UNIVERSITY OF OKLAHOMA
GRADUATE COLLEGE

# TOWARDS SOUND-INTEGRATED DIGITAL TWIN FOR HAZARD DETECTION AND RESPONSE

A THESIS
SUBMITTED TO THE GRADUATE FACULTY
in partial fulfillment of the requirements for the
Degree of
MASTER OF SCIENCE

By
ALAYNA MENDOZA
Norman, Oklahoma 2026

TOWARDS SOUND-INTEGRATED DIGITAL TWIN FOR HAZARD DETECTION AND RESPONSE

A THESIS APPROVED FOR THE

SCHOOL OF INDUSTRIAL AND SYSTEMS ENGINEERING

BY THE COMMITTEE CONSISTING OF


Dr. Ziho Kang, Chair

Dr. Soheil Hemmati

Dr. Byeong-Min Roh

# Acknowledgements

First, I would like to thank Dr. Ziho Kang, the chair of my thesis committee, for his mentorship and expertise throughout this entire research. His classes regarding human factors and research in eye tracking really sparked my interest in working with him. I also would like to thank Dr. Soheil Hemmati and Dr. Byeong-Min Roh for their willingness to join my committee. All contributions are greatly appreciated as I continue to grow into a skilled industrial and systems engineer.

Lastly, I would like to thank the School of ISE here at the University of Oklahoma. All ISE faculty has shaped me in ways that will last a lifetime but especially in professional and personal growth.

# Table of Contents

**Acknowledgements** ........ IV
**Table of Contents** ........ V
**List of Tables** ........ VI
**List of Figures** ........ VII
**Abstract** ........ VIII
**1. Introduction** ........ 1
**2. Methods** ........ 4
**2.1. Participants** ........ 4
**2.2. Apparatus** ........ 5
**2.3. Scenarios** ........ 5
**2.4. Task and Procedures** ........ 8
**2.5. Variables** ........ 8
**2.6. Data Analysis** ........ 9
**2.7. Self-report Data Analysis** ........ 11
**2.7.1. NASA-TLX Forms** ........ 11
**2.7.2. User-feedback Surveys** ........ 11
**3. Results** ........ 12
**4. Self-report Results** ........ 19
**4.1. NASA-TLX Forms** ........ 19
**4.2. User-feedback Surveys** ........ 20
**4.2.1. Perceived Benefit of Sound for Task Completion** ........ 21
**4.2.2. Helpfulness of Triggered Sounds** ........ 21
**4.2.3. Improvements to the VR app** ........ 22
**5. Discussion** ........ 24
**6. Limitations and Future Research** ........ 27
**7. References** ........ 29

# List of Tables

**Table 1.** Audio condition Mann-Whitney test results on task completion time.......................... 17
**Table 2.** Difficulty level Kruskal-Wallis test results on task completion time. .......................... 17
**Table 3.** Difficulty level Post-hoc analyses using Mann-Whitney test results on task completion time. .......................................................................................................................... 17
**Table 4.** Audio condition Linear Regression test results on task completion time..................... 17
**Table 5.** Audio condition Mann-Whitney test results on eye fixation duration.......................... 17
**Table 6.** Difficulty level Kruskal-Wallis test results on eye fixation duration. .......................... 17
**Table 7.** Difficulty level Post-hoc analyses using Mann-Whitney test results on eye fixation duration. ..................................................................................................................... 18
**Table 8.** Audio condition Linear Regression test results on eye fixation duration. .................... 18
**Table 9.** Audio condition Mann-Whitney test results on haptic interaction time. ....................... 18
**Table 10.** Difficulty level Kruskal-Wallis test results on haptic interaction time....................... 18
**Table 11.** Difficulty level Post-hoc analyses using Mann-Whitney test results on haptic interaction time. ........................................................................................................... 18
**Table 12.** Audio condition Linear Regression test results on haptic interaction time. ................ 18
**Table 13.** Audio condition Mann-Whitney test results on NASA-TLX workload dimensions and overall NASA-TLX values. .................................................................................... 20
**Table 14.** Seven perceived benefits of sound for task completion with example quotations. ..... 21
**Table 15.** Six perceived helpfulness of triggered sounds with example quotations. .................. 22
**Table 16.** Five improvements to the VR app with example quotations. ..................................... 23

# List of Figures

**Figure 1.** Examples of the type of objects used for each category. (a-c) Food objects. (d-f) Kitchen objects. (g-i) Hazardous objects. ........ 6

**Figure 2.** The VR environment at an initial stage for all three scenarios. ........ 7

**Figure 3.** A participant interacting in the VR environment relative to completing the task for level two. ........ 7

**Figure 4.** Average task completion times. ........ 13

**Figure 5.** Average eye fixation durations on AOIs. ........ 13

**Figure 6.** Average haptic interaction times with AOIs. ........ 13

**Figure 7.** Scenario specific average eye fixation durations on AOIs. (a) L1 average. (b) L2 average. (c) L3 average. ........ 15

**Figure 8.** Scenario specific average haptic interaction times with AOIs. (a) L1 average. (b) L2 average. (c) L3 average. ........ 16

**Figure 9.** Average rated points for workload dimensions and overall TLX values. ........ 20

# Abstract

Auditory feedback is commonly used in virtual reality (VR) to support user performance, particularly when visual attention is highly engaged. However, most VR applications rely on event-based or spatially triggered sounds, leaving limited understanding of how eye fixations can trigger auditory feedback. This study investigates the effects of eye fixation triggered sounds on task performance in a VR categorization task. Participants classified virtual objects while auditory feedback was triggered by eye fixations. Task performance was evaluated using task completion time as well as eye fixation duration and haptic interaction time. Results showed that eye fixation triggered auditory feedback significantly reduced task completion time, eye fixation duration, and haptic interaction time compared to the no-audio condition. These findings demonstrate that eye fixation triggered sound can serve as an effective mechanism for supporting performance in visually demanding VR tasks and highlights its potential for future training and hazard-related applications.



**Keywords:** eye fixation, auditory feedback, virtual reality, cognitive load, detection and response performance


Note: The contents of this thesis have been submitted as a journal paper.

# 1. Introduction

The use of eye fixation triggered sounds have been explored across a range of applications beyond VR, particularly in social interaction, developmental and learning disorders, and clinical or assistive contexts involving visual impairments. In social attention research, eye tracking has been used to examine how children and adults follow another person's gaze in situations involving compound audiovisual stimuli, where spoken information was either congruent or incongruent with a visual target; as social interactions involved more people, both groups showed reduced eye fixations on faces and targets, indicating increased difficulty maintaining gaze-based attention (Oka & Omori, 2023).

In the context of learning disorders, a narrative review of studies published between 1999 and 2025 highlighted that eye tracking and gaze-contingent interventions in school age children consistently revealed oculomotor signatures of inefficient visual sampling, often uncovering processing difficulties not detectable through accuracy or response time measures alone (Diotaiuti et al., 2026).

Gaze-contingent auditory feedback has also been used to guide visual attention when visual cues are unavailable or undesirable, with sonification-based guidance enabling users to exploit auditory feedback effectively and achieve significantly faster visual search performance compared to unguided conditions (Losing et al., 2014).

Similarly, gaze-contingent auditory sensory substitution systems mapping depth information to sound have been shown to support spatial perception in blind and visually impaired users, enabling depth perception and eliciting intermodal audiovisual processing (Twardon et al., 2013). Beyond perceptual guidance, continuous real-time sonification of eye movements has been explored as a means to improve oculomotor control and awareness, with auditory feedback shown to alter oculomotor behavior and enhance intrinsic perception of eye movements (Boyer et al., 2017).

While research on eye fixation triggered sounds in VR and other application domains has demonstrated benefits for understanding social interaction, perceptual and attentional control, and visual impairment, prior work has also shown that sound more broadly can enhance human task performance in VR. Immersive audio has been investigated as a means of compensating for reduced visual immersion when VR-like content is presented on non-VR displays, with results showing significant increases in viewers' concentration, emotional involvement, and gaze engagement, indicating that auditory enhancement can play a more critical role than visual enhancement in promoting flow and attentional engagement (Huang & Li, 2025). Gaze behavior has also been examined in relation to listener performance and cognitive effort during sound-localization tasks in VR, with eye-gaze patterns found to be strongly associated with task effort and cognitive load, highlighting gaze data as a valuable indicator of listener behavior during spatial-audio tasks (Moraes et al., 2022).

In addition, minimally disruptive, barely audible auditory cues have been shown to reliably degrade visual performance in VR by eliciting an audiovisual suppression effect, demonstrating that even subtle sounds can significantly influence visual perception (Jiménez-Navarro et al., 2024). Complementing these findings, studies on substitute multisensory feedback have shown that participants perform tasks faster and report higher levels of presence when auditory and tactile cues are provided, underscoring the role of sound in enhancing task performance and user experience despite reduced physical fidelity (Cooper et al., 2018). Finally, Soundspace VR demonstrated that spatialized auditory cues alone can effectively support navigation and spatial learning in the absence of visual information, with navigation performance influenced by task difficulty and global cognitive functioning (Fialho et al., 2021). Collectively, these studies highlight the substantial role of sound in improving task performance, attention, and spatial behavior in VR environments.

As discussed above, VR has been increasingly adopted across a range of research domains, demonstrating its value as an effective tool for training and learning. In particular, research combining VR and digital twin technologies for hazard detection, emergency response, and evacuation training has reported consistently positive outcomes. A systematic review of studies published between 2014 and 2024 highlighted the use of VR-based evacuation simulations in public buildings, transportation hubs, and high-risk workplaces, showing that VR enables realistic evacuation scenarios and improves preparedness, safety compliance, and evacuation performance (Hung et al., 2025). Complementing this work, VR experiments examining visual and auditory guidance cues in road-tunnel fire evacuations found that participants reliably followed guidance when a single cue was present, and that audio instructions and temporary cues were more influential than static visual signage when cues conflicted (Zhang et al., 2025).

Similarly, VR-based airport evacuation simulations demonstrated that multimodal safety alarms, particularly those combining auditory and visual cues, significantly influenced evacuation decision-making and reduced evacuation time compared to single-modality alarms (Zeng et al., 2025). Beyond civilian evacuation, VR-based situated digital twins integrating photogrammetry, VR, and augmented reality have been proposed to train first responders for rapidly evolving, high-risk incidents, with case studies showing scalable and adaptable training solutions that enhance situational awareness and decision-making (Partarakis et al., 2025).

VR-based evacuation training has also been explored in educational settings, where studies report strong student engagement and perceived learning benefits from immersive and gamified experiences, while educators and emergency professionals emphasize VR's role as a complement to, rather than a replacement for, traditional evacuation drills (Keya et al., 2025). Collectively, this body of work demonstrates that VR-based simulations, digital twins, and multimodal cueing strategies provide powerful and flexible tools for hazard related training.

Although prior work has demonstrated that auditory cues and gaze-contingent feedback can enhance attention, navigation, and perceptual stability in virtual environments, little is known about how eye fixation triggered auditory feedback can support high-level cognitive tasks such as categorization and decision-making. In particular, existing studies have not examined whether auditory cues mapped to task-relevant categories and triggered by eye fixation can facilitate detection accuracy and response performance. Closely related research has primarily focused on detection performance in VR-based visual search tasks, where auditory cues provide spatial guidance only, indicating where an object may be located rather than conveying information about what the object represents (Flanagan et al., 1998; Kim et al., 2008; Nelson et al., 1998; Van der Burg et al., 2008). Moreover, the effects of such auditory cues on response performance and decision timing have not been systematically explored in these studies.

Thus, this paper investigates how eye fixation triggered auditory cues influence user performance in a VR-based categorization task. I hypothesized that eye fixation triggered sounds mapped to task-relevant categories will significantly enhance user task performance.

# 2. Methods

## 2.1. Participants

A total of fifty-two participants at the University of Oklahoma were recruited for this study. Their ages ranged between 18 and 47 years old (*M* = 20, *S.D.* = 5.8). Inclusion criteria required normal vision and hearing, as well as sufficient cognitive ability to follow simple instructions.

Prior experience with virtual reality (VR) was not required. However, their experience was recorded on a scale from one to five where one indicated no experience and five indicated being very familiar with VR (*M* = 2.5, *S.D.* = 1.3). No additional exclusion criteria were applied.

## 2.2. Apparatus

The study utilized the HTC Vive Pro Eye virtual reality headset (which has a visual angle error of 0.4 degrees) and the left and right VR controllers. The VR application was developed using Python Script on the Vizard (accessible at www.worldviz.com) platform. Recorded data for each participant was saved separately as a text file. Participants' eye movement data, specifically raw eye gaze data that provides (x, y, z) coordinates at 100hz (or every 10 milliseconds), were collected using the HTC Vive Pro Eye device. In addition, the (x, y, z) coordinates of each VR controller (i.e. left controller and right controller) were also recorded at 100hz.

## 2.3. Scenarios

Three scenarios were utilized in this study. Scenarios were identified as difficulty levels one, two, and three consisting of six, nine, and twelve different objects, respectively. Within each scenario, there were three categories to place the associated objects in: (1) food (e.g., pizza, salad, ramen, etc.), (2) kitchen utensils (e.g., whisk, teapot, blender, etc.), and (3) hazardous objects (e.g., bleach, gasoline, Clorox wipes, etc.) (see **Figure 1**). Objects were evenly dispersed among the three categories for each scenario. Participants experienced scenarios either with the addition of sound or without the addition of sound, depending on which one was implemented last during experimenting. All objects in each scenario where participants experienced the addition of sound were given a trigger sound based on the three categories: (1) chewing, (2) cooking, and (3) a siren, respectively. Trigger sounds would engage when looking at an object.


**Figure 1.** Examples of the type of objects used for each category. (a-c) Food objects. (d-f) Kitchen objects. (g-i) Hazardous objects.

A table, kitchen counter, and black shelfing were displayed to place the different objects on as seen in **Figure 2**. The table was used for objects related to food, and the kitchen counter was used for objects related to kitchen utensils. Similarly, the black shelfing was used for objects related to hazard. Objects could be placed by holding the left or right triggers down on either VR controller when hovering over an object and then released once in the desired location. **Figure 3** depicts the process in which a participant places the different objects in their respective locations based on the category associated with it. Half of the participants (26) participated in all three scenarios with the addition of sound while the other half participated in all three scenarios without the addition of sound.

**Figure 2.** The VR environment at an initial stage for all three scenarios.

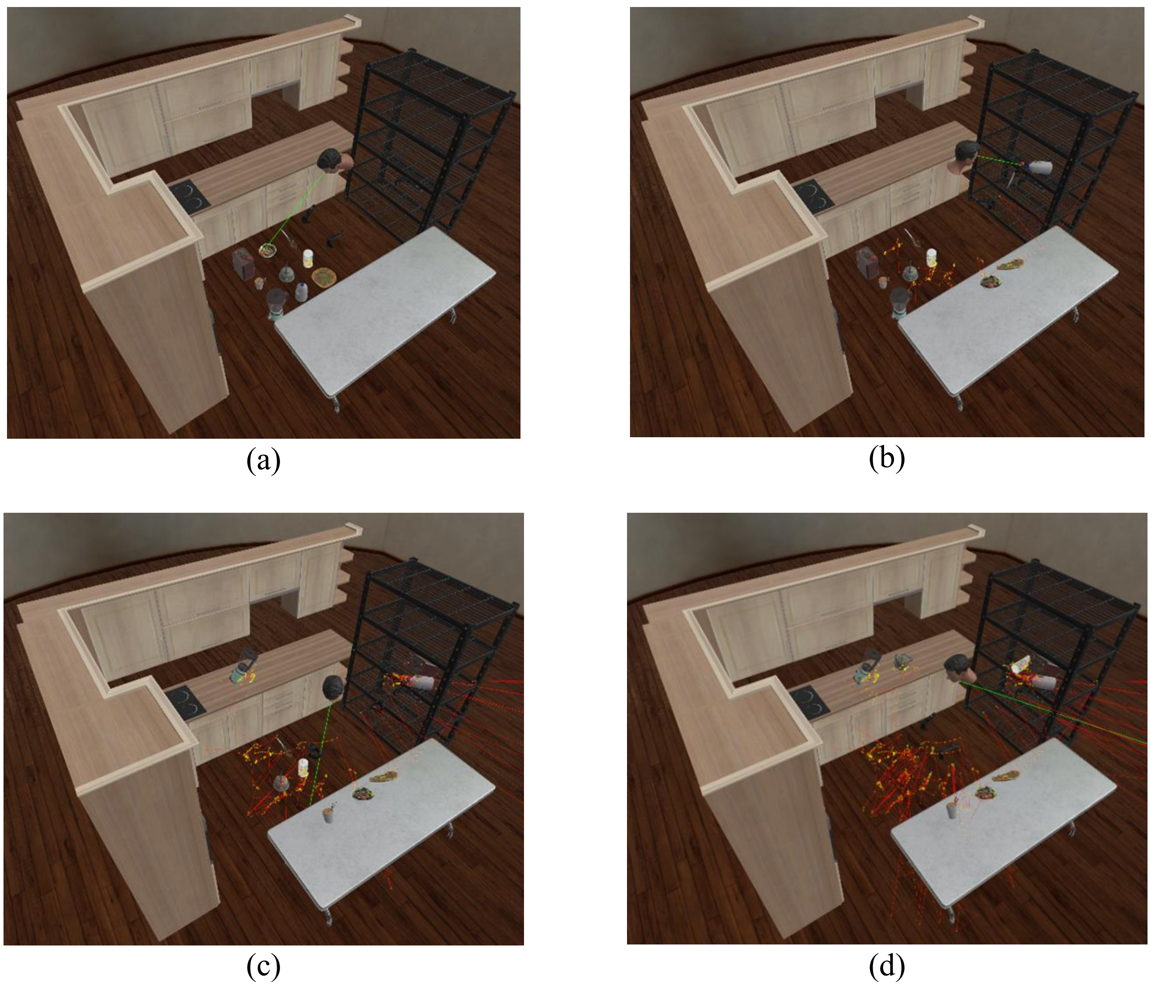



**Figure 3.** A participant interacting in the VR environment relative to completing the task for level two.

## 2.4. Task and Procedures

For each scenario, all participants were tasked with placing the different objects in their respective locations within the VR environment as quickly as possible. Participants were first informed of the task to be completed during experimenting, including how to use the left and right VR controllers once in the VR environment. Next, participants then calibrated the eye tracker inside the HTC Vive Pro Eye headset. Once calibration was complete, they entered the VR environment and proceeded with the task in a practice scenario. The practice scenario was modeled after the experimental scenarios for participants to complete as many times as needed until they felt comfortable interacting with the VR environment. Upon completing the task and data collection process for all three scenarios, participants completed a NASA Task Load Index (TLX) form and a user-feedback survey. Once both the form and survey were completed, participants were dismissed. All participants signed a consent form prior to beginning the experiment.

## 2.5. Variables

The independent variables in the study were audio condition and difficulty level. Audio condition consisted of two treatments: no sound and sound. That is, all scenarios were performed either with the addition of trigger sounds or without the addition of trigger sounds. Difficulty level consisted of three treatments: level one (six objects), level two (nine objects), and level three (twelve objects). Three dependent variables were measured in the study: (1) task completion time, (2) eye fixation duration on AOIs, and (3) haptic interaction time with AOIs. Task completion time involves the duration in which it took participants to complete the task of the experiment. Eye fixation duration is relative to how long participants were gazing at any of the objects. Haptic interaction time is the time it took participants to move each object to their respective locations. All dependent variables were measured in seconds.

The control variable in the study was the VR environment structure and conditions. In all three scenarios, every participant experienced the same starting position upon entering the VR environment and the layout of the environment itself. Additional factors such as the type of objects used in each scenario, participant instructions, and the two surveys filled out by participants after experimenting were also kept constant across all participants to reduce external variability.

## 2.6. Data Analysis

A between-subject experimental design was used in this study, where each participant was randomly assigned to one of the two treatments of audio condition (i.e., no sound or sound). Similarly, the order in which each participant completed the three scenarios was also randomly assigned to avoid possible confounding effects that could arise due to a learning effect. The raw data collected in the individual text files had to be processed to obtain accurate measures of the three dependent variables. All processed data was organized in tabular form in individual Excel files for each measure.

For task completion time, this was done by finding the difference between the first second recorded and the final second recorded. The result was utilized over the final second recorded because the first second recorded differs for each participant. Therefore, offsetting that difference among participants was crucial to obtain accurate task completion times. Eye fixation duration was collected using the summation of seconds relative to the AOIs (the objects). Replay videos of the participants interacting in the VR environment were utilized to collect haptic interaction times by calculating the difference between the second an object was picked up and the second the same object was placed in its respective location.

Summary charts were created to visualize initial trends. Three bar charts were produced to display the average task completion times, eye fixation durations on AOIs, and haptic interaction times with AOIs across all participants for each difficulty level. An additional six bar charts were created to illustrate each difficulty levels average eye fixation durations on AOIs and haptic interaction times with AOIs across all participants for the three categories (food, kitchen, and hazard). All bar charts were constructed on the basis of audio condition.

Considering the study is a two-factor experimental design, the data for each measure was first evaluated on the basis of a two-way ANOVA test. At least one of the three assumptions of ANOVA (i.e., independence of observations, normality of data, and homogeneity of variances) did not hold for each measure, indicating that ANOVA cannot be used for statistical testing of the data. Therefore, both factors (i.e., audio condition and difficulty level) must be tested independently of each other.

To analyze the relationship between audio condition and each measure, a Mann-Whitney test was performed. This test evaluated whether each measure significantly differed between participants who had no sound versus sound. Similarly, a Kruskal-Wallis test was performed to analyze the relationship between difficulty level and each measure. This test evaluated whether each measure significantly differed between participants in difficulty levels one, two, and three. Post-hoc analyses were performed using three pairwise Mann-Whitney comparison tests to evaluate which difficulty level groups differ from each other, considering there were significant differences found in the Kruskal-Wallis test.

Lastly, linear regression analyses were performed for each audio condition to assess whether changes in difficulty level were associated with changes in each measure.

## 2.7. Self-report Data Analysis

### 2.7.1. NASA-TLX Forms

Hart and Staveland's NASA-TLX form is a subjective workload assessment tool used to measure how demanding a task is on a person. Participants rate the task on six scales (workload dimensions):

1) *Mental Demand*: How mentally demanding was the task?
2) *Physical Demand*: How physically demanding was the task?
3) *Temporal Demand*: How hurried or rushed was the pace of the task?
4) *Performance*: How successful were you in accomplishing what you were asked to do?
5) *Effort*: How hard did you have to work to accomplish your level of performance?
6) *Frustration*: How insecure, discouraged, irritated, stressed, and annoyed were you?

Each workload dimension was rated on a 0-100 point scale in increments of five points. A bar chart was produced to display the average points rated across participants for each dimension as well as the average overall NASA-TLX values (i.e., the average across the six dimensions for each participant). Participant workload assessments were analyzed for each dimension along with the overall NASA-TLX values to evaluate any significant differences between participants who had no sound versus sound. Similar to the three measures, each dimension and the overall NASA-TLX values were first evaluated on the three assumptions of ANOVA. Normality of data did not hold for all six dimensions, therefore Mann-Whitney tests were performed to analyze the relationship between audio condition and each workload dimension. However, the three assumptions of ANOVA did hold for the overall NASA-TLX values. Therefore, a T-test was performed to analyze the relationship between audio condition and overall NASA-TLX values.

### 2.7.2. User-feedback Surveys

Two analysts evaluated the self-reported user-feedback surveys. For each participant who experienced no sound, the documented responses were examined for quotations indicating whether having sound would have helped with completing the task. Likewise, for each participant who experienced sound, the documented responses were examined for quotations relating to the helpfulness of the triggered sounds. In addition, quotations regarding how the VR app can be improved were identified for all participants.

By investigating and aggregating the terms used in the quotations, one analyst created an initial taxonomy indicating whether having sound would have helped with completing the task. A second analyst assigned the quotations using the classifications. The two analysts came to agreement on both the classifications and the quotation assignments. The other classifications were created similarly.

# 3. Results

All bar charts were constructed using the mean (average) and standard error (error bars).

**Figure 4** provides the average task completion times across all participants for each scenario. Similarly, **Figures 5** and **6** show the average eye fixation durations and haptic interaction times across all participants for each scenario, respectively. For no sound, the average eye fixation duration was slightly lower than the average haptic interaction time in scenario one (L1), and the average eye fixation durations were higher than the average haptic interaction times in scenarios two (L2) and three (L3). Whereas for sound, the average eye fixation duration was lower than the average haptic interaction time in scenario one (L1), and the average eye fixation durations were about the same (or slightly lower) than the average haptic interaction times in scenarios two (L2) and three (L3).

On average, participants had eye fixation durations and haptic interactions times about half the time of the average task completion times for each scenario, which is consistent for both no sound and sound. In any case, no sound is consistently longer in terms of the three measures regardless of the scenario.

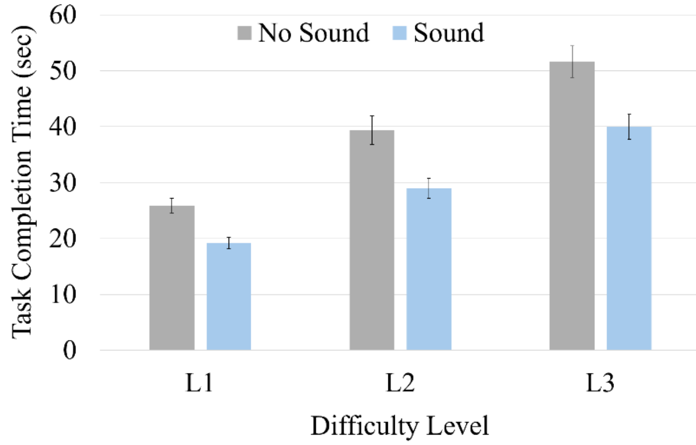


**Figure 4.** Average task completion times.

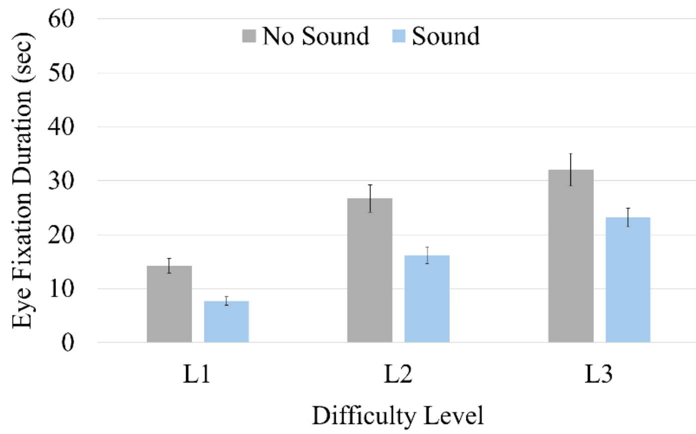


**Figure 5.** Average eye fixation durations on AOIs.

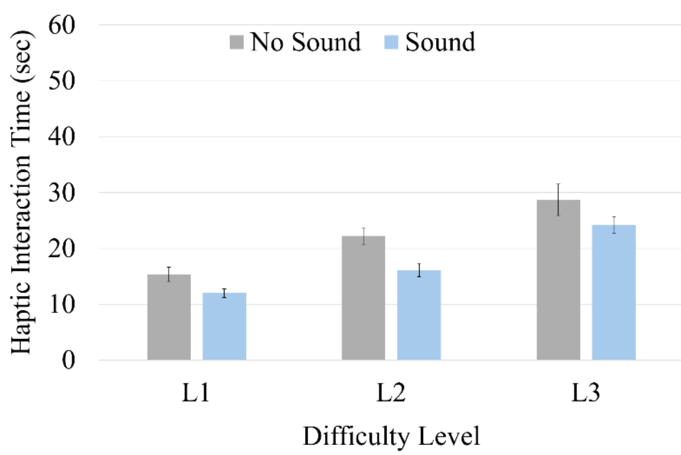
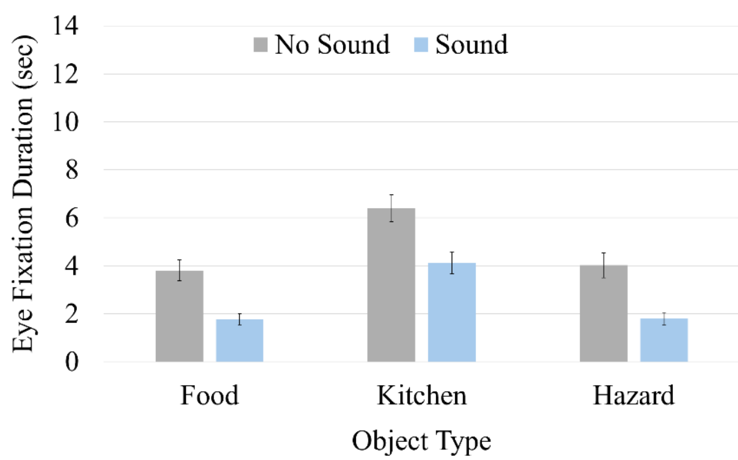
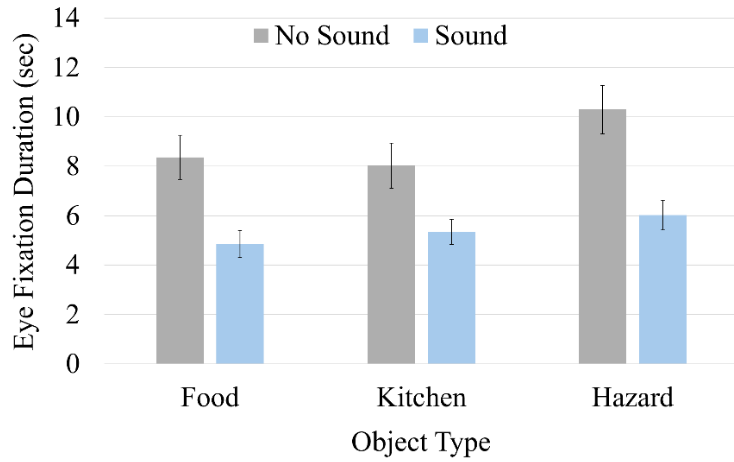


**Figure 6.** Average haptic interaction times with AOIs.

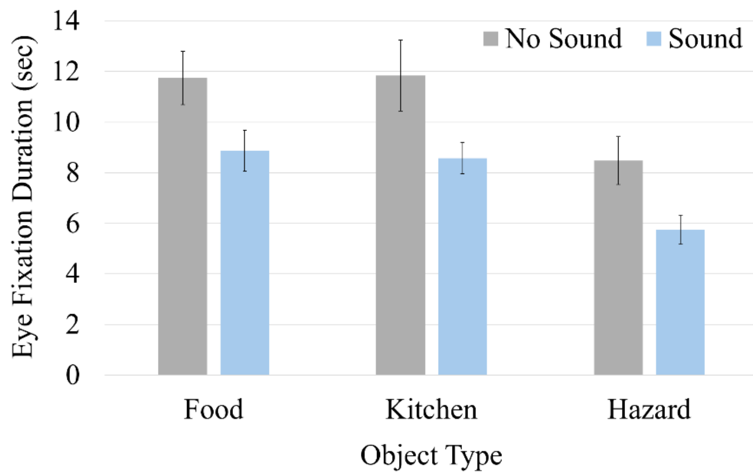
**Figure 7** provides the average eye fixation durations specific to each scenario and relative to the three categories in which to associate objects with. Similarly, **Figure 8** shows the average haptic interaction times specific to each scenario and relative to the three categories. The constant change in eye fixation durations between the three categories suggest there are no visual trends or patterns across the three scenarios. These findings propose that object type (food, kitchen, and hazard) has no effect on eye fixation duration. In contrast, hazardous objects consistently have longer haptic interactions times for both no sound and sound across the three scenarios. Furthermore, kitchen and food objects follow the same trend across the three scenarios, suggesting object type may have an effect on haptic interaction times. The correlation between eye fixation durations and haptic interaction times is not as strong as expected due to the constant change in eye fixation durations between the three categories. In any case, there are substantial differences between no sound and sound in that no sound consistently has longer eye fixation durations and haptic interaction times than sound.




(a)

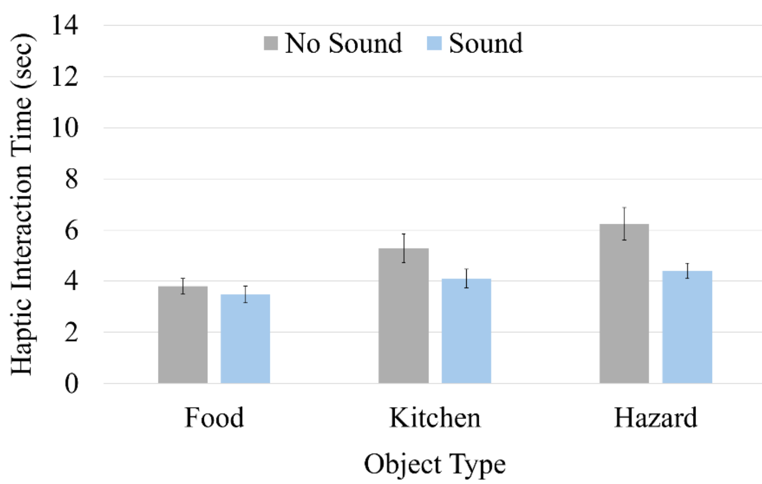



(b)

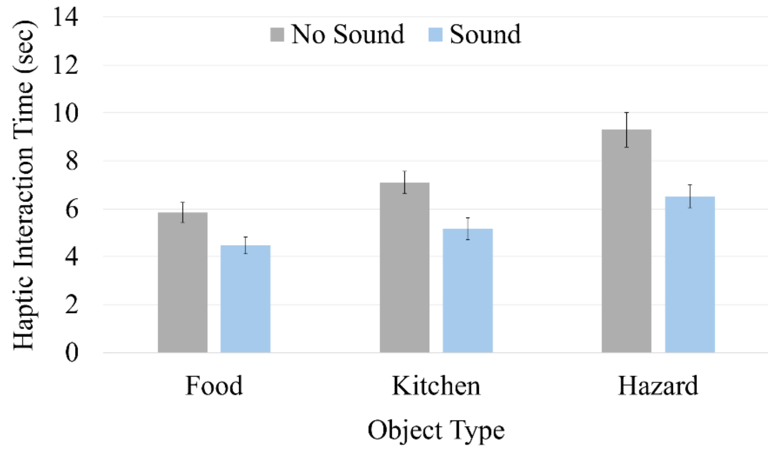



(c)

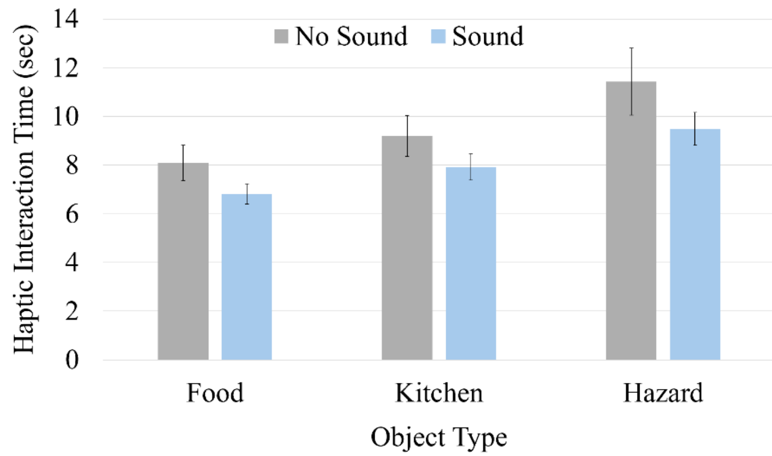
**Figure 7.** Scenario specific average eye fixation durations on AOIs. (a) L1 average. (b) L2 average. (c) L3 average.




(a)




(b)




(c)

**Figure 8.** Scenario specific average haptic interaction times with AOIs. (a) L1 average. (b) L2 average. (c) L3 average.

A significance level of $\alpha = 0.05$ was used for statistical tests.

Mann-Whitney tests indicated significant differences for audio condition on task completion time ($U = 539,\ p < 0.001$), eye fixation duration ($U = 529,\ p < 0.001$), and haptic interaction time ($U = 474,\ p = 0.01$) as seen in **Tables 1**, **5**, and **9**, respectively. Kruskal-Wallis tests showed significant differences for difficulty level (L1: six objects, L2: nine objects, and L3: twelve objects) on task completion time ($H = 75.53,\ p < 0.001$), eye fixation duration ($H = 53.51,\ p < 0.001$), and haptic interaction time ($H = 58.67,\ p < 0.001$) which can be found in **Tables 2**, **6**, and **10**, respectively. Post-hoc analyses were performed using Mann-Whitney tests with $\alpha = 0.017$ (or 0.05/3). For task completion time, significant differences were found for L1 versus L2 ($U = 523,\ p < 0.001$), L1 versus L3 ($U = 110,\ p < 0.001$), and L2 versus L3 ($U = 669,\ p < 0.001$) (see **Table 3**). For eye fixation duration, significant differences were found for L1 versus L2 ($U = 560,\ p < 0.001$), L1 versus L3 ($U = 299,\ p < 0.001$), and L2 versus L3 ($U = 947,\ p = 0.01$) (see **Table 7**). For haptic interaction time, significant differences were found for L1 versus L2 ($U = 674,\ p < 0.001$), L1 versus L3 ($U = 261,\ p < 0.001$), and L2 versus L3 ($U = 683,\ p < 0.001$) (see **Table 11**).

**Table 1.** Audio condition Mann-Whitney test results on task completion time.

| U Statistic | *p*-value |
|---|---|
| 539 | < 0.001 |

**Table 2.** Difficulty level Kruskal-Wallis test results on task completion time.

| H Statistic | *p*-value |
|---|---|
| 75.53 | < 0.001 |

**Table 3.** Difficulty level Post-hoc analyses using Mann-Whitney test results on task completion time.

| Pairwise Comparisons | U Statistic | *p*-value |
|---|---|---|
| (L1,L2) | 523 | < 0.001 |
| (L1,L3) | 110 | < 0.001 |
| (L2,L3) | 669 | < 0.001 |

**Table 4.** Audio condition Linear Regression test results on task completion time.

| Audio Condition | Regression Equation | $R^2$ |
|---|---|---|
| No Sound | Y = 4.29X + 0.33 | 0.450 |
| Sound | Y = 3.47X - 1.88 | 0.487 |

X = Difficulty Level; Y = Task Completion Time

**Table 5.** Audio condition Mann-Whitney test results on eye fixation duration.

| U Statistic | *p*-value |
|---|---|
| 529 | < 0.001 |

**Table 6.** Difficulty level Kruskal-Wallis test results on eye fixation duration.

| H Statistic | *p*-value |
|---|---|
| 53.51 | < 0.001 |

**Table 7.** Difficulty level Post-hoc analyses using Mann-Whitney test results on eye fixation duration.

| Pairwise Comparisons | U Statistic | *p*-value |
|---|---|---|
| (L1,L2) | 560 | < 0.001 |
| (L1,L3) | 299 | < 0.001 |
| (L2,L3) | 947 | 0.01 |

**Table 8.** Audio condition Linear Regression test results on eye fixation duration.

| Audio Condition | Regression Equation | $R^2$ |
|---|---|---|
| No Sound | Y = 2.97X - 2.39 | 0.263 |
| Sound | Y = 2.58X - 7.49 | 0.440 |

X = Difficulty Level; Y = Eye Fixation Duration

**Table 9.** Audio condition Mann-Whitney test results on haptic interaction time.

| U Statistic | *p*-value |
|---|---|
| 474 | 0.01 |

**Table 10.** Difficulty level Kruskal-Wallis test results on haptic interaction time.

| H Statistic | *p*-value |
|---|---|
| 58.67 | < 0.001 |

**Table 11.** Difficulty level Post-hoc analyses using Mann-Whitney test results on haptic interaction time.

| Pairwise Comparisons | U Statistic | *p*-value |
|---|---|---|
| (L1,L2) | 674 | < 0.001 |
| (L1,L3) | 261 | < 0.001 |
| (L2,L3) | 683 | < 0.001 |

**Table 12.** Audio condition Linear Regression test results on haptic interaction time.

| Audio Condition | Regression Equation | $R^2$ |
|---|---|---|
| No Sound | Y = 2.23X + 2.02 | 0.234 |
| Sound | Y = 2.04X - 0.95 | 0.410 |

X = Difficulty Level; Y = Haptic Interaction Time

Linear regression analyses of audio condition on task completion time, eye fixation duration, and haptic interaction time can be found in **Tables 4**, **8**, and **12**, respectively. For each audio condition, difficulty level was used as the predictor (X) and the three measures as outcomes (Y) to test whether difficulty level predicted any of the measures. Difficulty level explains 45% ($R^2$ = 0.450) of the variance in task completion time without sound and 48.7% ($R^2$ = 0.487) with sound (see **Table 4**). In both cases, difficulty level moderately explains the performance of task completion time. The increase in $R^2$ suggests a marginally stronger association between difficulty level and task completion time when sound is present.

For eye fixation duration, difficulty level explains 26.3% ($R^2$ = 0.263) of the variance without sound and 44% ($R^2$ = 0.440) with sound (see **Table 8**). Likewise, difficulty level explains 23.4% ($R^2$ = 0.234) of the variance in haptic interaction time without sound and 41% ($R^2$ = 0.410) with sound (see **Table 12**). Difficulty level modestly explains the performance of both measures when no sound is present, whereas difficulty level moderately explains the performance of both measures when sound is present. The increase in $R^2$ suggests a stronger association between difficulty level and both measures when sound is present.

Overall, sound appears to strengthen the alignment between difficulty level and all three measures compared to no sound.

# 4. Self-report Results

## 4.1. NASA-TLX Forms

**Figure 9** shows the average NASA-TLX points rated across all participants for each workload dimension and the average overall NASA-TLX values (i.e., the average across the six dimensions for each participant). The bar chart was created using the mean (average) and standard error (error bars). It can be interpreted such that higher values (points) indicate a greater workload. Sound predominantly has a marginally higher self-reported workload compared to no sound. The lower values of sound for the performance and frustration dimensions may suggest that participants felt more successful and less stressed in completing the task versus no sound. Temporal demand had the highest self-reported workload for both sound and no sound where participants reported how hurried or rushed the pace of the task was. The large increase in temporal demand for sound may suggest that participants felt more rushed to complete the task as quickly as possible, further validating the significant differences found in task completion time. Overall, although sound produced a slightly higher self-reported workload, both audio conditions reflect relatively low workload levels for the task participants were required to complete during the experiment.

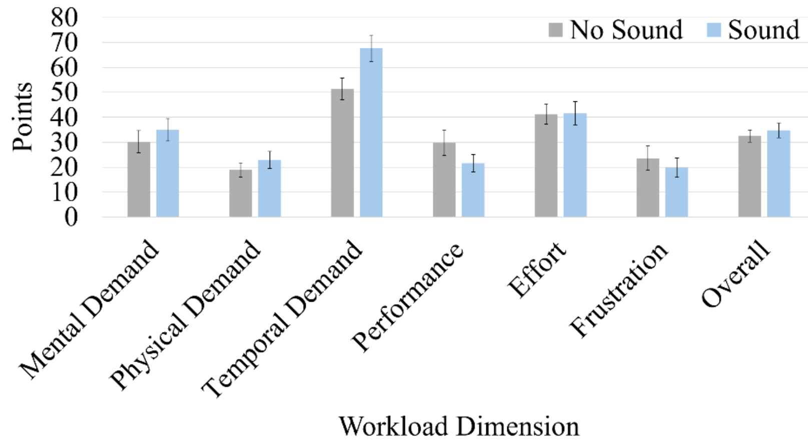



**Figure 9.** Average rated points for workload dimensions and overall TLX values.

A significance level of $\alpha = 0.05$ was used for statistical tests. Mann-Whitney tests indicated no significant differences for audio condition on the following workload dimensions: mental and physical demand, performance, effort, frustration, and the overall NASA-TLX values (see **Table 13**). However, a Mann-Whitney test showed significant differences for audio condition on temporal demand ($U = 190,\ p = 0.007$).

**Table 13.** Audio condition Mann-Whitney test results on NASA-TLX workload dimensions and overall NASA-TLX values.

| Dimension | U statistic | *p*-value |
|---|---|---|
| Mental Demand | 296 | 0.45 |
| Physical Demand | 296 | 0.44 |
| Temporal Demand | 190 | 0.007 |
| Performance | 390.5 | 0.34 |
| Effort | 337.5 | 1.00 |
| Frustration | 361 | 0.67 |
| Overall | 291 | 0.39 |

### 4.2. User-feedback Surveys

The following classifications and quotations refer to the self-reported responses provided by participants in the user-feedback surveys.

#### 4.2.1. Perceived Benefit of Sound for Task Completion

Participants who interacted in the VR app without sound were asked the following question: "Do you think having sound, triggered by gaze on the object, would have helped you in the completion of the task in the experiment?" Participants provided yes or no responses and their reasoning for such a response. Based on the self-reports, seven classifications were identified:

1) "Speed/efficiency" refers to any quotations where "fast" or "quick" was used to describe a response to an object or understanding of an object.
2) "Identification/recognition" means recognizing or knowing what the objects are.
3) "Cognitive load/workload" means spending less time having to think what each object is.
4) "Condition depends on sound design" refers to conditional quotations where participants felt that having sound might have helped them in the completion of the task depending on the sound design.
5) "Rejection" means sound was not needed.
6) "Overwhelm" means too many sounds being present to focus or overstimulating.
7) "Distraction/interference" means being stressed or feeling that sound would hinder the ability to think quickly.

**Table 14.** Seven perceived benefits of sound for task completion with example quotations.

| **Response to Question** (yes, no, or maybe) | **Classification** | **Example Quotation** |
|---|---|---|
| Yes | Speed/efficiency | "It would have made grabbing objects quicker." |
| | Identification/recognition | "It can possibly help me know what object I am currently picking up." |
| | Cognitive load/workload | "I wouldn't have had to expend as much thought on recognizing each thing." |
| Maybe | Condition depends on sound design | "It depends on what noise would be triggered." |
| No | Rejection | "Don't need sound." |
| | Overwhelm | "I think it would be overstimulating." |
| | Distraction/interference | "I think having a sound when I look at an object will have stressed me out." |

**Table 14** shows example quotations in these classifications. Quotations in the table were parsed according to the classification descriptions.

#### 4.2.2. Helpfulness of Triggered Sounds

Participants who interacted in the VR app with sound were presented with the following statement: "Please provide the helpfulness of the triggered sound from 1 to 9, with 1 being not helpful at all and 9 being extremely helpful." Participants provided their response on the nine-point

Likert scale and gave reasoning to why that response was chosen. Based on the self-reports, six classifications were identified:

1) "Primary/confirmatory" means participants actively used sound as a primary or confirmatory signal to recognize objects faster than by visuals alone.
2) "Navigation/placement guidance" means sounds guided where to put objects, reducing uncertainty about destination or orientation.
3) "Cognitive offloading & speed" means sound reduced memory load and accelerated decision-making.
4) "Category dependent" means helpfulness depended on the object type.
5) "Visual dominance" means participants relied on visuals or memory, treating sound as background with little impact on performance.
6) "Distraction/interference" means competing or excessive cues led to confusion and distraction.

**Table 15.** Six perceived helpfulness of triggered sounds with example quotations.

| Helpfulness | Classification | Example Quotation |
| --- | --- | --- |
| Very or Extremely Helpful (scores between 7-9) | Primary/confirmatory | "At some point in the testing, I started to focus on sound alone rather than visuals." |
| | Navigation/placement guidance | "There were times where if I didn't know exactly what the object was the sound would lead me to the right location." |
| | Cognitive offloading & speed | "When I hover over the object the sound helped me identify which object I was getting faster." |
| Helpful (scores between 5-6) | Category dependent | "I could see what the objects were, so the sound was neither helpful nor unhelpful, it was just there." |
| Not Helpful or Somewhat Helpful (scores between 1-4) | Visual dominance | "I don't think I paid much attention to the sound when sorting." |
| | Distraction/interference | "VR is visually intensive, so sound really just distracted me" |

**Table 15** shows example quotations in these classifications. The average score across the twenty-two responses was 5.68 points with a standard deviation of 2.12. Quotations in the table were parsed according to the classification descriptions.

### 4.2.3. Improvements to the VR app

All participants were asked the following question: "How can the virtual reality app be improved?" Based on the self-reports, five classifications were identified:

1) "Physics" means adding basic physics, so objects don't float and maintain sensible orientation when picked up or set down.

2) “Spatial design” means reduce dead space and space objects to minimize distraction and crowding.
3) “Ergonomics” means calibrating user height and reduce bending by enabling pointer-based pickup or raising object height
4) “Focus & accessibility” means increasing the overall visual clarity and provide focus/fit options for users who wear glasses.
5) “Sound design” means make sounds more distinguishable.

**Table 16.** Five improvements to the VR app with example quotations.

| Classifications | Example Quotation |
| --- | --- |
| Physics | “Add gravity.” |
| Spatial design | “The app could be improved if there was less dead space between the kitchen boundaries and walls.” |
| Ergonomics | “If the pointer was activated so that way you don’t have to bend down to pick up objects, you could just point the pointer at the object and press the back button.” |
| Focus & accessibility | “Have settings to fit people that need glasses.” |
| Sound design | “The triggered sounds were hard to distinguish when I rushed to compete the task.” |

**Table 16** shows example quotations in these classifications. Participant responses that did not relate to improving the VR app were omitted. Quotations in the table were parsed according to the classification descriptions.

# 5. Discussion

In summary, this study explored whether the presence of sound facilitates user task performance compared to no sound in a VR-based categorization task. The results indicate that participants who had sound had significantly faster task completion times. Average task completion time, eye fixation duration on AOIs, and haptic interaction time with AOIs were the three measures used to analyze the effects of sound presence versus absence on rapid detection and response.

Task completion time was significantly shorter for the sound treatment, suggesting that the presence of sound, triggered by gaze on the object, may have facilitated faster detection and response, and the results align with the recent research of Peng et al. (2024) which showed that latency was reduced when visual and auditory stimulus were provided together compared to providing visual stimulus alone for a traffic scenario perception task. Furthermore, **Figure 4** (above in the Results section) shows a proportional increase in task completion time across the three difficulty levels for both sound and no sound conditions. Such observed tendency indicates that task difficulty affected completion time similarly across conditions, and more importantly, for each difficulty the audio-conditioned treatment always outperformed the no-audio treatment.

Additionally, eye fixation duration on AOIs was used to examine differences in visual processing between the sound and no sound conditions. Specifically, this measure was used to assess whether the addition of sound enhanced the efficiency of visual processing of task-relevant information and facilitated faster target localization. The significant differences observed suggest that participants exposed to sound may have required less cognitive effort to identify or recognize objects, resulting in faster task completion times. The presence of sounds triggered in relation to an object's category may have enhanced visual processing, thereby leading to shorter eye fixation durations. While variations in eye fixation duration across object categories (food, kitchen, and hazard) suggest that object type did not substantially influence fixation duration, increases in fixation duration were consistent across the three difficulty levels. This consistency indicates that task difficulty exerted a similar effect across all levels. Overall, fixation durations were consistently lower in the sound condition across all difficulty levels, reinforcing the conclusion that sound may have enhanced visual processing during object detection, thereby facilitating faster responses.

Similarly, haptic interaction time with AOIs was used to examine differences in decision-making between the sound and no sound conditions. The significant differences observed may indicate that the reduction in haptic interaction time in the presence of sound is attributable to quicker decision-making or more decisive interactions once a target object has been identified. This finding suggests that the sounds triggered by each object type may have facilitated faster decision-making, as participants repeatedly heard the same auditory cue while visually attending to an object. Consistent with the task completion time results, haptic interaction time increased proportionally across the three difficulty levels for both the sound and no sound conditions. This pattern indicates that task difficulty exerted a comparable effect across both conditions, further highlighting the significant effect of difficulty level on haptic interaction time.

Linear regression modeling was conducted in which each measure (Y) was modeled as a function of task difficulty (X), separately for the sound and no sound conditions. Across all three measures, the regression parameters were comparable between the sound and no sound models, indicating that task difficulty exerted a similar effect across conditions and difficulty levels. In addition, the sound condition yielded higher $R^2$ values for all three measures, suggesting that performance was more predictable when sound was present. Collectively, these findings imply that the presence of triggered sounds reduced variability or noise in participant performance relative to the no sound condition.

NASA-TLX forms were used to assess the effects of the task on participants perceived workload. **Figure 9** (above in the Results section) presents the average self-reported scores for each workload dimension, as well as overall NASA-TLX scores, for both the sound and no sound conditions. Overall, the sound condition was associated with higher workload ratings, particularly for temporal demand, which reflects the extent to which participants felt hurried or rushed during task execution. This finding suggests that the presence of triggered sounds may have increased participants' perceptions of time pressure, which aligns with the task objective of categorizing and placing objects as quickly as possible. The elevated temporal demand observed in the sound condition may be attributable to the sounds facilitating faster task completion, thereby heightening participants' awareness of task pacing. This interpretation is consistent with the significant differences observed for temporal demand. In contrast, no significant differences were found for the remaining workload dimensions, indicating that the task imposed a moderate and comparable workload across the sound and no sound conditions. Furthermore, although no significant differences were observed for overall NASA-TLX scores, **Figure 9** suggests that, on average, the task was perceived as imposing a relatively low workload in both the sound ($M$ = 34.74, $S.E.$ = 2.96) and no sound ($M$ = 32.50, $S.E.$ = 2.47) conditions.

User-feedback survey responses provided additional insight into participants' subjective experiences with the VR app and the use of triggered sounds. Participants' qualitative feedback mostly aligned with the observed performance improvements which indicated that the triggered sounds supported object identification and task execution. Feedback regarding the usefulness and clarity of the triggered sounds highlights the importance of intentional sound design in VR environments, particularly for tasks requiring rapid detection and decision-making.

Ultimately, the presence of triggered sounds enhanced participant performance across all three measures compared to the no sound condition, underscoring the potential value that well-designed VR applications can provide in real-world contexts. The integration of purposeful auditory cues (e.g., alerts, confirmation sounds, or spatialized audio) within VR environments may accelerate users' detection and response processes. Such enhancements may be particularly beneficial in applications involving time-critical decision-making, including training scenarios for first responders or air traffic controllers, as well as interfaces designed for safety-critical tasks. By consistently pairing auditory cues with task-relevant events, users may develop conditioned perceptual and cognitive responses that support faster detection and more efficient decision-making when transitioning to real-world scenarios.

# 6. Limitations and Future Research

Future research should focus on improving both the VR application and the task design. Based on user-feedback surveys, a primary recommendation from participants was to incorporate physics, specifically gravity, into the VR app to improve object orientation and stability after placement. Implementing gravity would reduce the need for users to repeatedly adjust object orientation after placing them down, helping ensure that task completion times more accurately reflect task performance rather than corrective actions.

Additional improvements could also be made to the sound design. Participants noted that the frequencies of the eye fixation triggered sounds were similar, making it difficult to distinguish between them in a time pressure situation. Increasing the spatial separation between objects or varying sound properties based on distance from an object may help users better differentiate auditory feedback. These adjustments could reduce distraction and cognitive overload, potentially leading to improved task performance.

Furthermore, while the categorization task used in this study provides evidence that eye fixation triggered sounds can enhance user task performance and establishes a foundation for applications in training and hazard-related contexts, the task itself remains limited in assessing their effects in more complex or dynamic environments. Future research should therefore investigate the use of eye fixation triggered auditory feedback in real-world scenarios to better understand its impact on performance under more demanding conditions.

Finally, investigation of eye-tracking-triggered visual information feedback is needed to compare its effect against eye-tracking–triggered auditory feedback. In detail, eye fixations have been widely used to trigger adaptive responses such as highlighting visual elements, modifying display layouts, or revealing task-relevant information when users gaze at specific items. Reviews of eye tracking research in learning and instructional settings show that eye fixation triggered feedback has been widely used to adapt content in response to sustained gaze on text, diagrams, or visual elements, helping guide attention, reduce cognitive load, and support learning processes in VR (Ke et al., 2024), a well as enable gaze-based interaction by treating eye fixations as signals of user intent or cognitive focus (Adhanom et al., 2023). Furthermore, prior work demonstrates that eye fixation triggered feedback can meaningfully increase performance. In VR tasks where visual properties of objects changed based on where users fixated, the eye fixation triggered adaptations guided attention and influenced reaction time and accuracy without requiring explicit user actions (Limballe et al., 2022). Therefore, further investigation is needed to see whether visual feedback alone, or auditory feedback alone, or a combined visual and auditory feedback might results in different performances.

# 7. References


Adhanom, I. B., MacNeilage, P., & Folmer, E. (2023). Eye Tracking in Virtual Reality: a Broad Review of Applications and Challenges. *Virtual Reality*, *27*. https://doi.org/10.1007/s10055-022-00738-z

Boyer, E. O., Portron, A., Bevilacqua, F., & Lorenceau, J. (2017). Continuous Auditory Feedback of Eye Movements: An Exploratory Study toward Improving Oculomotor Control. *Frontiers in Neuroscience*, *11*. https://doi.org/10.3389/fnins.2017.00197

Cooper, N., Milella, F., Pinto, C., Cant, I., White, M., & Meyer, G. (2018). The effects of substitute multisensory feedback on task performance and the sense of presence in a virtual reality environment. *PLOS ONE*, *13*(2), e0191846. https://doi.org/10.1371/journal.pone.0191846

Fialho, L., Oliveira, J., Filipe, A., & Luz, F. (2021). Soundspace VR: spatial navigation using sound in virtual reality. *Virtual Reality*. https://doi.org/10.1007/s10055-021-00597-0

Flanagan, P., McAnally, K. I., Martin, R. L., Meehan, J. W., & Oldfield, S. R. (1998). Aurally and Visually Guided Visual Search in a Virtual Environment. *Human Factors the Journal of the Human Factors and Ergonomics Society*, *40*(3), 461–468. https://doi.org/10.1518/001872098779591331

Huang, Q., & Li, Z. (2025). Examining immersive audio's role in flow and gaze behavior in pseudo-VR environments. *Frontiers in Virtual Reality*, *6*. https://doi.org/10.3389/frvir.2025.1691405

Hung, M.-C., Lin, C.-Y., & Hsiao, G. L.-K. (2025). Virtual Reality in Building Evacuation: A Review. *Fire*, *8*(2), 80. https://doi.org/10.3390/fire8020080

Jiménez-Navarro, D., Serrano, A., & Malpica, S. (2024). Minimally disruptive auditory cues: their impact on visual performance in virtual reality. *The Visual Computer*. https://doi.org/10.1007/s00371-024-03707-6

Ke, F., Liu, R., Zlatko Sokolikj, Ibrahim Dahlstrom-Hakki, & Israel, M. (2024). Using eye-tracking in education: review of empirical research and technology. *Educational Technology Research and Development*, *72*. https://doi.org/10.1007/s11423-024-10342-4

Kim, R. S., Seitz, A. R., & Shams, L. (2008). Benefits of Stimulus Congruency for Multisensory Facilitation of Visual Learning. *PLoS ONE*, *3*(1), e1532. https://doi.org/10.1371/journal.pone.0001532

Limballe, A., Kulpa, R., Vu, A., Mavromatis, M., & Bennett, S. J. (2022). Virtual reality boxing: Gaze-contingent manipulation of stimulus properties using blur. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.902043

Losing, V., Rottkamp, L., Zeunert, M., & Pfeiffer, T. (2014). *Guiding visual search tasks using gaze-contingent auditory feedback*. https://doi.org/10.1145/2638728.2641687

Lukas Twardon, Hendrik Koesling, Finke, A., & Ritter, H. (2013). Gaze-contingent audio-visual substitution for the blind and visually impaired. *Proceedings of the ICTs for Improving Patients Rehabilitation Research Techniques*. https://doi.org/10.4108/icst.pervasivehealth.2013.252018

Moraes, A. N., Flynn, R., & Murray, N. (2022). Analysing Listener Behaviour Through Gaze Data and User Performance during a Sound Localisation Task in a VR Environment. *2022 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. https://doi.org/10.1109/ismar-adjunct57072.2022.00102

Nelson, W. T., Hettinger, L. J., Cunningham, J. A., Brickman, B. J., Haas, M. W., & McKinley, R. L. (1998). Effects of Localized Auditory Information on Visual Target Detection Performance Using a Helmet-Mounted Display. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *40*(3), 452–460. https://doi.org/10.1518/001872098779591304

Nikolaos Partarakis, Theodoros Evdaimon, Menelaos Katsantonis, & Xenophon Zabulis. (2025). Training First Responders Through VR-Based Situated Digital Twins. *Computers*, *14*(7), 274–274. https://doi.org/10.3390/computers14070274

Oka, M., & Omori, M. (2023). Why is it difficult for children and adults to follow a person's eye gaze in polynomial social relationships with compound audio-visual stimuli: An eye-tracking study. *PLOS ONE*, *18*(8), e0289404. https://doi.org/10.1371/journal.pone.0289404

Peng, X., Zhang, Y., Jiménez-Navarro, D., Serrano, A., Myszkowski, K., & Sun, Q. (2024). Measuring and Predicting Multisensory Reaction Latency: A Probabilistic Model for Visual-Auditory Integration. *IEEE Transactions on Visualization and Computer Graphics*, 1–11. https://doi.org/10.1109/tvcg.2024.3456185

Pierluigi Diotaiuti, Siena, F. D., Vitiello, S., Di, A., & Mancone, S. (2026). Eye-tracking in school-age specific learning disorders: a practice-oriented narrative review of assessment and gaze-contingent interventions. *Frontiers in Education*, *10*. https://doi.org/10.3389/feduc.2025.1759299

Rashika Tasnim Keya, Heldal, I., Patel, D., Murano, P., & Wijkmark, C. H. (2025). Implementing Virtual Reality for Fire Evacuation Preparedness at Schools. *Computers*, *14*(7), 286–286. https://doi.org/10.3390/computers14070286

Van der Burg, E., Olivers, C. N. L., Bronkhorst, A. W., & Theeuwes, J. (2008). Pip and pop: Nonspatial auditory signals improve spatial visual search. *Journal of Experimental Psychology: Human Perception and Performance, 34*(5), 1053–1065. https://doi.org/10.1037/0096-1523.34.5.1053

Zeng, J., Rebelo, F., He, R., Noriega, P., Vilar, E., & Wang, Z. (2025). Using virtual reality to explore the effect of multimodal alarms on human emergency evacuation behaviors. *Virtual Reality*, *29*(2). https://doi.org/10.1007/s10055-025-01141-0

Zhang, Y., Kinateder, M., Templeton, A., & Warren, W. H. (2025). A Virtual Reality Experiment on Visual and Auditory Guidance for Egress in Road Tunnel Fires. *Fire Technology*, *61*. https://doi.org/10.1007/s10694-024-01668-9