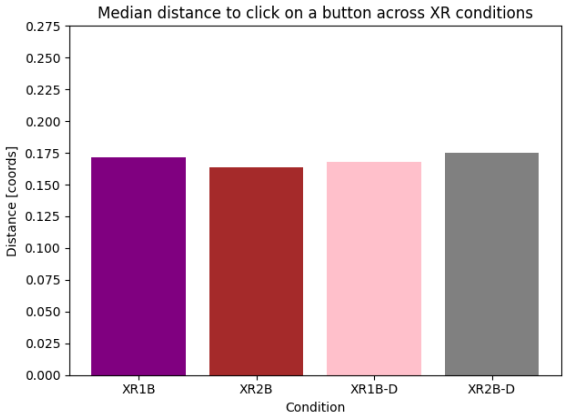
Finally, participants took significantly longer under 2-button activation conditions compared to 1-button activation conditions, and under diminished conditions compared to non-diminished conditions. These findings suggest that both the complexity of the task (i.e., the number of buttons to be activated) and the presence of visual elements can significantly impact the time it takes participants to respond.

3.2.5 Per-button distance to target

An important detail to note is the difference in coordinates systems between the physical (🖵) conditions (which based the mouse's and buttons' coordinates on absolute points of reference based on the physical monitor) and XR (🥽) conditions (which based the mouse's and the buttons' coordinates on relative points of reference based on the parent panel on which all buttons were placed). For this reason the data analysis was split into two data subsets for data gathered under physical and XR conditions, respectively.



(a) Median distance from the target button at its activation, divided by condition (Physical conditions).



(b) Median distance from the target button at its activation, divided by condition (XR conditions).

Figure 3.2: *Bar graphs illustrating the difference in absolute values when comparing the measured distance between the physical and XR conditions, making them incomparable.*

A Shapiro-Wilk test revealed that the data on the distance from the mouse pointer to the active target button at the moment of its activation was not normally distributed ($p \in \{5.02 \cdot 10^{-16}, 8.96 \cdot 10^{-15}, 2.22 \cdot 10^{-13}, 9.55 \cdot 10^{-13}, 2.11 \cdot 10^{-11}, 8.58 \cdot 10^{-11}, 3.20 \cdot 10^{-09}, 3.27 \cdot 10^{-09}\}$ depending on the exact condition). Levene's test showed inhomogeneity ($W \approx 1284.92, p \approx 0.00$). This required the use of non-parametric statistical tests. Friedman's test showed significant differences ($p \approx 5.29 \cdot 10^{-23}$). Post-hoc Benjamini-Hochberg Corrected Conover's tests found the following significant differences:

- When comparing 1-button activation (🍒) vs 2-button activation (🍒🍒) conditions with physical environments (🖵) without visual elements (🖂), a significant difference was found of 63.72 pixels more under 1-button activation conditions, on average ($p \approx 8.81 \cdot 10^{-3}$)

For the complete results of the pairwise post-hoc tests, see Figure A.4.

Explanation

Participants had a significantly larger distance to cover under 1-button activation conditions in physical environments without visual elements compared to 2-button activation conditions. This could suggest that the additional cognitive load of having to identify and click on one of two active buttons may have led participants to position their mouse pointer in a more central location, reducing the average distance to the active buttons.

3.3 Subjective Data

This section will deal with the results and the data analysis of the self-reported metrics: SSQ, NASA-TLX, SUS, CAQ, Personal ranking of the conditions, Perceived level of distraction (see Section 2.7.2). It is necessary to provide a brief explanation for the statistical tests that have been conducted in this section. Some datasets will be found to be normally distributed, but non-parametric tests will be used anyway, as they require the fewest assumptions, simply to avoid the inconvenience of running different tests on different splits of the same dataset, as well as running the risk of no longer being able to compare results for specific questionnaires because different tests with different statistical powers have been used. Finally, before performing any sort of statistical test on the data derived from the SSQ, NASA-TLX, SUS, or CAQ, their scores were computed as described in their respective papers. Each section will provide a concise overview of the specific computation.

For a simpler visualisation of the data, please refer to Figure A.5

3.3.1 Simulator Sickness Questionnaire

Scores for individual reports of the SSQ were computed according to the original paper's formula [50]. Each symptom's rating was assigned to one of three categories: *Nausea*, *Oculomotor*, and *Disorientation*. For each category, the sum of the reported scores for each symptom was calculated, and then all three were summed once more, this time with weights of 1.0, 1.0, and 3.74, respectively.

A Shapiro-Wilk test revealed that the data gathered from the SSQ was not normally distributed ($p \in \{9.77 \cdot 10^{-4}, 3.20 \cdot 10^{-3}, 7.18 \cdot 10^{-3}, 1.38 \cdot 10^{-2}\}$ depending on the exact condition). Levene's test showed homogeneity ($W \approx 0.52, p \approx 0.67$). This required the use of non-parametric statistical tests. Friedman's test showed no significant differences ($p \approx 6.71 \cdot 10^{-2}$).

3.3.2 NASA Task Load Index

Scores for individual reports of the NASA-TLX were computed according to the original paper's formula [51]. The average of all questions' score was taken.

A Shapiro-Wilk test revealed that the data gathered from the NASA-TLX was mostly normally distributed ($p \in \{5.13 \cdot 10^{-2}, 9.16 \cdot 10^{-2}, 0.22\}$ depending on the exact condition), except for the physical condition (🖵) with additional visual elements (🖻, $p \approx 4.67 \cdot 10^{-2}$). Levene's test showed homogeneity ($W \approx 0.11, p \approx 0.96$). This required the use of non-parametric statistical tests. Friedman's test showed no significant differences ($p \approx 0.85$).

3.3.3 System Usability Scale

Scores for individual reports of the SUS were computed according to the original paper's formula [52]. Since even-numbered questions are inverted (expressing a lack of usability, instead), their scores needed to be inverted accordingly. Thus, the score for even questions was computed as $5 - r_i$, where r_i was the reported value for $i \in [1, 10], i \bmod 2 = 0$, whereas the score for odd-numbered questions was $r_j - 1$, where r_j was the reported value for $j \in [1, 10], j \bmod 2 = 1$. Afterwards, the scores were summed, and the total multiplied by 2.5.

A Shapiro-Wilk test revealed that the data gathered from the SUS was mostly non-normally distributed ($p \in \{2.89 \cdot 10^{-4}, 2.55 \cdot 10^{-3}, 3.54 \cdot 10^{-2}\}$ depending on the exact condition), except for the XR condition (🥽) with diminished visual elements (🖼, $p \approx 0.16$). Levene's test showed homogeneity ($W \approx 1.55, p \approx 0.21$). This required the use of non-parametric statistical tests. Friedman's test showed significant differences ($p \approx 1.08 \cdot 10^{-3}$). Post-hoc Benjamini-Hochberg Corrected Conover's tests found the following significant differences:

- When comparing physical (🖵) vs XR (🥽) conditions with diminished visual elements (🖼), a significant difference was found of 12.5 points (on a 0-100 scale) more under physical conditions, on average ($p \approx 4.99 \cdot 10^{-2}$)

Post-hoc Bonferroni Corrected Wilcoxon signed-rank tests found the following significant differences:

- When comparing physical (🖵) vs XR (🥽) conditions, a significant difference was found of 8.47 points (on a 0-100 scale) more under physical conditions, on average ($p \approx 1.59 \cdot 10^{-2}$)

Explanation

Participants rated the system as significantly more usable under physical conditions compared to XR conditions, both when visual elements were present and when they were diminished. This could suggest that participants found the physical environment more intuitive or comfortable to interact with, or that the XR environment introduced additional challenges or complexities that affected perceived usability.

Interestingly, when visual elements were diminished, the difference in usability between physical and XR conditions was even more pronounced. This could indicate that the process of removing or diminishing visual elements in an XR environment using DR might have been more disorienting or challenging for participants, further impacting perceived usability.

3.3.4 Cognitive Absorption Questionnaire

Scores for individual reports of the CAQ were computed according to the original paper's formula [53]. The 4th question was negatively worded, so its scale was reversed by taking its value r_4 for each participant and computing $r'_4 = 8 - r_4$, then the average of all questions' score was taken.

A Shapiro-Wilk test revealed that the data gathered from the SUS was mostly non-normally distributed ($p \in \{2.30 \cdot 10^{-5}, 1.78 \cdot 10^{-3}, 3.36 \cdot 10^{-2}\}$ depending on the exact condition), except for the XR condition (🥽) with additional visual elements (🖼, $p \approx 0.12$). Levene's test showed homogeneity ($W \approx 0.56, p \approx 0.64$). This required the use of non-parametric statistical tests. Friedman's test showed significant differences ($p \approx 3.30 \cdot 10^{-4}$). Post-hoc Benjamini-Hochberg Corrected Conover's tests found no further significant differences; however, post-hoc Bonferroni Corrected Wilcoxon signed-rank tests found the following:

- When comparing non-diminished (🖼) vs diminished (🖼) conditions, a significant difference was found of 0.51 points (on a 0-7 scale) more under diminished conditions, on average ($p \approx 9.10 \cdot 10^{-3}$)

Explanation

Participants reported significantly higher cognitive absorption under diminished conditions compared to non-diminished conditions. This could suggest that the removal or hiding of visual elements might have led to increased focus and immersion in the task, thereby enhancing cognitive absorption. However, it's important to note that no further significant differences were found in the post-hoc Benjamini-Hochberg Corrected Conover's tests. This could indicate that while the presence or absence of visual elements can impact cognitive absorption, other factors such as the physical vs. XR environment or the number of active buttons might not have a significant effect.

3.3.5 Conditions Ranking

From the participants' reported data (see Appendix A.10), we can clearly see that the physical condition with diminished visual elements (🖥🖼) was the clear favourite, and the XR condition with additional visual elements (🥽🖼) was the least favourite. In order to quantify the data in a rigorous manner, an Inverse Weighted Sum (see Section A.11) was used to compute a score that would allow the comparison of the participants' preferences on a 22-88 scale. This yielded the following rankings:

- 1-button activations
    1. Physical with diminished visual elements: 82/88 points
    2. XR with diminished visual elements: 60/88 points
    3. Physical with additional visual elements: 47/88 points
    4. XR with additional visual elements: 31/88 points
- 2-button activations
    1. Physical with diminished visual elements: 85/88 points
    2. XR with diminished visual elements: 56/88 points
    3. Physical with additional visual elements: 48/88 points
    4. XR with additional visual elements: 31/88 points

For further information on the scores, see Figure A.6.

3.3.6 Perceived distraction

As indicated by the self-reports, the majority of participants did not perceive the additional visual elements as a source of distraction. The majority of the people who participated rated their perceived level of distraction as 1 or 2, especially for 1-button activation conditions.

A Shapiro-Wilk test revealed that the data self-reported by the participants' on their perceived distraction by the visual elements was non-normally distributed ($p_{\text{sb}} \approx 3.41 \cdot 10^{-5}, p_{\text{db}} \approx 7.97 \cdot 10^{-4}$). Levene's test

showed homogeneity ($W \approx 1.45, p \approx 0.24$). This required the use of non-parametric statistical tests. However, Friedman's test showed no significant differences ($p \approx 0.16$).

For further information on the scores, see Figure A.7.

3.4 Hypothesis Verification

- Hypothesis H_1: Additional visual elements have no effect on performance.
  - Outcome: **Rejected**
  - Findings show that additional visual elements had an effect on the number of misclicks (see Section 3.2.2)
- Hypothesis H_2: Additional visual elements have no effect on mental load.
  - Outcome: **Failed to reject**
- Hypothesis H_3: Higher task difficulty have no effect on performance.
  - Outcome: **Rejected**
  - Findings show that higher task difficulty (in this case increase the number of active buttons from 1 to 2) has an effect on the time required to complete the task (see Section 3.2.4)
- Hypothesis H_4: Higher task difficulty have no effect on mental load.
  - Outcome: **Failed to reject**
- Hypothesis H_5: Physical environments are not more "immersive" than XR environments.
  - Outcome: **Failed to reject**
- Hypothesis H_6: Physical environments are not more "usable" than XR environments.
  - Outcome: **Rejected**
  - Findings show that physical conditions were reported to be more usable on the SUS.
- Hypothesis H_7: A higher task load has no effect on performance.
  - Outcome: no difference in task load was found, and we could thus not be sure.
- Hypothesis H_8: No difference under any of the metrics can be found between physical removal and removal using DR.
  - Outcome: **Rejected**
  - Findings show that DR removal has more misclicks and a higher time to click on a button compared to physical removal.

3.5 Limitations

The research conducted presents several limitations that should be taken into account when interpreting the results. Firstly, the use of different mice for the clicking tasks in physical and XR conditions could

have influenced the participants' performance. The physical clicking task utilized a slimmer, less heavy USB mouse, whereas the XR clicking task employed a larger, heavier Bluetooth mouse. The difference in size, weight, and perhaps even responsiveness between these two mice could have affected the ease and speed with which participants were able to complete the task, potentially confounding the results.

Secondly, the fidelity of the photogrammetry used to create the mesh of the desk was not always well received by the participants. Informal feedback from two participants indicated that they found the mesh unconvincing or even more distracting than the actual visual elements that were supposed to be the main focus of the task. This suggests that the quality of the virtual environment and the believability of the diminished reality techniques used could significantly impact the user experience and task performance.

Further, the use of the CAAV database as the source of visual elements might not accurately represent the types of visual elements encountered in an average person's day-to-day life. While the CAAV database provides a wide range of standardized stimuli, it may not fully capture the diversity and complexity of real-world visual environments. This could limit the generalisability of the findings to real-world settings.

Moreover, the results presented here are likely limited to the Meta Quest 3 HMD and the See-Through DR implementation. Given the rapid advancements in HMD technology, it is possible that these results may change in the future due to factors such as enhanced visual fidelity for VST, improved real-world blending, changes in overall design and weight, and varying screen resolution, which could influence the outcomes. Furthermore, once a general DR implementation framework is developed, it could serve as a common ground for all future work on the subject, with the potential to reach near-indistinguishable levels of DR illusion. At that point, our results are likely to be surpassed. Furthermore, our study exclusively focused on visual elements, excluding other potential distractions such as auditory, olfactory and haptic stimuli, which also influence user perception [55]. Additionally, we did not consider visual elements that were not confined to a screen, such as an object, virtual or physical, entering the user's field of view in 3D, such as someone moving their arm to reach out to them. Finally, the nature of the task only had two difficulty settings, without any gradation from one to the other, i.e. moving from a relatively simple task that did not necessitate high concentration to one that required visual search and was completed in a considerably longer time. It is probable that other tasks of varying degrees of difficulty, potentially ones that progress in difficulty gradually, would yield different results.

4

Conclusion and Future Work

4.1 Conclusion

This thesis has explored the intricate dynamics of visual elements within see-through systems, highlighting their significant impact on user experience and interaction. Through comprehensive literature review, experimentation, and user studies, we have attempted to study whether performance and mental load were affected by adding or removing visual elements, and if DR-based removal was a valid approach, analogous to physical removal. The findings underscore the importance of thoughtful design in implementing DR applications, even more so if they are being designed as stand-ins or replacements for reality, as there are many factors that contribute to believability, and special care must be put in not accidentally affecting the user's performance or task load.
We have found that physical environments are still more usable than XR ones, even if they are simply replicating reality, and that, although DR is not perfect, and currently does not represent a visual element removal mechanism comparable to physical elimination due to affecting various objective and subjective metrics differently from it, present-day capabilities highlight its potential future developments that could bring it closer to a perfect illusion. As we contemplate the future, there remains ample opportunity for further research to refine these insights, explore new technologies, and develop innovative applications that leverage the principles of DR. By continuing to investigate the interplay between visual design and user interaction, we can contribute to the evolution of immersive technologies and their applications in various fields, which is still to this day a novel research direction.

4.2 Future Work

The findings from this research open up several avenues for future work. We have found evidence of visual elements having an effect on the user, so one interesting direction could be to investigate whether additional visual elements affect physical tasks differently from XR tasks. This could involve designing a study where participants perform similar tasks in both physical and XR environments, with and without additional visual elements. Furthermore, the impact of these elements on different metrics or other,

less simple, kinds of tasks could be analyzed to understand their impact in different contexts and under different points of view, and provide a more complete picture of their effect, in conjunction with the present work. Different natures of visual elements could be shown to have different effects, such as distractive ones, meant to capture the participants' attention and direct it away from their main task. For instance, the mobile phone notifications mentioned in Section 1 are designed to do so. Another avenue for further research could be alternative visualisations of the data in the user's main task, which could help to offer additional information in a way that is always available to the user if needed, effectively acting as an additional screen, which could aid the task but nevertheless reduce its mental load. By removing visual elements with different objectives, the impact is likely to be different.

Moreover, our study was limited to the See-Through implementation of DR (see Section 1.1.2), but others might be viable options to accomplish different approaches towards hiding visual elements. For instance, computer vision could be used to identify the outline of the visual element to be diminished, and possibly apply an in-painting technique to diminish it in real-time without need for setup from the experimenter, possibly yielding a more convincing illusion and/or different results.

Another promising area for future research could be to explore the potential benefits of this approach for individuals with Autism Spectrum Disorder (ASD) or Attention Deficit Hyperactivity Disorder (ADHD). Both conditions are characterised by difficulties with attention regulation and sensory processing, and thus the ability to control and manipulate visual elements in an environment, as demonstrated in this study, could potentially be used as a therapeutic tool to help individuals with ASD or ADHD better manage their attention and sensory experiences. This could involve developing tailored XR environments and tasks that take into account the specific needs and challenges of these individuals.

These potential research directions highlight the versatility and potential of XR and DR technologies, and the value of further exploring their applications in different contexts and populations. As technology continues to advance, the possibilities for future work in this area are vast.

Bibliography

[1] T. S. Foulger, "The 21st-century teacher educator and crowdsourcing," *Journal of Digital Learning in Teacher Education*, vol. 30, pp. 110–110, January 2014.

[2] S.-K. Kim, S.-Y. Kim, and H.-B. Kang, "An analysis of the effects of smartphone push notifications on task performance with regard to smartphone overuse using erp," *Computational Intelligence and Neuroscience*, vol. 2016, pp. 1–8, 2016.

[3] S. Ohly and L. Bastin, "Effects of task interruptions caused by notifications from communication applications on strain and performance," *Journal of Occupational Health*, vol. 65, January 2023.

[4] D. Heaney, "Quest 3 depth api released for mixed reality dynamic occlusion," October 2023. Available at `https://www.uploadvr.com/quest-3-depth-api-mixed-reality-dynamic-occlusion/` [Online; accessed 26-July-2024].

[5] C. Gsaxner, S. Mori, D. Schmalstieg, J. Egger, G. Paar, W. Bailer, and D. Kalkofen, "Deepdr: Deep structure-aware rgb-d inpainting for diminished reality," in *2024 International Conference on 3D Vision (3DV)*, pp. 750–760, Institute of Electrical and Electronics Engineers (IEEE), March 2024.

[6] Y. Qin, J. Su, H. Qin, and Y. Tian, "Exploring the role of video playback visual cues in object retrieval tasks," *Sensors*, vol. 24, p. 3147, May 2024.

[7] Y. Du, X. Huang, and D. El-Zanfaly, "Subtle visual cues in mixed reality: Influencing user perception and facilitating interaction," in *Creativity and Cognition*, pp. 556–560, ACM, June 2024.

[8] F. Mallek, T. Mazhar, S. F. A. Shah, Y. Y. Ghadi, and H. Hamam, "A review on cultivating effective learning: synthesizing educational theories and virtual reality for enhanced educational experiences," *PeerJ Computer Science*, vol. 10, p. e2000, May 2024.

[9] O. K. Xanthidou, N. Aburumman, and H. Ben-Abdallah, "Investigating trainee perspectives on virtual reality environments: An in-depth examination of immersive experiences with haptic feedback vibration," in *2024 IEEE International Systems Conference (SysCon)*, pp. 1–8, Institute of Electrical and Electronics Engineers (IEEE), April 2024.

[10] H. Mittal, *Virtual Reality Applications in Healthcare*, pp. 50–62. CRC Press, August 2023.

[11] T. Zigart, G. Kormann-Hainzl, H. Lovasz-Bukvova, M. Hölzl, T. Moser, and S. Schlund, "From lab to industry: lessons learned from the evaluation of augmented and virtual reality use cases in the austrian manufacturing industry," *Production & Manufacturing Research*, vol. 11, January 2023.

[12] G. Marsicano, L. Casartelli, A. Federici, S. Bertoni, L. Vignali, M. Molteni, A. Facoetti, and L. Ronconi, "Prolonged neural encoding of visual information in autism," *Autism Research*, vol. 17, pp. 37–54, January 2024.

[13] P. K. Panda, A. Ramachandran, V. Kumar, and I. K. Sharawat, "Sensory processing abilities and their impact on disease severity in children with attention-deficit hyperactivity disorder," *Journal of Neurosciences in Rural Practice*, vol. 14, p. 509, August 2023.

[14] P. B. DeGuzman, S. Abooali, H. Sadatsafavi, G. Bohac, and M. Sochor, "Back to basics: Practical strategies to reduce sensory overstimulation in the emergency department identified by adults and caregivers of children with autism spectrum disorder," *International Emergency Nursing*, vol. 72, p. 101384, February 2024.

[15] J. J. S. Kooij and D. Bijlenga, "High prevalence of self-reported photophobia in adult adhd," *Frontiers in Neurology*, vol. 5, January 2014.

[16] S. Mann and J. Fung, "Videoorbits on eye tap devices for deliberately diminished reality or altering the visual perception of rigid planar patches of a real world scene," in *International Symposium on Mixed Reality (ISMR2001)*, March 2001.

[17] S. Mori, S. Ikeda, and H. Saito, "A survey of diminished reality: Techniques for visually concealing, eliminating, and seeing through real objects," *IPSJ Transactions on Computer Vision and Applications*, vol. 9, p. 17, December 2017.

[18] S. Mann, "'wearcam' (the wearable camera): personal imaging systems for long-term use in wearable tetherless computer-mediated reality and personal photo/videographic memory prosthesis," in *Digest of Papers. Second International Symposium on Wearable Computers (Cat. No.98EX215)*, pp. 124–131, Institute of Electrical and Electronics Engineers (IEEE), 2001.

[19] N. Kawai, T. Sato, and N. Yokoya, "Diminished reality based on image inpainting considering background geometry," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, pp. 1236–1247, March 2016.

[20] M. Kari, T. Grosse-Puppendahl, L. F. Coelho, A. R. Fender, D. Bethge, R. Schutte, and C. Holz, "Transformr: Pose-aware object substitution for composing alternate mixed realities," in *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 69–79, Institute of Electrical and Electronics Engineers (IEEE), October 2021.

[21] V. Drescher, "Inpainting: Removing distracting objects in hi res images | img.ly," May 2023. Available at `https://img.ly/blog/image-inpainting/` [Online; accessed 26-July-2024].

[22] Y. F. Cheng, H. Yin, Y. Yan, J. Gugenheimer, and D. Lindlbauer, "Towards understanding diminished reality," in *CHI Conference on Human Factors in Computing Systems*, pp. 1–16, ACM, April 2022.

[23] C. Li, S. Yeom, J. Dermoudy, and K. de Salas, "Cognitive load measurement in the impact of vr intervention in learning," in *2022 International Conference on Advanced Learning Technologies (ICALT)*, pp. 325–329, Institute of Electrical and Electronics Engineers (IEEE), July 2022.

[24] C. Groening and C. Binnewies, "The more, the merrier? - how adding and removing game design elements impact motivation and performance in a gamification environment," *International Journal of Human–Computer Interaction*, vol. 37, pp. 1130–1150, July 2021.

[25] J. Tsurukawa, M. Al-Sada, and T. Nakajima, "Filtering visual information for reducing visual cognitive load," in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers - UbiComp '15*, pp. 33–36, ACM, 2015.

[26] G. A. M. Vasiljevic and L. C. de Miranda, "The influence of graphical elements on user's attention and control on a neurofeedback-based game," *Entertainment Computing*, vol. 29, pp. 10–19, March 2019.

[27] B. Johnson, "Visual distraction effects on deliberate focus work," *Haworth*, 2017.

[28] E. Redlinger, B. Glas, and Y. Rong, "Impact of visual game-like features on cognitive performance in a virtual reality working memory task: Within-subjects experiment," *JMIR Serious Games*, vol. 10, p. e35295, April 2022.

[29] K. Kim and I. H. Jo, "The effect of visual cues on cognitive load depending on self-regulation in video-based learning," in *Proceedings of the 13th International Conference on Educational Data Mining, EDM 2020* (A. N. Rafferty, J. Whitehill, C. Romero, and V. Cavalli-Sforza, eds.), pp. 776–780, International Educational Data Mining Society, 2020.

[30] M. Workman, *Cognitive Load Research and Semantic Apprehension of Graphical Linguistics*, vol. 4799, pp. 375–388. Springer, 2007.

[31] K. Yu, I. Prasad, H. Mir, N. Thakor, and H. Al-Nashash, "Cognitive workload modulation through degraded visual stimuli: a single-trial eeg study," *Journal of Neural Engineering*, vol. 12, p. 046020, August 2015.

[32] A. Hoesl, M. Alic, and A. Butz, *On the Effects of Progressive Reduction as Adaptation Strategy for a Camera-Based Cinematographic User Interface*, vol. 10513, pp. 513–522. Springer, 2017.

[33] I. Dahlstrom-Hakki, Z. Alstad, J. Asbell-Clarke, and T. Edwards, "The impact of visual and auditory distractions on the performance of neurodiverse students in virtual reality (vr) environments," *Virtual Reality*, vol. 28, p. 29, March 2024.

[34] Y. Deng, Y. Wang, M. She, Y. Zhang, and Z. Li, *A Study on Visual Workload Components: Effects of Component Combination and Scenario Complexity on Mental Workload in Maritime Operation Tasks*, vol. 11571 LNAI, pp. 20–28. Springer, 2019.

[35] C. Chung, A. Kadan, Y. Yang, A. Matsuoka, J. Rubin, and M. Chechik, "The impact of visual load on performance in a human-computation game," in *Proceedings of the 12th International Conference on the Foundations of Digital Games*, vol. Part F130151, pp. 1–4, ACM, August 2017.

[36] M. Kari, R. Schütte, and R. Sodhi, "Scene responsiveness for visuotactile illusions in mixed reality," in *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–15, ACM, October 2023.

[37] J. Lee and L. H. Kim, "Augmenting reality to diminish distractions for cognitive enhancement," *arXiv*, March 2024.

[38] R. W. Engle, "Working memory capacity as executive attention," *Current Directions in Psychological Science*, vol. 11, pp. 19–23, February 2002.

[39] J. C. Raven and J. H. Court, *Raven's progressive matrices and vocabulary scales*. Oxford Psychologists Press Oxford, 1998.

[40] A. D. Crosta, P. L. Malva, C. Manna, A. Marin, R. Palumbo, M. C. Verrocchio, M. Cortini, N. Mammarella, and A. D. Domenico, "The chieti affective action videos database, a resource for the study of emotions in psychology," *Scientific Data*, vol. 7, p. 32, January 2020.

[41] Z. Li, Y. Wang, J. Guo, L.-F. Cheong, and S. Z. Zhou, "Diminished reality using appearance and 3d geometry of internet photo collections," in *2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 11–19, Institute of Electrical and Electronics Engineers (IEEE), October 2013.

[42] H. Ke and H. Wang, "Poster: Real-time object substitution for mobile diminished reality with edge computing," in *Proceedings - 2023 IEEE/ACM Symposium on Edge Computing, SEC 2023*, pp. 279–281, Institute of Electrical and Electronics Engineers (IEEE), 2023.

[43] P. Li, L. Liu, C.-B. Schönlieb, and A. I. Aviles-Rivero, "Optimised propainter for video diminished reality inpainting," *arXiv*, June 2024.

[44] J. J. Gross and R. W. Levenson, "Emotion elicitation using films," *Cognition & Emotion*, vol. 9, pp. 87–108, January 1995.

[45] J. Rottenberg, R. Ray, and J. Gross, "Emotion elicitation using films in: Coan ja, allen jjb, editors. the handbook of emotion elicitation and assessment," 2007.

[46] K. C. Smith and R. A. Abrams, "Motion onset really does capture attention," *Attention, Perception, & Psychophysics*, vol. 80, pp. 1775–1784, October 2018.

[47] P. L. Malva, I. Ceccato, A. D. Crosta, A. Marin, M. Fasolo, R. Palumbo, N. Mammarella, R. Palumbo, and A. D. Domenico, "Updating the chieti affective action videos database with older adults," *Scientific Data*, vol. 8, p. 272, October 2021.

[48] J. A. Russell and L. F. Barrett, "Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant.," *Journal of Personality and Social Psychology*, vol. 76, pp. 805–819, 1999.

[49] R. Pfister, A. Kiesel, and J. Hoffmann, "Learning at any rate: action–effect learning for stimulus-based actions," *Psychological Research*, vol. 75, pp. 61–65, January 2011.

[50] R. S. Kennedy, N. E. Lane, K. S. Berbaum, and M. G. Lilienthal, "Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness," *The International Journal of Aviation Psychology*, vol. 3, pp. 203–220, July 1993.

[51] S. G. Hart, "Task load index (nasa-tlx) v 1.0 ames research center," 1986.

[52] J. Brooke, "Sus: a 'quick and dirty' usability scale," 1996.

[53] R. Agarwal and E. Karahanna, "Time flies when you're having fun: Cognitive absorption and beliefs about information technology usage," *MIS Quarterly*, vol. 24, p. 665, December 2000.

[54] I. S. MacKenzie, "Fitts' law as a research and design tool in human-computer interaction," *Human–Computer Interaction*, vol. 7, pp. 91–139, March 1992.

[55] R. A. Gougeh and T. H. Falk, "Multisensory immersive experiences: A pilot study on subjective and instrumental human influential factors assessment," in *2022 14th International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–6, Institute of Electrical and Electronics

Engineers (IEEE), September 2022.

A

Appendix

A.1 User Study Counterbalancing Flow Chart Diagram

The following diagram shows the course of a user study, and how counterbalancing was implemented. This corresponds to the description provided in Section 2.7.3, where three fair coins were tossed to non-deterministically decide the ordering of the conditions, yielding a total of 8 possible orderings. In the flow chart below, the terms "Physical" and "XR" symbolise the blocks of 2 physical (🖵) and 2 XR (🥽) conditions, respectively. Within each block, the ordering of conditions with (🏞) and without (🏞) visual elements could also vary.

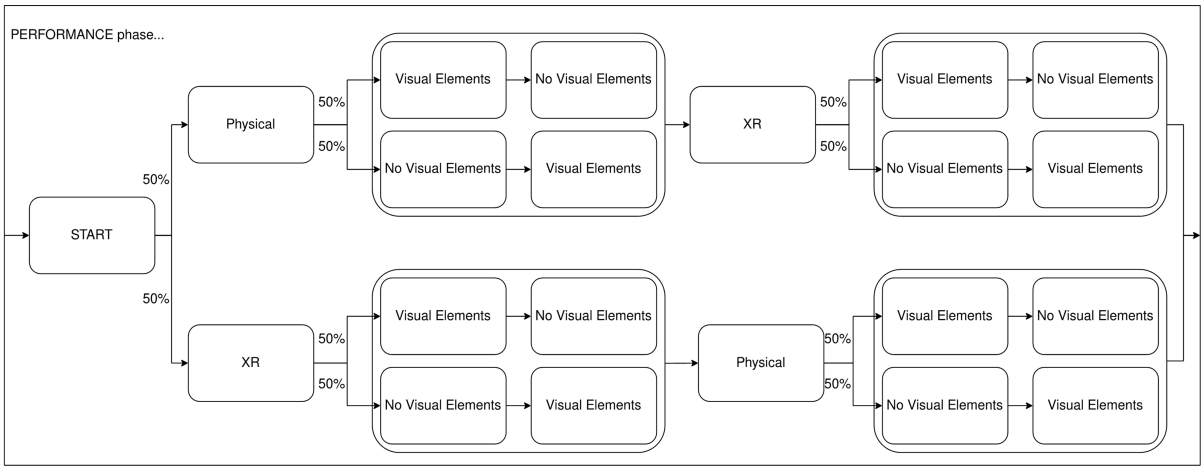


Figure A.1: *Flow chart of the counterbalancing procedure during the course of the participant's participation.*

A.2 Counterbalancing results

Participant ID	P-PD-XR-XRD	PD-P-XR-XRD	P-PD-XRD-XR	PD-P-XRD-XR	XR-XRD-P-PD	XR-XRD-PD-P	XRD-XR-P-PD	XRD-XR-PD-P
1				✓				
2	✓							
3				✓				
4			✓					
5						✓		
6								✓
7								✓
8		✓						
9								✓
10						✓		
11						✓		
12					✓			
13		✓						
14	✓							
15						✓		
16		✓						
17							✓	
18								✓
19							✓	
20			✓					
21								✓
22		✓						
23		✓						

Table A.1: *Table displaying the results of the counterbalancing, i.e. the orderings assigned to each participant. For the context of this table,* P *and* PD *refer to the physical conditions with and without visual elements, respectively, and* XR *and* XRD *refer to the XR conditions with and without visual elements, respectively. A column "A-B-C-D" represents the order of conditions, symbolising an ordering of conditions A, then B, then C, and finally D.*

A.3 Timeouts Distribution Across Conditions

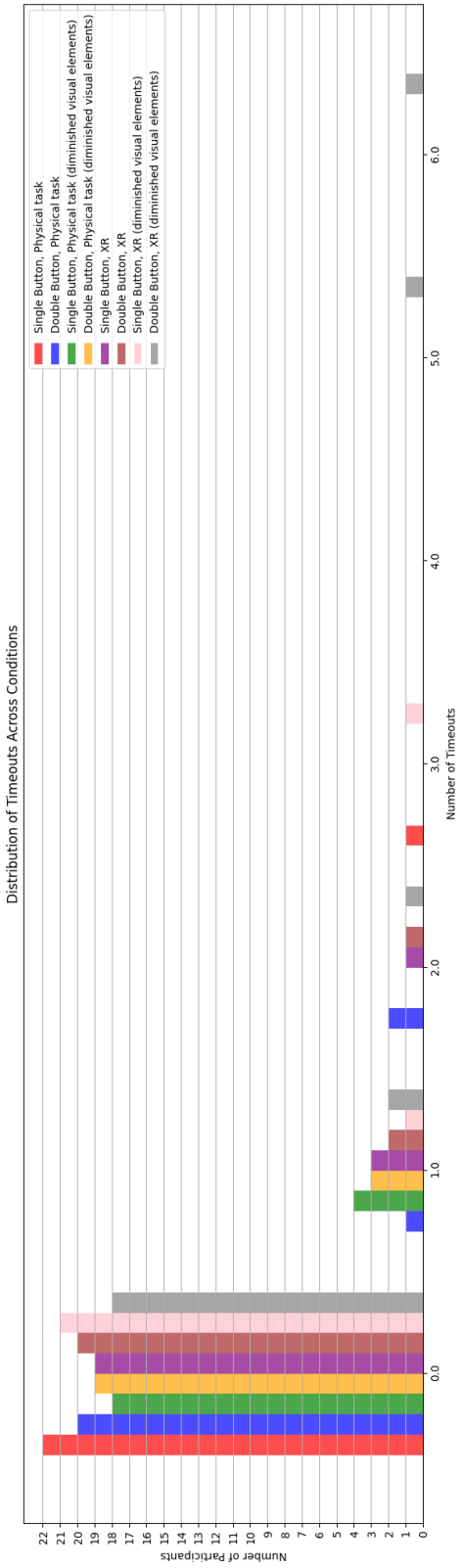


Figure A.2: *Bar graph showcasing the frequency of timeouts by condition.*

A.4 Misclicks Distribution Across Conditions

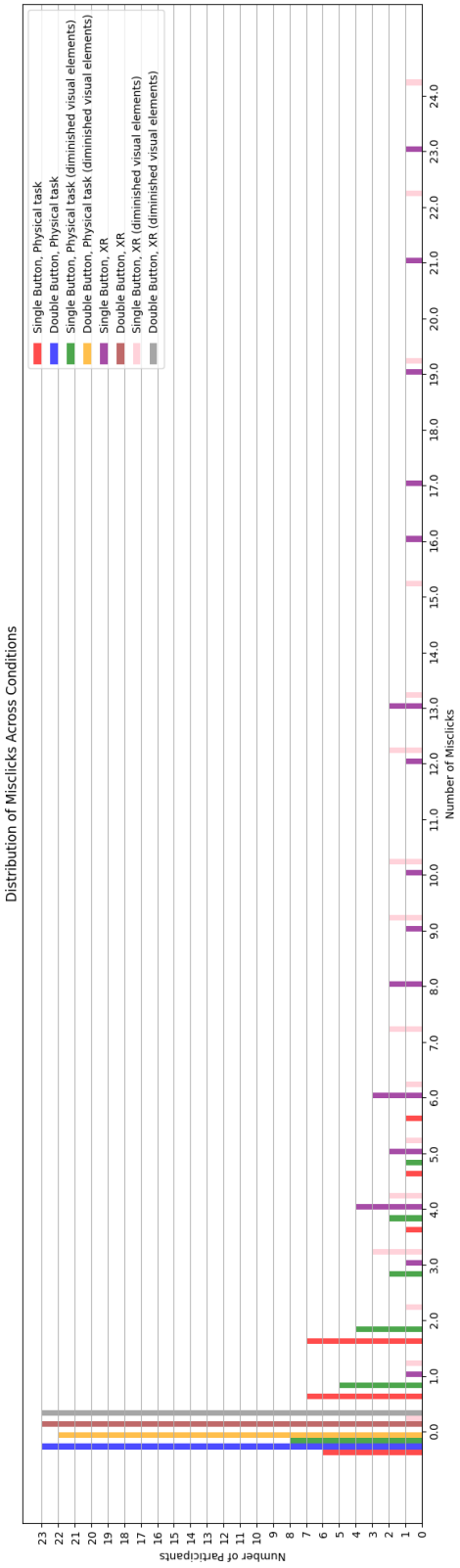


Figure A.3: *Bar graph showcasing the frequency of misclicks by condition.*

A.5 Misclicks Conover's Test Post-Hoc Differences Table

For the following table, "diff." stands for the median difference between the value under the condition represented by the row, and the value under the condition represented by the column. Bold text represents a difference that was significant with a significance level $\alpha = 0.05$.

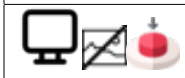
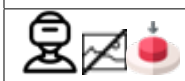
				
	1.00	$7.72 \cdot 10^{-1}$	$\mathbf{9.51 \cdot 10^{-10}}$ (diff. $= -7.91$)	$\mathbf{1.09 \cdot 10^{-8}}$ (diff. $= -7.32$)
	$7.72 \cdot 10^{-1}$	1.00	$\mathbf{5.07 \cdot 10^{-10}}$ (diff. $= -8.05$)	$\mathbf{4.01 \cdot 10^{-9}}$ (diff. $= -7.45$)
	$\mathbf{9.51 \cdot 10^{-10}}$ (diff. $= -7.91$)	$\mathbf{5.07 \cdot 10^{-10}}$ (diff. $= -8.05$)	1.00	$5.82 \cdot 10^{-1}$
	$\mathbf{1.09 \cdot 10^{-8}}$ (diff. $= -7.32$)	$\mathbf{4.01 \cdot 10^{-9}}$ (diff. $= -7.45$)	$5.82 \cdot 10^{-1}$	1.00

Table A.2: *Table illustrating the results of the post-hoc Conover's test, indicating the difference between experimental conditions in terms of misclicks.*

A.6 Fitts' Law Analysis

The index of difficulty was computed by first calculating the angle of approach between the mouse cursor and the target button using basic trigonometry:

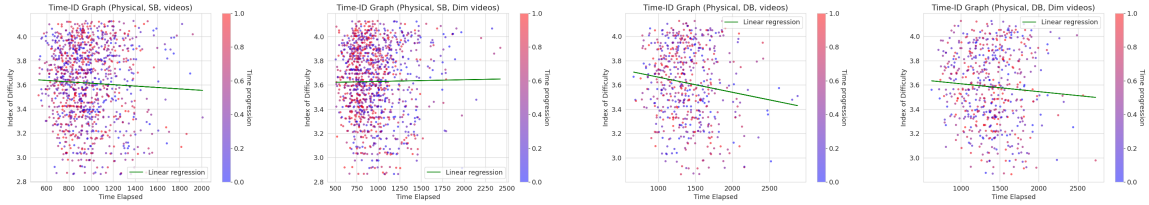
$$\theta = \arctan\left(\frac{y_{target} - y_{mouse}}{x_{target} - x_{mouse}}\right).$$

This angle was then used to determine whether the mouse was approaching the target vertically or horizontally by comparing it to the angle of the diagonal of the button, computed using the aspect ratio of the button $\text{AR} = \arctan\left(\frac{w}{h}\right)$ (the ratio of the target's width to its height), where w is the target's width and h is its height. Based on this information, the width w' of the target button was determined, which corresponded to either its height (h) or width (w) depending on the angle of approach, i.e., $w' = w$ if $\theta \leq \text{AR}$, and $w' = h$ if $\theta > \text{AR}$. The index of difficulty was then computed using the Shannon formulation of Fitts' Law [54]:

$$\text{ID} = \log_2\left(\frac{d}{w'} + 1\right),$$

where d is the distance to the target, which was computed as the Euclidean distance between the mouse cursor and the target button, i.e., $d = \sqrt{(x_{target} - x_{mouse})^2 + (y_{target} - y_{mouse})^2}$. Subsequently, the ID values were plotted against the time required to click on the target button, with each participant's data points colored according to their time progression since the beginning of the task, to visualise possible learning effects, and a linear regression line was also plotted, resulting in the following 8 plots. N.b., in the following plots' captions, "SB" indicates conditions where a single button would activate (), and

"DB" indicates conditions where two buttons would activate at once (🍒🍒). Furthermore, "Dim videos" is used to refer to conditions where the additional visual elements were diminished (🖼), whereas simply "Videos" refers to conditions where the additional visual elements were present (🖼).

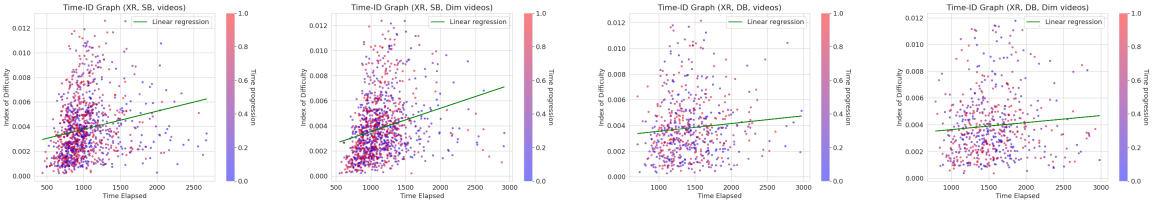


(a) *Time-ID plot for the clicks under the physical condition with single-button activation and additional visual elements. R^2 score for linear regression accuracy:* 0%

(b) *Time-ID plot for the clicks under the physical condition with single-button activation and diminished visual elements. R^2 score for linear regression accuracy:* 0%

(c) *Time-ID plot for the clicks under the physical condition with double-button activation and additional visual elements. R^2 score for linear regression accuracy:* 2%

(d) *Time-ID plot for the clicks under the physical condition with double-button activation and diminished visual elements. R^2 score for linear regression accuracy:* 0%



(e) *Time-ID plot for the clicks under the XR condition with single-button activation and additional visual elements. R^2 score for linear regression accuracy:* 4%

(f) *Time-ID plot for the clicks under the XR condition with single-button activation and diminished visual elements. R^2 score for linear regression accuracy:* 6%

(g) *Time-ID plot for the clicks under the XR condition with double-button activation and additional visual elements. R^2 score for linear regression accuracy:* 1%

(h) *Time-ID plot for the clicks under the XR condition with double-button activation and diminished visual elements. R^2 score for linear regression accuracy:* 1%

Figure A.4: *Time-ID plots for clicks recorded under the 8 experimental conditions*

A.7 Time Elapsed Conover's Test Post-Hoc Differences Table

For the following table, "diff." stands for the median difference between the value under the condition represented by the row, and the value under the condition represented by the column. Bold text represents a difference that was significant with a significance level $\alpha = 0.05$.

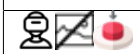
								
	1.00	$\mathbf{8.64 \cdot 10^{-170}}$ (diff. $= -378.23$)	$\mathbf{1.83 \cdot 10^{-2}}$ (diff. $= 17.58$)	$\mathbf{7.66 \cdot 10^{-165}}$ (diff. $= -372.32$)	$\mathbf{10.00 \cdot 10^{-4}}$ (diff. $= -52.50$)	$\mathbf{3.19 \cdot 10^{-187}}$ (diff. $= -423.52$)	$\mathbf{2.96 \cdot 10^{-54}}$ (diff. $= -175.88$)	$\mathbf{8.79 \cdot 10^{-248}}$ (diff. $= -541.75$)
	$\mathbf{8.64 \cdot 10^{-170}}$ (diff. $= -378.23$)	1.00	$\mathbf{6.23 \cdot 10^{-192}}$ (diff. $= 378.28$)	$6.85 \cdot 10^{-1}$	$\mathbf{8.19 \cdot 10^{-141}}$ (diff. $= 321.39$)	$1.92 \cdot 10^{-1}$	$\mathbf{2.55 \cdot 10^{-55}}$ (diff. $= 211.62$)	$\mathbf{9.47 \cdot 10^{-9}}$ (diff. $= -165.79$)
	$\mathbf{1.83 \cdot 10^{-2}}$ (diff. $= 17.58$)	$\mathbf{6.23 \cdot 10^{-192}}$ (diff. $= 378.28$)	1.00	$\mathbf{7.92 \cdot 10^{-187}}$ (diff. $= -372.22$)	$\mathbf{1.28 \cdot 10^{-8}}$ (diff. $= -70.02$)	$\mathbf{2.74 \cdot 10^{-210}}$ (diff. $= -423.42$)	$\mathbf{3.86 \cdot 10^{-71}}$ (diff. $= -193.47$)	$\mathbf{3.77 \cdot 10^{-273}}$ (diff. $= -543.11$)
	$\mathbf{7.66 \cdot 10^{-165}}$ (diff. $= -372.32$)	$6.85 \cdot 10^{-1}$	$\mathbf{7.92 \cdot 10^{-187}}$ (diff. $= -372.22$)	1.00	$\mathbf{3.17 \cdot 10^{-136}}$ (diff. $= 315.57$)	$9.00 \cdot 10^{-2}$	$\mathbf{2.40 \cdot 10^{-52}}$ (diff. $= 206.04$)	$\mathbf{8.23 \cdot 10^{-10}}$ (diff. $= -168.29$)
	$\mathbf{10.00 \cdot 10^{-4}}$ (diff. $= -52.50$)	$\mathbf{8.19 \cdot 10^{-141}}$ (diff. $= 321.39$)	$\mathbf{1.28 \cdot 10^{-8}}$ (diff. $= -70.02$)	$\mathbf{3.17 \cdot 10^{-136}}$ (diff. $= 315.57$)	1.00	$\mathbf{5.21 \cdot 10^{-157}}$ (diff. $= -366.77$)	$\mathbf{1.84 \cdot 10^{-34}}$ (diff. $= -123.64$)	$\mathbf{7.02 \cdot 10^{-214}}$ (diff. $= -487.96$)
	$\mathbf{3.19 \cdot 10^{-187}}$ (diff. $= -423.52$)	$1.92 \cdot 10^{-1}$	$\mathbf{2.74 \cdot 10^{-210}}$ (diff. $= -423.42$)	$9.00 \cdot 10^{-2}$	$\mathbf{5.21 \cdot 10^{-157}}$ (diff. $= -366.77$)	1.00	$\mathbf{5.24 \cdot 10^{-66}}$ (diff. $= 257.24$)	$\mathbf{9.22 \cdot 10^{-6}}$ (diff. $= -121.21$)
	$\mathbf{2.96 \cdot 10^{-54}}$ (diff. $= -175.88$)	$\mathbf{2.55 \cdot 10^{-55}}$ (diff. $= 211.62$)	$\mathbf{3.86 \cdot 10^{-71}}$ (diff. $= -193.47$)	$\mathbf{2.40 \cdot 10^{-52}}$ (diff. $= 206.04$)	$\mathbf{1.84 \cdot 10^{-34}}$ (diff. $= -123.64$)	$\mathbf{5.24 \cdot 10^{-66}}$ (diff. $= 257.24$)	1.00	$\mathbf{2.44 \cdot 10^{-107}}$ (diff. $= -379.84$)
	$\mathbf{8.79 \cdot 10^{-248}}$ (diff. $= -541.75$)	$\mathbf{9.47 \cdot 10^{-9}}$ (diff. $= -165.79$)	$\mathbf{3.77 \cdot 10^{-273}}$ (diff. $= -543.11$)	$\mathbf{8.23 \cdot 10^{-10}}$ (diff. $= -168.29$)	$\mathbf{7.02 \cdot 10^{-214}}$ (diff. $= -487.96$)	$\mathbf{9.22 \cdot 10^{-6}}$ (diff. $= -121.21$)	$\mathbf{2.44 \cdot 10^{-107}}$ (diff. $= -379.84$)	1.00

Table A.3: *Table illustrating the results of the post-hoc Conover's test, indicating the difference between experimental conditions in terms of time elapsed to click on a target.*

A.8 Distance to Target Conover's Test Post-Hoc Differences Table

For the following table, "diff." stands for the median difference between the value under the condition represented by the row, and the value under the condition represented by the column. Bold text represents a difference that was significant with a significance level $\alpha = 0.05$.

								
	1.00	$9.51 \cdot 10^{-1}$	$4.93 \cdot 10^{-1}$	$6.07 \cdot 10^{-2}$	**0.00** (diff. = 2214.71)	**0.00** (diff. = 2219.69)	**0.00** (diff. = 2215.17)	**0.00** (diff. = 2220.86)
	$9.51 \cdot 10^{-1}$	1.00	$5.14 \cdot 10^{-1}$	$1.23 \cdot 10^{-1}$	**0.00** (diff. = 2214.41)	**0.00** (diff. = 2214.41)	**0.00** (diff. = 2214.41)	**0.00** (diff. = 2216.37)
	$4.93 \cdot 10^{-1}$	$5.14 \cdot 10^{-1}$	1.00	$\mathbf{8.81 \cdot 10^{-3}}$ (diff. = 63.72)	**0.00** (diff. = 2233.01)	**0.00** (diff. = 2227.54)	**0.00** (diff. = 2233.68)	**0.00** (diff. = 2228.82)
	$6.07 \cdot 10^{-2}$	$1.23 \cdot 10^{-1}$	$\mathbf{8.81 \cdot 10^{-3}}$ (diff. = 63.72)	1.00	**0.00** (diff. = 2163.81)	**0.00** (diff. = 2163.82)	**0.00** (diff. = 2163.81)	**0.00** (diff. = 2165.42)
	**0.00** (diff. = 2214.71)	**0.00** (diff. = 2214.41)	**0.00** (diff. = 2233.01)	**0.00** (diff. = 2163.81)	1.00	$7.08 \cdot 10^{-1}$	$9.42 \cdot 10^{-1}$	$4.93 \cdot 10^{-1}$
	**0.00** (diff. = 2219.69)	**0.00** (diff. = 2214.41)	**0.00** (diff. = 2227.54)	**0.00** (diff. = 2163.82)	$7.08 \cdot 10^{-1}$	1.00	$6.62 \cdot 10^{-1}$	$3.61 \cdot 10^{-1}$
	**0.00** (diff. = 2215.17)	**0.00** (diff. = 2214.41)	**0.00** (diff. = 2233.68)	**0.00** (diff. = 2163.81)	$9.42 \cdot 10^{-1}$	$6.62 \cdot 10^{-1}$	1.00	$5.14 \cdot 10^{-1}$
	**0.00** (diff. = 2220.86)	**0.00** (diff. = 2216.37)	**0.00** (diff. = 2228.82)	**0.00** (diff. = 2165.42)	$4.93 \cdot 10^{-1}$	$3.61 \cdot 10^{-1}$	$5.14 \cdot 10^{-1}$	1.00

Table A.4: *Table illustrating the results of the post-hoc Conover's test, indicating the difference between experimental conditions in terms of distace from the mouse pointer to the active target at the moment of its activation.*

A.9 Questionnaires Scores

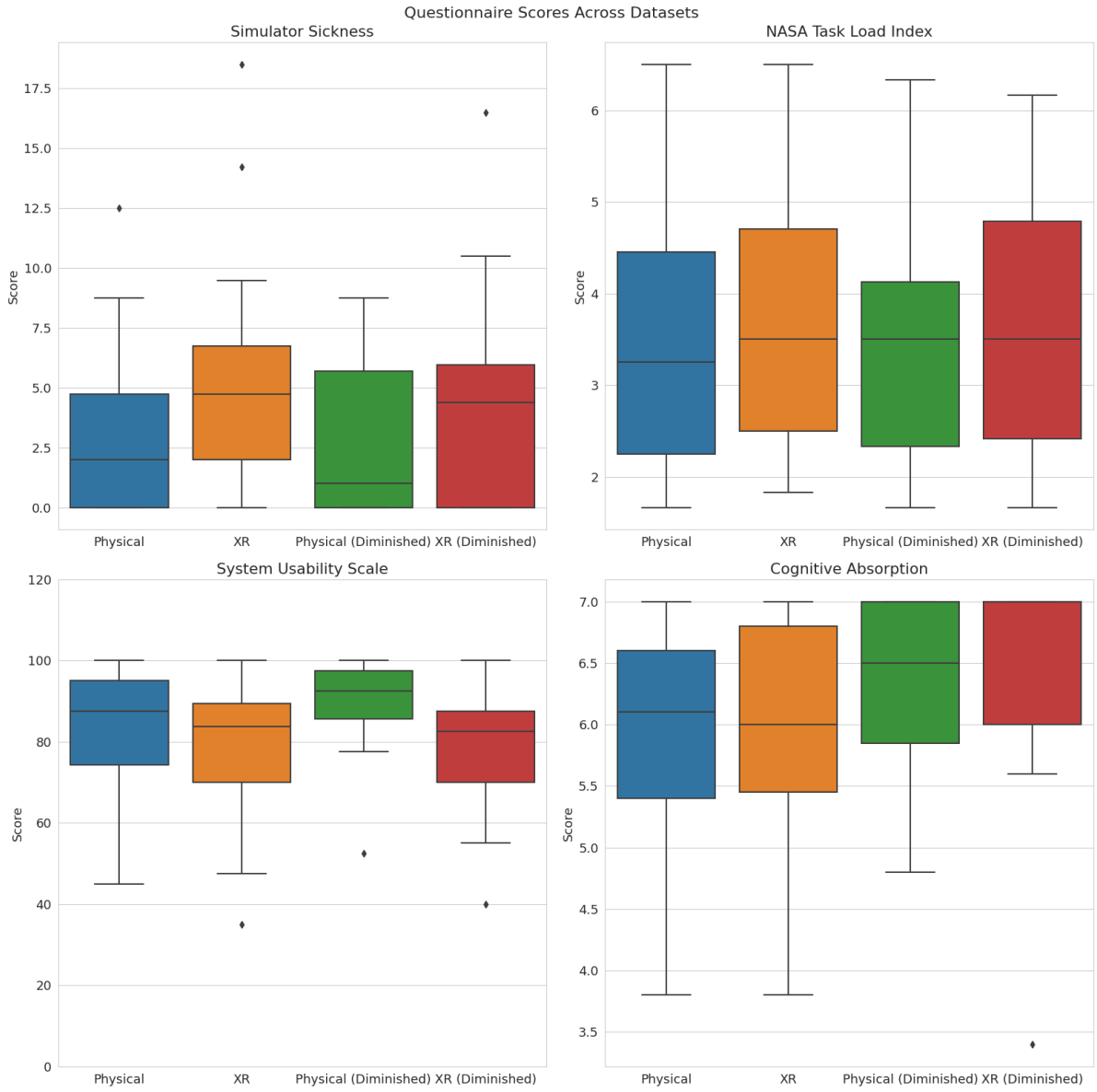
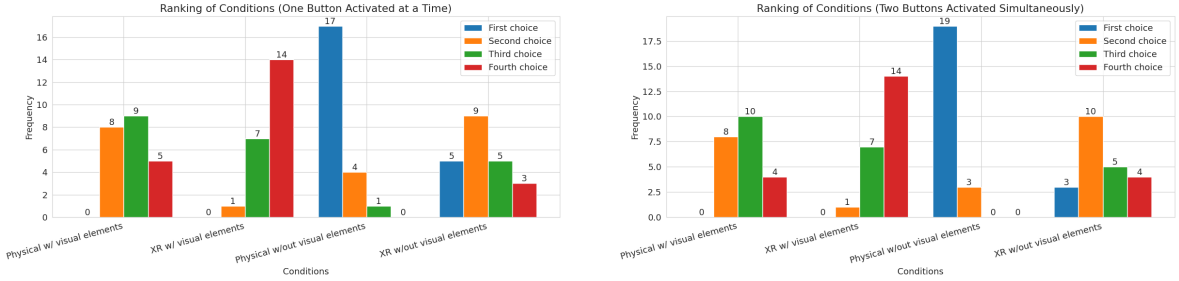


Figure A.5: *Bar graph showcasing the frequency of timeouts by condition.*

A.10 Conditions Ranking - Full Data



(a) Participants' self-reported ranking in preference for under which condition to perform the clicking task with 1-button activation.

(b) Participants' self-reported ranking in preference for under which condition to perform the clicking task with 2-button activation.

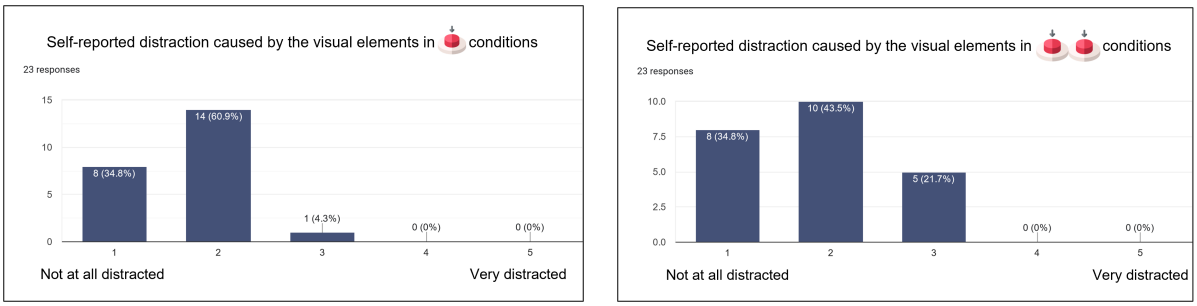
Figure A.6: *Bar graphs illustrating the participants' self-reported preference for the conditions under which to perform the clicking task.*

A.11 Inverse Weighted Sum

To calculate the inverse weighted sum score s_c for one of the N conditions $c \in C$ (for this user study, $N = 8$), where C is the set that contains our 4 basic study conditions, and n_c^i is the amount of users that have expressed condition c as being their i-th favourite out of all conditions.

$$s_c = \sum_{i=1}^{N} n_c^i \cdot (n - i)$$

A.12 Perceived Distraction



(a) Participants' self-reported level of distraction perceived from the CAAV clips while they were performing the task with 1-button activation.

(b) Participants' self-reported level of distraction perceived from the CAAV clips while they were performing the task with 2-button activation.

Figure A.7: *Bar graphs illustrating the frequency of participants' report for each level of distraction.*