UTILIZING XR TECHNOLOGY TO DELIVER INFORMATION TO PARTICIPANTS THROUGH VISUAL, AUDITORY, AND HAPTIC SENSORY CHANNELS

by

SOLAF MOHAMMAD YOUSEF ATHAMNAH

A thesis submitted in partial fulfillment of the
requirements for the degree of

MAJOR IN ELECTRICAL AND COMPUTER ENGINEERING

2026

Oakland University
Rochester, Michigan

Thesis Advisory Committee:

Wing Yue Geoffrey Louie, Ph.D., Chair
Osamah Rawashdeh, Ph.D., Co-chair
Shadi Alawneh, Ph.D.

To my mother and father

ACKNOWLEDGMENTS

I acknowledge the mentorship and help I received from Prof. Wing Yue Geoffrey Louie, Prof. Jennifer Vonk, Prof. Osamah Rawashdeh, and Mark Brudnak, PhD, also the help I received from Andrea Macklem-Zabel and Absalat Getachew. A special thanks to Prof. Omar Hiari. Without all of your support, this work would not have been possible.

Additionally, this work is part of a bigger project between Oakland University (OU) and Automotive Research Center (ARC), the project members from ARC are Mark Brudnak, PhD, and Ryan Wood and from Oakland University are Prof. Wing Yue Geoffrey Louie (Distinguished Associate Professor, Co-PI, ECE), Prof. Jennifer Vonk (Professor, Psychology), Andrea Macklem-Zabel (Graduate Student, Principal Investigator, CSE), Absalat Getachew (Graduate Student, ISE), and I. They started working on the motivation, design of the study before I joined. I joined the project at the task design and development part, so my contributions to the project are mainly the development of the tasks using unreal engine and helping with the tasks/Cues theoretical design. I also helped with running participant to collect the data used in this thesis.

The authors wish to acknowledge the technical and financial support of the Automotive Research Center (ARC) in accordance with Cooperative Agreement W56HZV-24-2-0001 U.S. Army DEVCOM Ground Vehicle Systems Center (GVSC) Warren, MI.

Solaf Athamnah

ABSTRACT

UTILIZING XR TECHNOLOGY TO DELIVER INFORMATION TO PARTICIPANTS THROUGH VISUAL, AUDITORY, AND HAPTIC SENSORY CHANNELS

by

SOLAF MOHAMMAD YOUSEF ATHAMNAH

Adviser: Wing Yue Geoffrey Louie, Ph.D.

In complex, high-stakes tasks such as remote drone and vehicle operation, effective Human-Machine interfaces are critical for ensuring operators comprehend task-critical information and maintain high situation awareness. However, flooding operators with information can overwhelm their senses, causing critical details to be lost. This raises the question: how should designers determine which sensory channel best suits each information cue? A 3×3×3 study was conducted, with information-to-sensory-channel mapping as a between-subjects factor and task type as a within-subjects factor. Tasks were performed in an Unreal Engine simulation. Preliminary analysis of 18 participants revealed that the visual channel consistently outperformed auditory and haptic channels in both accuracy and reaction time. Additionally, the nature of the information cue had an effect on comprehension. These findings demonstrate that effective multimodal interface design requires strategically mapping information cues to appropriate sensory channels rather than distributing them arbitrarily.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

XR	Extended Reality
SA	Situation Awareness
MRT	Multiple Resource Theory
HMD	Head Mounted Display
FoV	Field of View
RT	Reaction Time
HUD	Heads-Up Displays

CHAPTER ONE

INTRODUCTION

Extended Reality (XR) technologies have come a long way in recent years, and their use cases are increasing day by day. The demand for effective human-computer interaction is higher than ever. Whether in remote drone teleoperation, advanced vehicular navigation, or immersive simulation training [10, 5]. Operators in these use cases are required to process large amounts of information, and systems need to deliver this task critical information without overwhelming them.

Traditionally, human-machine interface design has exhibited a pronounced bias toward the visual sensory channel. Critical system data, environmental states, and spatial alerts are predominantly rendered via heads-up displays (HUDs), complex control panels, and intricate data overlays. However, human cognitive processing is inherently constrained by distinct, physiologically bounded resource pools. When an operator is forced to process an overwhelming amount of information through a single sensory modality, it creates an "input bottleneck," resulting in sensory channel saturation, increased cognitive load, and a rapid degradation of performance. While modern XR systems and simulation environments have the technical capability to deliver high-fidelity spatial audio and complex haptic feedback, these modalities are often relegated to secondary, they are used as redundant information source rather than a primary one.

Despite the multi-sensory capabilities of modern XR systems (including advanced headsets, surround sound headphones, and full-body haptic feedback suits), a significant gap exists in the formalized understanding of how to optimally map specific types of

critical information to these available sensory channels. Current interface design often defaults to visual dominance or arbitrary multimodal distribution without a rigorous empirical framework to predict how different information-to-modality mappings affect operator comprehension, reaction time, and overall accuracy in dynamic environments [6, 7].

To address this gap in human-computer interaction [8] and interface design, this thesis investigates the efficacy of distributing information across multiple sensory channels to reduce cognitive bottlenecking. This research is driven by the following primary research question:

"To what extent does the sensory modality (Visual, Auditory, Haptic) influence the comprehension accuracy of information cues ("What?", "Where?", "When?")."

Ultimately, this thesis aims to construct a scientifically grounded, empirically validated taxonomy that dictates not only the efficacy of multimodal integration but prescribes precise mappings between data types and sensory channels. The resulting principles are designed to inform the development of next-generation, high-performance XR interfaces that systematically enhance situational awareness and optimize cognitive load management.

CHAPTER TWO

BACKGROUND AND RELATED WORK

To effectively design a system that tests the optimization of information delivery, it is necessary to ground the experimental approach in established cognitive psychology frameworks. This chapter reviews the theoretical models that underpin human information processing, specifically focusing on Situation Awareness (SA) and Multiple Resource Theory (MRT), and how they relate to the design of complex task environments.

2.1 Situation Awareness (SA)

Situation Awareness is a foundational concept in human factors engineering, originally formalized by Endsley [1]. SA describes a human operator's internal understanding of their dynamic environment and is defined across three hierarchical levels: Level 1 (Perception): The perception of the elements in the environment within a volume of time and space. Level 2 (Comprehension): The synthesis of disjointed Level 1 elements to understand their collective meaning. Level 3 (Projection): The ability to forecast future states of the environment based on Level 2 comprehension [1, 11].

This thesis focuses entirely on the critical “input bottleneck” of Level 1 SA (Perception). For an operator to achieve higher-level comprehension and decision-making, they must first successfully perceive the foundational data points of the environment.

Using Endsley’s framework, we identified the essential pieces of information that an operator must move from their working memory to long-term memory to build an accurate mental model (or schema) of a task. This includes understanding the task setting, the goal, the available action space, and the specific meaning of environmental cues.

Because the working memory has limited capacity, operators must focus their attention efficiently. Therefore, the experimental design of this study incorporates structured tutorial sessions for each task, ensuring participants fully encode the meaning of individual cues before being tested on their ability to perceive and synthesize multiple cues simultaneously.

2.2 Multiple Resource Theory (MRT)

While Situation Awareness dictates what information must be perceived, Wickens' Multiple Resource Theory (MRT) formulates the neurocognitive constraints governing how that information can be simultaneously processed [2]. MRT fundamentally rejects the notion of a singular, monolithic pool of general cognitive resources. Instead, it posits that human attention acts as a dynamic allocator across discrete, semi-independent resource reservoirs dedicated to distinct sensory modalities (e.g., visual versus auditory) and cognitive processing codes (e.g., spatial versus verbal processing) [10].

A central tenet of MRT is that two tasks or streams of information will interfere with one another (causing cognitive bottlenecking and performance degradation) if they compete for the same resource pool. For example, forcing a user to process three simultaneous streams of complex visual information (stacking) will cause visual channel saturation. However, if those three streams of information are distributed across different sensory channels (such as presenting the "What" visually, the "Where" auditorily, and the "When" haptically) MRT suggests that the operator can process the data in parallel with significantly less interference.

2.3 The (What, Where, and When) Framework

To systematically test Level 1 SA perception, the information presented to the user must be categorized. Drawing inspiration from the journalistic "5W1H" framework (What,

Where, When, Who, Why, How), this study isolates three fundamental questions that operators must answer to understand their immediate environment: “What?” (Characteristics): Information related to the identity of an object or an environmental status (e.g., Is it a civilian or a supply box? Is it raining or clear?). “Where?” (Localization): Information establishing spatial relationships relative to the operator or the controlled object (e.g., Is the target above, below, left, or right?). “When?” (Temporal Aspect): Information detailing the timeline of events that affect immediate action (e.g., Is the event happening now, or is it delayed?).

By standardizing the information delivered across three distinct tasks into these three categories, the study ensures that the resulting data regarding sensory modalities is highly generalizable, rather than being tied to the specific mechanics of a single simulation. Additionally, only using 3 information cue types would reduce the number of participants needed for the study.

2.4 XR Hardware Constraints and Saturation Thresholds

While foundational cognitive theories frame the human limitations of sensory processing, the physical constraints of contemporary XR hardware further exacerbate these bottlenecks. Visual rendering in Head-Mounted Displays (HMDs) is frequently bounded by restricted Fields of View (FOV) and the vergence-accommodation conflict, forcing operators to expend unnatural cognitive effort to focus on overlapping digital artifacts. Furthermore, while auditory spatialization algorithms are highly advanced, acoustic overlap in simulated environments risks destructive audio masking. Similarly, haptic actuator matrices face physical resolution limits; localized tactile feedback becomes entirely ambiguous when stimuli are deployed in excessively dense physiological clusters.

Therefore, optimizing multimodal distribution is not merely a theoretical cognitive exercise, but a mandatory engineering necessity to bypass the strict physical rendering thresholds of modern XR apparatuses.

CHAPTER THREE

METHODS

To test our research question, a novel 3 task virtual environment was developed and implemented for an in person human subject research study. The following sections describe the study design, task design, and cues design, and their development.

3.1 Study design

Our study is answering this research question: Does individual performance differ in terms of reaction time and/or accuracy as a function of information to sensory channel mappings across multiple trials within each task? To answer this question, a mixed experimental design was employed with a 3 × 3 × 3 structure. Three independent variables, 3 information cues (where, what, and when) to 3 sensory channels (Visual, Auditory, Haptic) mapping space (between subjects), with 3 task types as within-subjects factor. The study design can be found in Figure 1. The research question will be answered by measuring participant performance, using two dependent variables accuracy and reaction time.

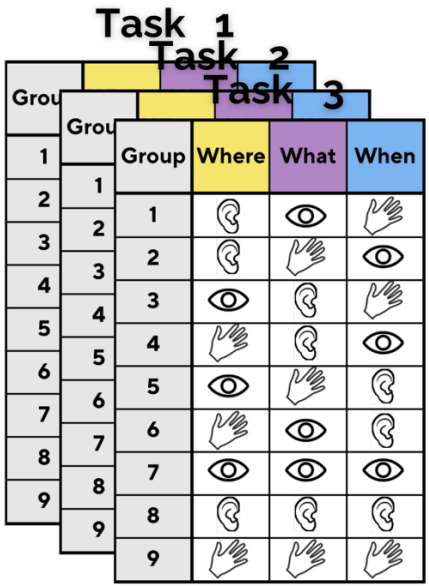

Figure 1 Study design

Participants

The study requires a total of 108 participants divided into 9 groups, 12 each randomly assigned. 6 of the groups get their information cues distributed on multiple sensory channels and 3 of the groups gets all three cues stacked on one sensory channel.

Exclusion Criteria. Participants were excluded if they met any of the following conditions, many of which align with the safety requirements of the Teslasuit haptic feedback system. Individuals who were not fluent in English were ineligible, as task instructions required adequate language comprehension. Those with severe visual, auditory, or tactile impairments were excluded, as the study required intact perception across all three sensory channels. Medical exclusion criteria encompassed cardiac diseases (hypertension, arrhythmia), muscular diseases (myasthenia, myopathy), circulatory disorders (coagulopathies, thrombophilia), nervous system diseases (peripheral

neuropathies, epilepsy, recent strokes), pulmonary conditions (asthma, COPD), skin diseases (dermatitis, psoriasis), metabolic disorders (diabetes, chronic kidney disease), oncological conditions, organ transplantation history, and mental health disorders. Individuals with implanted metallic materials or electronic devices were excluded due to potential interaction with the suit's electrical stimulation. Additional exclusions applied to those with a history of severe motion sickness, vertigo, or photosensitive epilepsy; those in acute medical states (recent injuries, post-surgery, recent tattoos); those experiencing dehydration, sleep deprivation, or substance influence; anyone with a low pain or electrical stimulation threshold or fear of electrical contact; and women during menstruation, with intrauterine devices, or who were pregnant. Participants taking specific medications including antipsychotics, anticholinergics, antidepressants, corticosteroids, and neuromuscular blocking agents were also excluded. Finally, participants have to be over 18 or under 65 years of age, are able to comfortably wear the VR headset, headphones, and haptic jacket for approximately 40 minutes, not with any non-removable jewelry or body piercings, and their whose chest or waist measurements below the haptic jacket's maximum supported dimensions (46 inches chest, 42.5 inches waist).

Procedures

Before coming to the lab, participants had to fill out an exclusion form, if they qualified, they would book an appointment with us. Upon arrival the participant is presented with the consent form and if they choose to proceed, they are assigned randomly to one of the 9 groups and a specific task order (counterbalanced), we change the task order once we have one participant per group for that order, ensuring that no single task consistently appeared first or last. They then fill out a demographic form, including age,

gender, VR experience, video game experience. Afterwards they are introduced to the XR technology, a Varjo XR3 VR head mounted display (HMD), a Sony WH-1000XM5 headphones, and a Teslasuit haptic feedback suit, explaining the electro muscle stimulation technology. After putting on the haptic suit we start with calibrating the feedback using the Teslasuit software section by section, the feedback is set by the participant to a comfortable level where they can feel every section without the feedback being distracting, the feedback level was limited to a maximum of 20% by the institution review board. Then they put the rest of the gear on. The participant starts with a task based on the task order they were assigned at the start.

Each task starts with an intro that sets the task goal and overall setting of the task, then introduces the user interface, then the 3 information cues (where, what, and when) needed to perform the task, one at a time. To limit the number of participants needed for the study, the information cues (Where, What, and When) answers had to be binary. Meaning, each information cue can have two possibilities. When introducing the cues one at a time, the participant is presented with the 2 possibilities of that cue followed by a test where they are presented with one of the two possibilities at random and cannot proceed unless they get 4/5 correct, if they don't, their score would be reset, and they must try again. In that test, if the system detects that they got 3 of the same cue, it will force the other possibility to show up as the 4^{th} event, then goes back to random. Once they are done with each cue, they will learn all 8 possible combinations and which action is correct for that specific combination. Then the participant starts a training session where they get 16 events, the 8 possible combinations 2 times in random order, and a combination does not repeat until the 8 possible happen. Here the participant gets if their action was correct or

not and what the correct action was. Lastly in the task, the participant goes through the final test where they get a similar thing to the training, 16 events, all 8 possible combination 2 times in random order without repetition until all 8 happen. But this time they only get if what they did was correct or not. The overall task protocol can be seen in Figure 2.

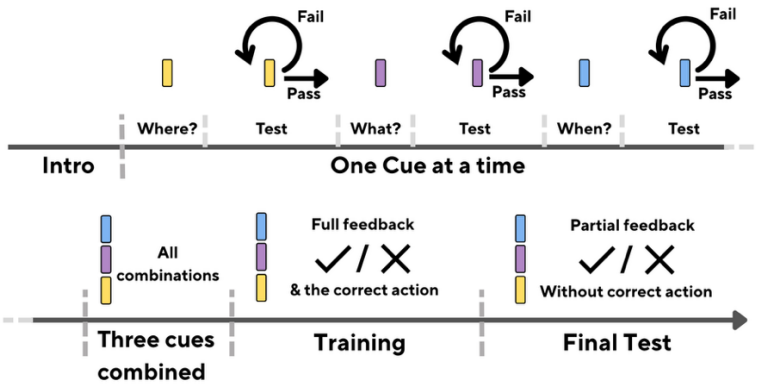

Figure 2 Task protocol

While running the study the researcher has a view of what the participant is seeing (Figure 3 (1)), it allows the researcher to know where the participant is and help them calibrate their position at the start of the experiment. Figure 3 (2) shows the camera cutout

the participants can see around the input controller, this helps those who are not familiar with game controllers see it and make the actions.

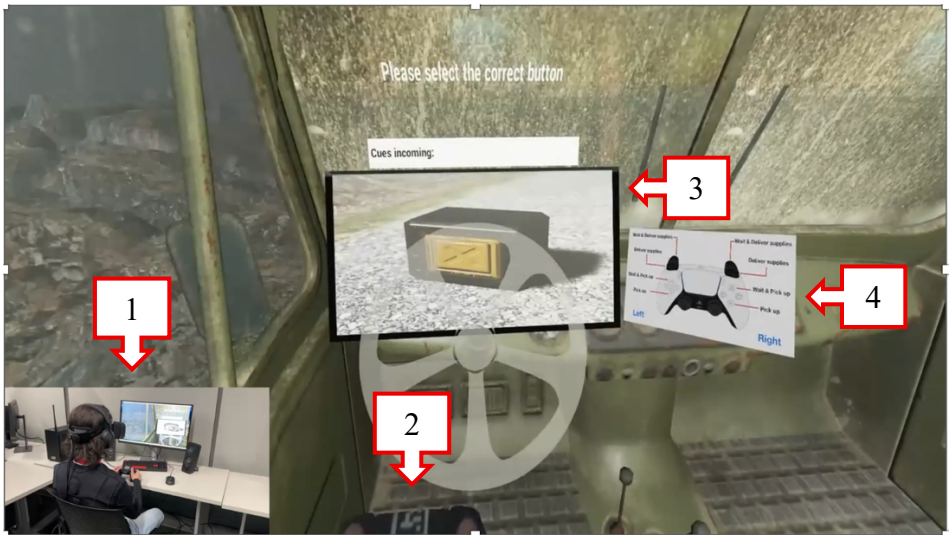

Figure 3 A Participant doing the study

After the participant is done with each task, they fill out a workload questionnaire and answer 3 quick interview questions. When the participant is done with all 3 tasks, they fill out 2 more questionnaires, the system usability questionnaire and the iq presence questionnaire.

3.2 Information Cues Design: Generalizability and Stackability

3.2.1 Generalizability

In the following section we will be describing how we approached each information cue for generalizability across the three tasks (**Task 1**, **Task 2**, and **Task 3**):

"What" generalizability

The "What" cues represented different characteristics of task relevant objects or conditions, ranging from the object identity (**Civilian** vs. **Supply in task 1**) to

environmental status affecting your action (**Storm** vs. **Light rain in task 2**, or **Asphalt** vs. **Gravel in task 3**).

"Where" generalizability

These cues establish spatial relationships to either the operator or the controlled object in the task. The spatial dimensions were, lateral (**Left** vs. **Right in task 1**), vertical (**Above** vs. **Below in task 2**), and longitudinal (**Front** vs. **Behind in task 3**) positioning. To ensure that we are providing accurate spatial cues, we looked at a comprehensive practical review done by Alessandro Carlini [14] that highlights the Elements influencing auditory localization. Such as, Sound frequency spectrum, Sound intensity, Pointing methods, Training, and Auditory localization illusions.

"When" generalizability

These cues were varied on the past present future timeline, leading to different actions based on the temporal aspect of the task. Starting with present vs. future (Now **(Arriving now)** vs. Delayed **(Arriving later) in task 1**), then close future vs. further away future (Happening soon **(Low battery)** vs. Happening later **(Full battery) in task 2**), lastly past vs. future (Happened **(Already flat)** vs. Will happen **(Deflating) in task 3**). To make those cues we had to use the location and movement, so if we wanted to represent as here now it would be relatively close and static, but to represent it as arriving later it would be relatively far and moving towards us.

3.2.2 Designing for Cues stackability

Because the experimental design required delivering cues to participants either distributed across multiple sensory channels or stacked on one channel. The cue had to be stackable. Stackability here means that three distinct pieces of information (What, Where,

and When) could be transmitted simultaneously on a single sensory channel (Visual, Auditory, or Haptic) without masking or interfering with one another.

To achieve stackability different approaches were used depending on the sensory channel, the available bandwidth and the things you could manipulate on the channel.

Visual cue stackability

The screen was divided into three distinct sections, allowing the "What" (e.g., an image of a civilian), the "Where" (e.g., a shadow on the left), and the "When" (e.g., a moving shadow) to be processed as independent visual elements. That decision came after debating showing multiple pieces of information on one cue, like showing a civilian on the left side walking towards the road side, where the civilian gave us the “what”, the “where”, and “when” pieces of information, or dividing the screen into three section and showing each section showing a piece of information on a different cue. This choice allowed us to compare the same cue when it is the only visual cue to when it was presented with other visual cues.

Auditory cue stackability

Stackability in the auditory channel is heavily constrained by human perceptual limits. Based on a Systematic Review done by Chanbeom Kwak [15], normal hearing listeners could perceive and differentiate approximately three to five sound sources, things like acoustic properties, spatial separation, background noise, aging, and attention capacity can all affect that number. Therefore, to prevent auditory masking, the three simultaneous audio cues were designed with distinct acoustic properties.

Haptic cue stackability

While the Teslasuit provides full-body haptic stimulation, the logistics of running 108 participants in-person necessitated limiting direct skin contact to the arms to maintain study efficiency and hygiene. When designing the haptic cues, we had to divide the available bandwidth (the arms) into 3 main sections, the upper arm, the mid arm, and the lower arm section, which allowed us to deliver the three cues without one cue overpowering or interfering with the other.

3.3 Study setup

Three tasks were developed for the study using Unreal Engine 5.1, a vehicle operation task, a remote drone operation task, and a vehicle damage assessment task. Having more than one task ensures that our findings can be more generalizable. While designing those tasks, we ended up following the same approach and setup for each one. So, the tasks have a similar underlying structure shown before in Figure 2 and Figure 4.

The participants get instructions on how to proceed in the tasks with the text banner shown in Figure 4 (1), those instructions are also being read to them through the headphones. Figure 4 (2) shows a banner that shows the participants which channel to expect a cue to come on, it flashes that channel before showing the cue. The display shown in Figure 4 (3) and Figure 3 (3) is where the participants get their visual cues, this display was moved to center of the participant's field of view from an earlier version were it was on the side, that change was to keep the cues available due to the nature of the visual sense. Figure 4 (4) and Figure 3 (4) shows the available action space and its mapping to the input device (PS5 Controller), this action space banner changes from a task to another.

The main differences between the tasks are the setting, the goal, the cues, and the action space and their mapping. The participants look for 3 cues around them and piece them together to make an action, having 3 cues gives us 8 unique combinations with each cue being different, that means that we needed 8 different actions to have an action tied to each single combination of cues in each task. After receiving the 3 cues, the participant uses the input device to choose the appropriate action.

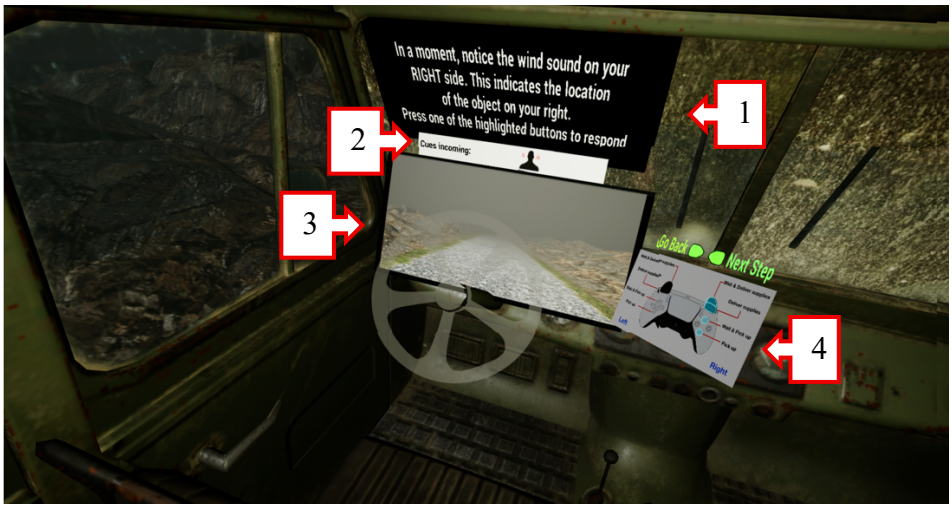

Figure 4 Virtual environment setup

3.4 Task design

The following subsections describe each task in detail. Each task description is accompanied by the task-specific action space table, controller mapping diagram, and representative images of the visual, auditory, and haptic cue presentations as experienced by participants.

Figure 5 shows an example of the visual cues from task 1 (Civilian (right image) vs. Supply (left image)), notice how they are presented on the display in front of the participant. This approach is also followed for the other tasks for the visual cues.

Figure 5 Task 1 visual cues example

Task 1: Vehicle operation task

In this task the participant is a delivery driver, their job is to pick up supplies and deliver them to civilians, if the civilian or the box are not there yet they have to choose actions with wait in them. So, you need to look for three things, where the thing is, what is it a civilian or a box, and when will it arrive. Table 1 shows the available action space and Figure 6 shows its mapping to the input controller.

Table 1 Task 1 Action Space

Event	What	Where	When	Cue Combination	Decision
1	Civilian (0)	Left (0)	Now (0)	000	Deliver supply on left
2	Civilian (0)	Left (0)	**Delayed (1)**	001	**wait** left & deliver supply
3	Civilian (0)	Right (1)	Now (0)	010	Deliver supply on right
4	Civilian (0)	Right (1)	**Delayed (1)**	011	**wait** right & deliver supply
5	Supplies (1)	Left (0)	Now (0)	100	Pick up supply on left
6	Supplies (1)	Left (0)	**Delayed (1)**	101	**wait** left & pick up supply
7	Supplies (1)	Right (1)	Now (0)	110	Pick up supply on right
8	Supplies (1)	Right (1)	**Delayed (1)**	111	**wait** right & pick up supply

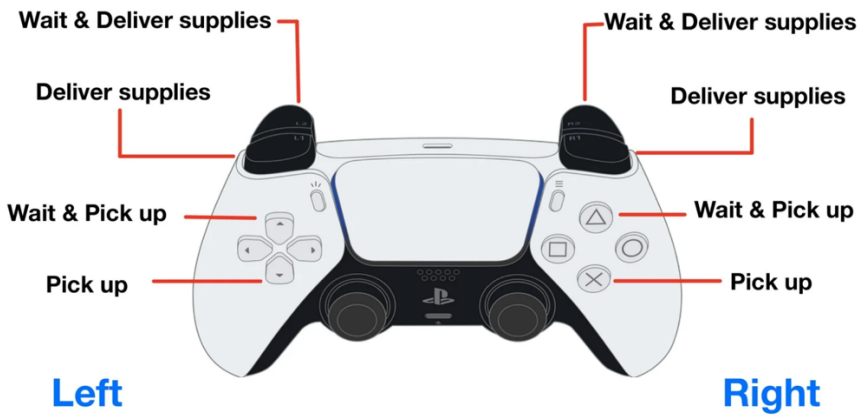

Figure 6 Action space mapping in task 1

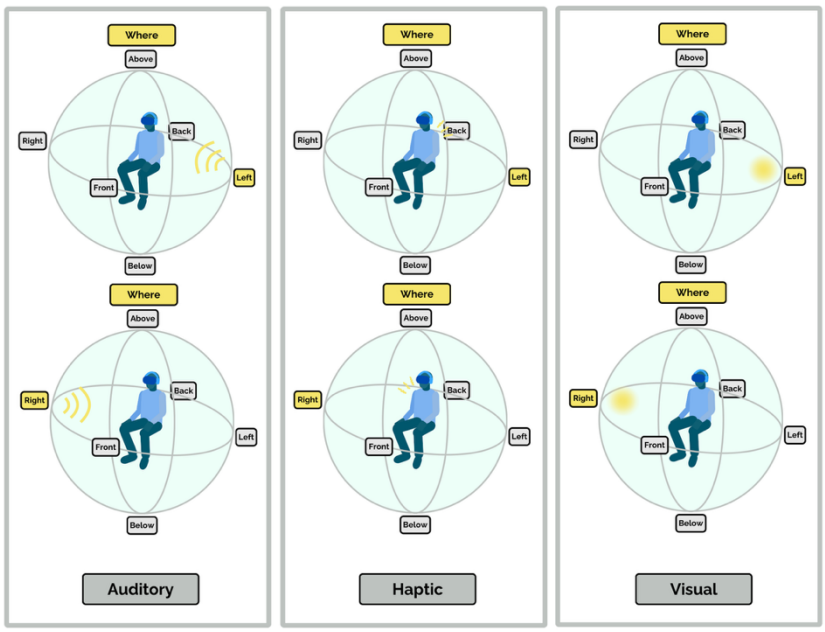

Figure 8 Task 1 "Where?" Cues, Left or Right

The "Where" cue in this task is either left or right, it indicates that the supply box or the civilian is on that side. Auditorily, the participant hears some wind rumbling on their left or right ear. Haptically, they feel feedback on the left or right shoulder. Visually, the participant sees a shadow on the display either on the left or right side.

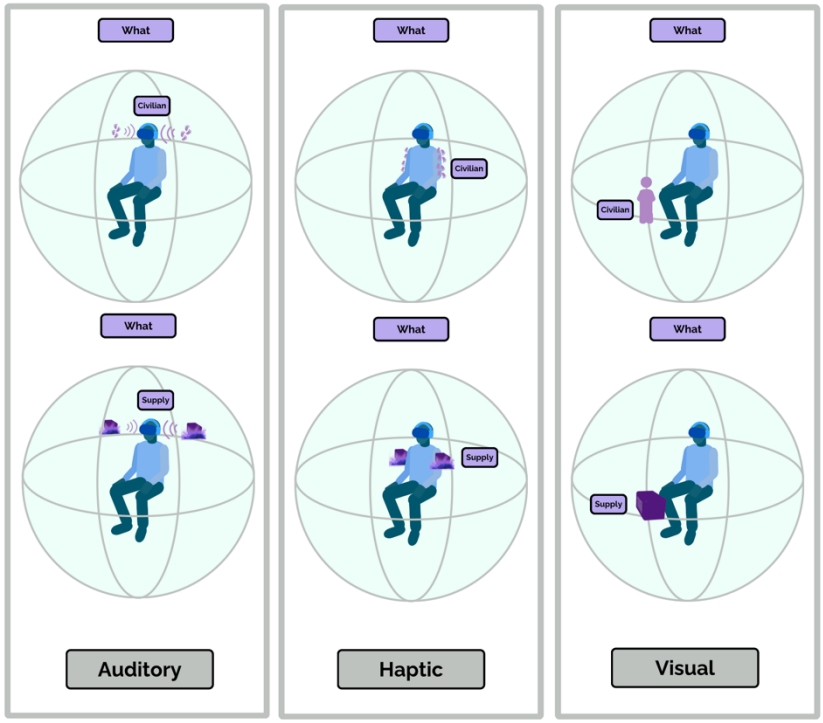

Figure 7 Task 1 "What?" cues, Civilian or Supply box

The "What" cue is either a civilian or a supply box. Auditorily, the participant hears footsteps for a civilian and a box drop sound for a supply box. Haptically, they feel a moving haptic signal for a civilian and a static strong haptic

signal for a supply box. Visually, they see an image of a civilian or an image of a supply box on the display in front of them.

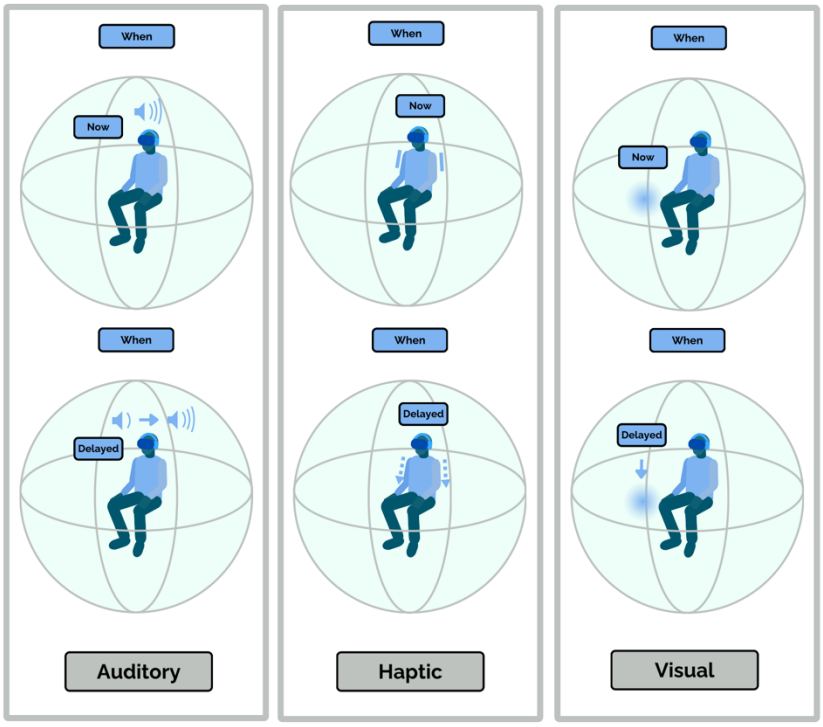

Figure 9 Task 1 "When?" cues, Now or Delayed

For the "When" cue it is either now or delayed, meaning that the civilian is on the side of the road or they are moving to it or the supply box is on the side of the road or being dropped so it could be picked up when it arrives. Auditorily, the participant hears a rumbling sound at a constant volume for now and a rumbling sound getting closer and increasing volume for delayed. Haptically, static feedback for now and moving feedback for delayed. Visually, they see a static shadow on the display for now or a moving shadow for delayed.

Task 2: Remote drone operation task

In this task the participant is a remote drone operator, they are trying to take picture of a vehicle. There are three things to look for, where the vehicle is, what the weather is like, and how much battery do we have left (when should I return the drone). The following Table 2 shows the available action space and Figure 10 shows its mapping to the controller.

Table 2 Task 2 Action Space

Event	What	Where	When	Cue Combination	Decision
1	Storm (0)	Above (0)	Low battery (0)	000	Take a picture above with filtered lens and return
2	Storm (0)	Above (0)	**Full** battery **(1)**	001	Take a picture above with filtered lens and **continue**
3	Storm (0)	Below (1)	Low battery (0)	010	Take a picture below with filtered lens and return
4	Storm (0)	Below (1)	**Full** battery **(1)**	011	Take a picture below with filtered lens and **continue**
5	Light Rain (1)	Above (0)	Low battery (0)	100	Take a picture above with normal lens and return
6	Light Rain (1)	Above (0)	**Full** battery **(1)**	101	Take a picture above with normal lens and **continue**
7	Light Rain (1)	Below (1)	Low battery (0)	110	Take a picture below with normal lens and return
8	Light Rain (1)	Below (1)	**Full** battery **(1)**	111	Take a picture below with normal lens and **continue**

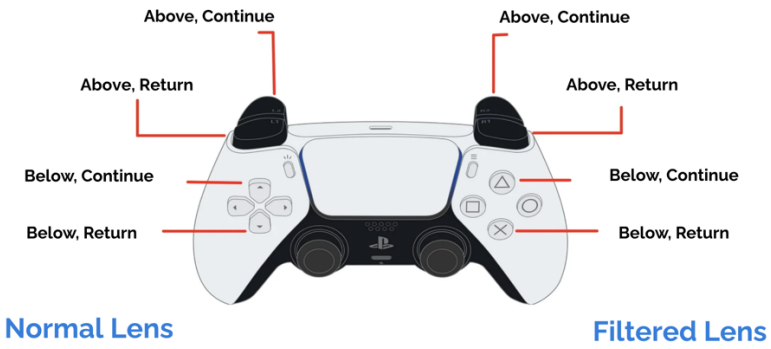

Figure 10 Action space mapping in task 2

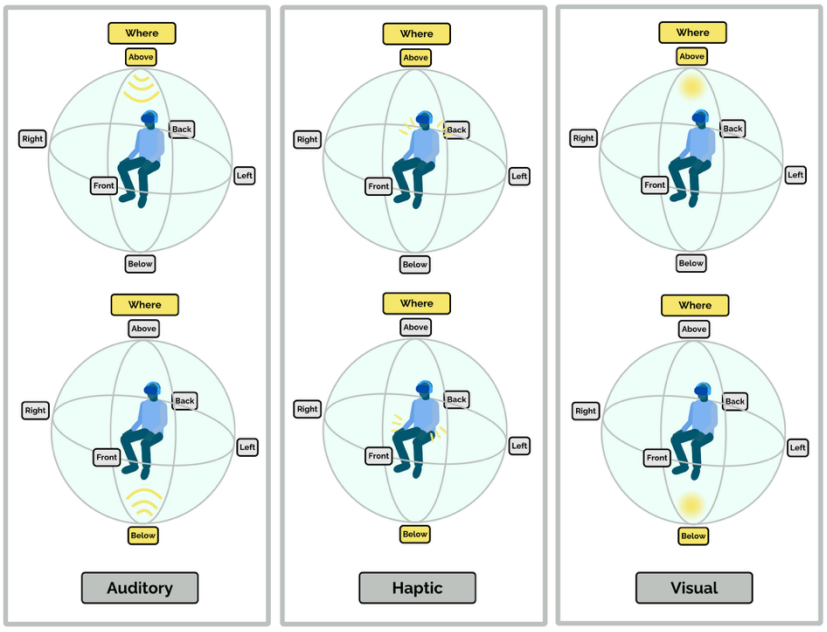

Figure 11 Task 2 "Where?" cues, Above or Below

The "Where" cue in this task is either above or below, that tells the participant where the vehicle they are trying to take a picture of is located. Auditorily, the participant hears car engine noise coming from above or below. Haptically, they get feedback at the top of their arm or the bottom side. Visually, the participant sees a picture of a car on top of a cliff or below the cliff.

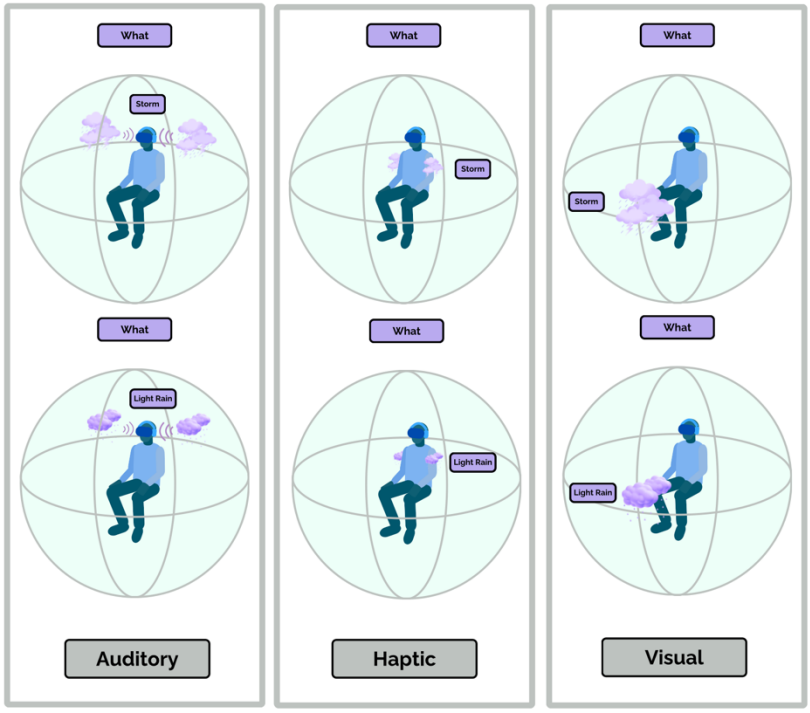

Figure 12 Task 2 "What?" cues, Storm or Light Rain

The "What" cue is related to the current weather, its either a storm or light rain. Auditorily, the participant hears the sound of a storm with thunder or the sound of light rain. Haptically, they either feel strong haptic feedback

resembling thunder or light feedback moving around on their arm representing rain drops. Visually, they see on the screen in front of them the field view of the drone, they see there an image or a thunderstorm or an image of light rain.

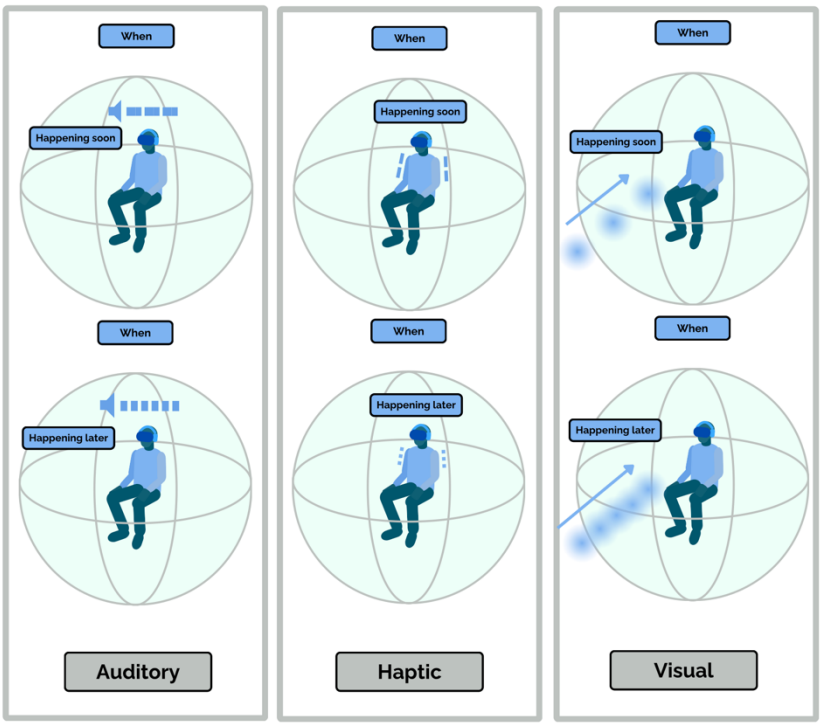

Figure 13 Task 2 "When?" cues, Happening soon or Happening later

The "when" cues for this task depend on the current status of the drone's battery and when will it run out of battery. Auditorily, the participant hears slow beeping indicating that the battery is low and it's running out soon, they hear fast beeping if the battery is full. Haptically, they feel slow pulses for low battery and fast pulses for a full battery. Visually, the display in front of them flashes slowly for low battery and flashes quickly for full battery.

Task 3: Vehicle damage assessment task

In this task you are a passenger in a vehicle and your job is to check if the tire needs to be changed or just patched. You need to look for 3 things, where is the bad tire, what type of road we are driving on and if the tire is flat or is leaking air (when will the tire be flat). Table 3 shows the available action space for this task and Figure 14 shows the mapping to the controller.

Table 3 Task 3 Action Space

Event	What	Where	When	Cue Combination	Decision
1	Asphalt (0)	Front (0)	Flat (0)	000	Change front tire and fill it to high pressure
2	Asphalt (0)	Front (0)	**Deflating (1)**	001	**Patch** front tire and fill it to high pressure
3	Asphalt (0)	Back (1)	Flat (0)	010	Change back tire and fill it to high pressure
4	Asphalt (0)	Back (1)	**Deflating (1)**	011	**Patch** back tire and fill it to high pressure
5	Gravel (1)	Front (0)	Flat (0)	100	Change front tire and fill it to low pressure
6	Gravel (1)	Front (0)	**Deflating (1)**	101	**Patch** front tire and fill it to low pressure
7	Gravel (1)	Back (1)	Flat (0)	110	Change back tire and fill it to low pressure
8	Gravel (1)	Back (1)	**Deflating (1)**	111	**Patch** back tire and fill it to low pressure

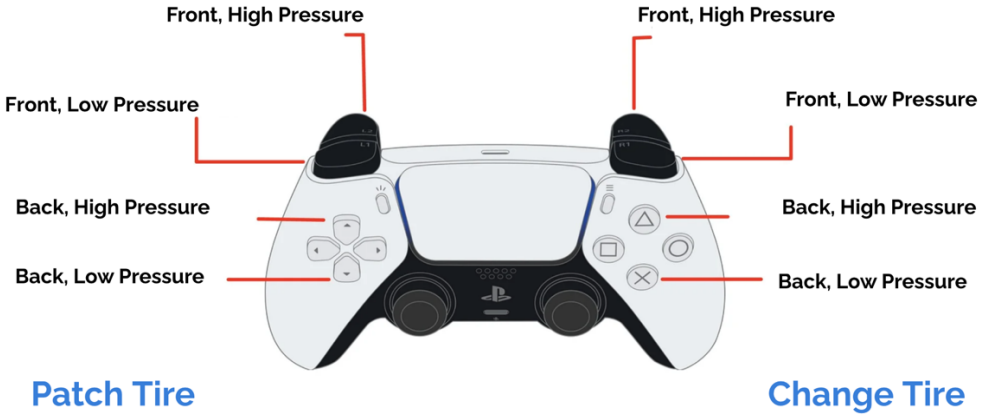

Figure 14 Action space mapping in task 3

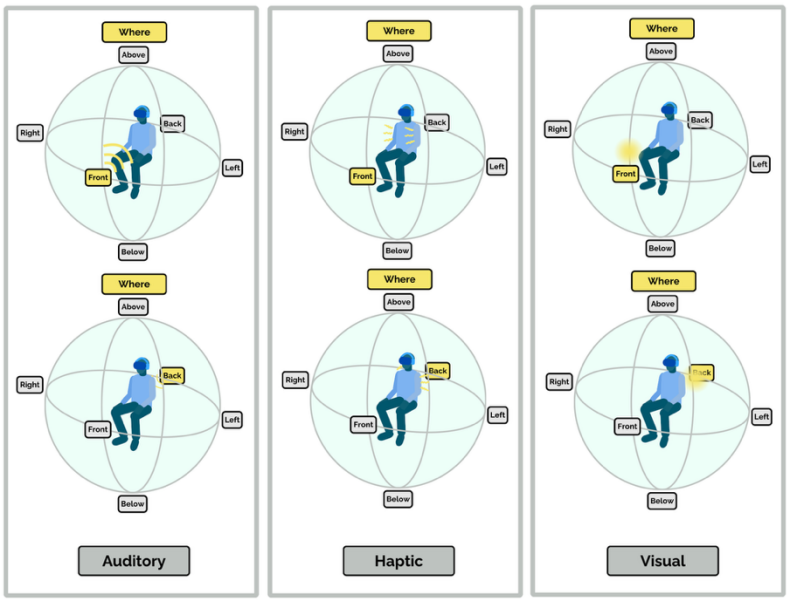

Figure 15 Task 3 "Where?" cues, Front or Back

The "Where" cue in this task is either front or back, this indicates that the bad tire is in the front part of the vehicle or the back part. Auditorily, the participant hears the sound of rumbling coming from the front or the back side of the vehicle. Haptically, they feel haptic feedback on the front of their arms or the back side of their arms. Visually, they see a picture of the vehicle on the screen leaning forward or backward.

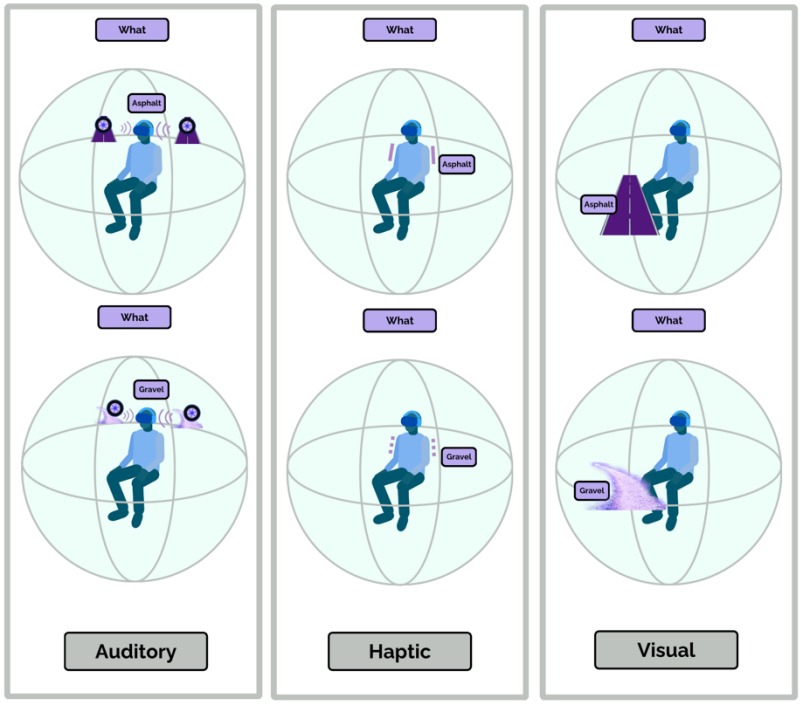

Figure 16 Task 3 "What?" cues, Asphalt or Gravel

For the "What" cue can be an asphalt road or a gravel road. Auditorily, the participant hears the sound of wheels drive on asphalt road or the sound of wheels driving on a gravel road. Haptically, they feel steady haptic feedback for an asphalt

road or bumpy interrupted feedback for a gravel road. Visually, they see an image of the asphalt road or the image of the gravel road on the cue display.

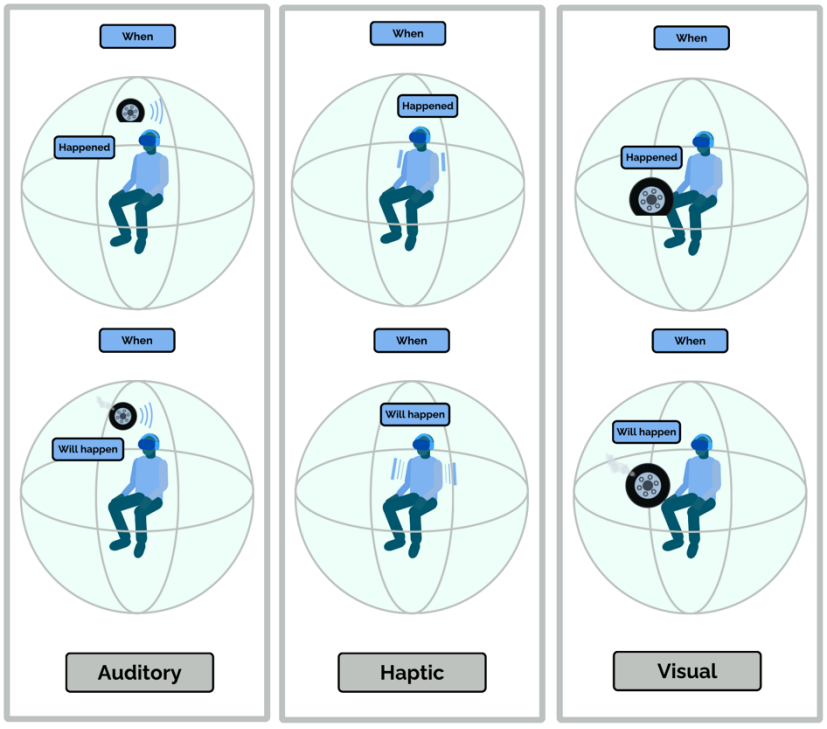

Figure 17 Task 3 "When?" cues, Happened or Will happen

For the "When" cue it is either a flat tire or a punctured tire. Auditorily, the participant hears the sound of a flat tire driving on the road or the hissing sound of air coming out of the tire. Haptically, they feel steady feedback for a flat tire or increasing feedback for deflating tire. Visually, they either see an image of a flat tire on the screen or an image of air coming out of a tire.

3.5 Measures

To answer our research question, we are only focusing on Accuracy and Reaction time that are associated with each information cue (Where, What, and When), whether in the one cue at a time sessions or in the combination of 3 different cues (training and final testing) sessions. For the one cue at a time sessions, we will be looking only at the first 4 events in the testing session, this is because if a participant performs perfectly, they will pass after doing only 4 events. In the training and final testing sessions we will be looking at all 16 task events for both. Figure 18 shows all data point for each task.

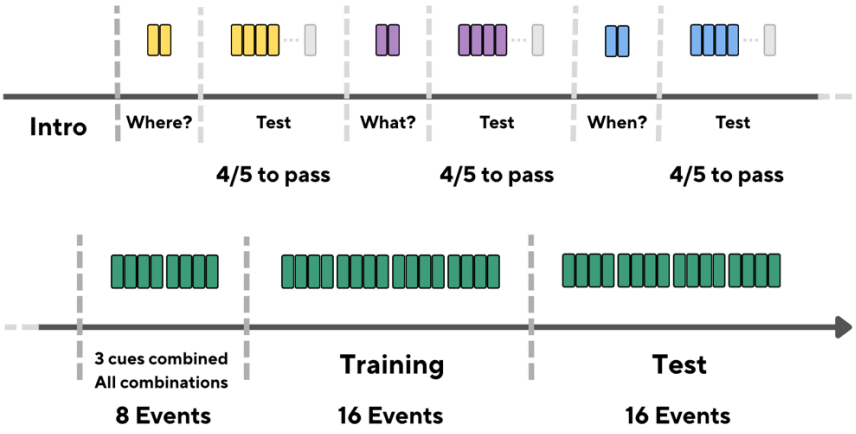

Figure 18 Task data points

3.5.1 Accuracy

For accuracy we are logging what action the participant has made and which cues were presented to them, when we present only one cue, several actions could be correct, but when they go through training and final testing, only one action would be correct, more on the actions in study setup section. Looking at the logs after the participant is done, we can know where exactly they made the mistake, meaning which information cue they could not comprehend. With that, we can give each information cue two accuracy scores, an accuracy score when it was presented alone and another when it was presented in combination with other cues. In addition, there is also the accuracy of answering the task events in training and final testing depending on which group the participant was assigned, but we will not get into it in this thesis.

3.5.2 Reaction time

Reaction time (RT) is the time between when the cue/cues were presented and when the participant made their action. We can get a reaction time for when we are presenting one cue at a time, this RT is how long it took them to comprehend that one cue and decide

which action to choose based on that cue alone, so it is only associated with that specific information cue (Figure 19 (a)). On the other hand, when the participant do the training and final testing sessions they will be presented with 3 cues at the same time, so the RT there include comprehending each cue and deciding which action to choose based on the 3 cues, with that being said, we are not able to disentangle how long each cue took to be comprehended individually (Figure 19 (b)). The two reaction times are illustrated in Figure 19.

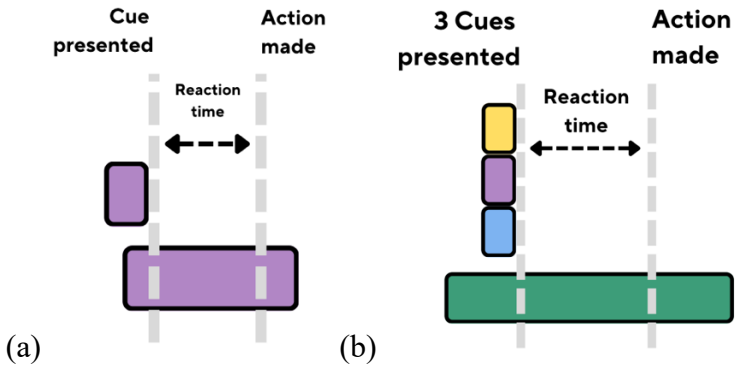

Figure 19 (a) Reaction time for individual cues (b) Reaction time for an Event

CHAPTER FOUR

SYSTEM ARCHITECTURE AND IMPLEMENTATION

This chapter describes the technical architecture underlying the experimental platform used in this study. While the three tasks differ in their contextual scenarios, cue content, and action spaces, they all share a common system architecture and event-driven execution pipeline. This unified design ensures procedural consistency across tasks and enables direct comparison of participant performance. The chapter first presents the technology stack, then describes the Unreal Engine map architecture that organizes the virtual environment, followed by the high-level task flow governing each session, and finally the event-level execution pipeline responsible for cue delivery, response capture, and data logging.

4.1 Technology Stack

The experimental platform was developed using Unreal Engine 5.1 (Epic Games), a real-time 3D creation engine widely adopted in both game development and academic XR research. Unreal Engine was selected for its high-fidelity rendering pipeline, robust physics simulation, and support for extended reality (XR) hardware.

Participants experienced the virtual environment through a Varjo XR-3 head-mounted display (HMD). The Varjo XR-3 was selected for its industry-leading visual fidelity and inside out camera head tracking. Auditory cues were delivered through Sony WH-1000XM5 headphones with Dolby Atmos 360 spatial audio. These headphones were chosen for their support of Dolby Atmos and active noise cancellation, which isolated participants from external laboratory sounds and ensured that auditory cues were perceived

in a controlled acoustic environment. The Dolby Atmos 360 spatial audio processing enabled true three-dimensional sound rendering, allowing directional cues to be perceived with accurate positional fidelity, sounds could be spatialized to originate from the left or right, above or below, and front or behind the participant. This capability was essential for delivering "Where" cues through the auditory channel, where the spatial origin of the sound encoded the cue's informational content. Haptic cues were delivered using the Teslasuit, a full-body haptic feedback suit that provides electrostimulation-based tactile feedback through an embedded actuator matrix. The Teslasuit was selected because it offers programmable, individually addressable haptic zones across the body, enabling the system to deliver spatially distinct and precisely timed haptic patterns. The suit was integrated with Unreal Engine via its SDK, allowing haptic events to be triggered in synchrony with visual and auditory cues at the engine level.

4.2 Unreal Engine Map Architecture

Figure 20 illustrates the map-level architecture of the Unreal Engine project. The application is organized into four distinct maps (levels): a central Start Map and three task-

specific maps, Vehicle Operation Task, Remote Drone Operation, and Vehicle Damage Assessment.

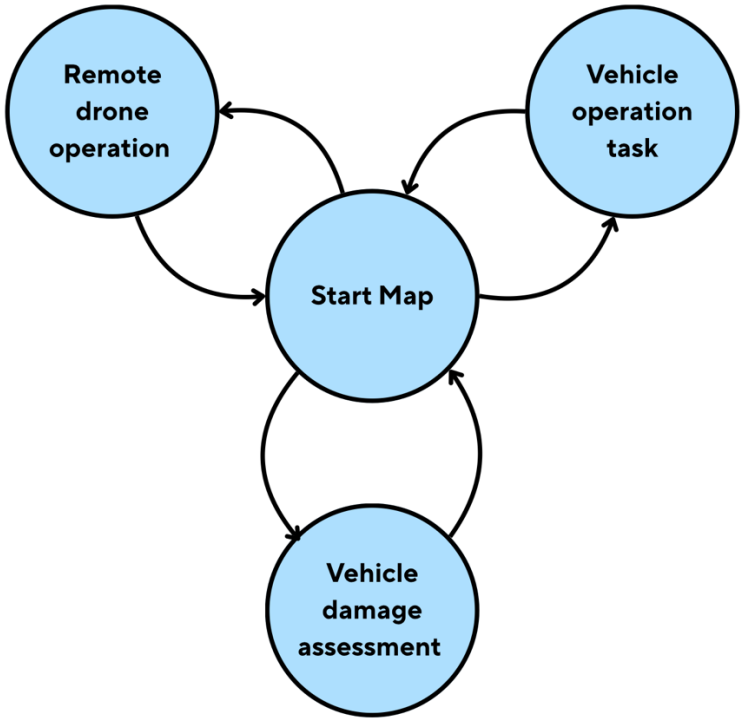

Figure 20 The map-level architecture of the Unreal Engine project

The Start Map serves two critical functions. First, it acts as the calibration environment where each participant begins their session. Upon entering the virtual environment, the participant's physical position and orientation are calibrated within the Start Map, and this calibrated position is saved and propagated to all subsequent task maps. This ensures that the participant's spatial reference frame remains consistent across all three tasks. Second, the Start Map functions as a transitional hub between tasks. After completing each task, the participant is returned to the Start Map before being loaded into

the next task map. This transition provides a natural break point between tasks, during which post-task questionnaires can be administered, and ensures that each task map is loaded from a clean, initialized state. It also allows us to play the tasks in any order we want.

Each of the three task maps contains the complete virtual environment for its respective scenario, including the 3D scene, task-specific cue assets, action space mappings, and event logic. Despite the differences in visual setting and narrative context, all three task maps share the same underlying Blueprint architecture for event sequencing, cue delivery, and data logging, as described in the following sections.

4.3 Task Flow Architecture

Figure 21 presents the high-level task flow architecture shared by all three experimental tasks. This flow governs the overall session structure from initialization through final data collection.

The task begins with an Init phase, during which the system loads the task-specific environment, configures the participant's assigned sensory group, and initializes all logging variables. This is followed by an Intro phase, where the participant receives an overview of the task scenario, goals, and context. Instructions are delivered both visually through an in-engine text banner and auditorily through the headphones.

The participant is then introduced to each of the three information cues (Where, What, and When) one at a time. For each cue, the system first presents two introductory events that familiarize the participant with the cue's two possible values (e.g., left vs. right for a "Where" cue), here the participant is able to test out the buttons and see if the selected action is a right response for that cue. This is followed by a Test phase consisting of an

undefined number of events where the criteria for passing is getting four out of five events right, if the participant makes a mistake their score resets.

Once all three cues have been individually introduced and tested, the participant enters the 3 Cues All Combinations phase. Here, all eight possible combinations of the three binary cues are presented, requiring the participant to make an action for each combination to proceed ensuring they experienced that combination, the participant moves backward or forward manually. This is followed by a Training session consisting of 16 events using all cue combinations twice, allowing participants to build familiarity with multi-cue integration under their assigned sensory channel configuration, the 16 events start automatically in this and the following session, if the participant did not make an action for 15s the next event starts. After training, a pause screen provides a brief break before the Testing session, which also consists of 16 events. With that the task ends and the participants are move to the start level map.

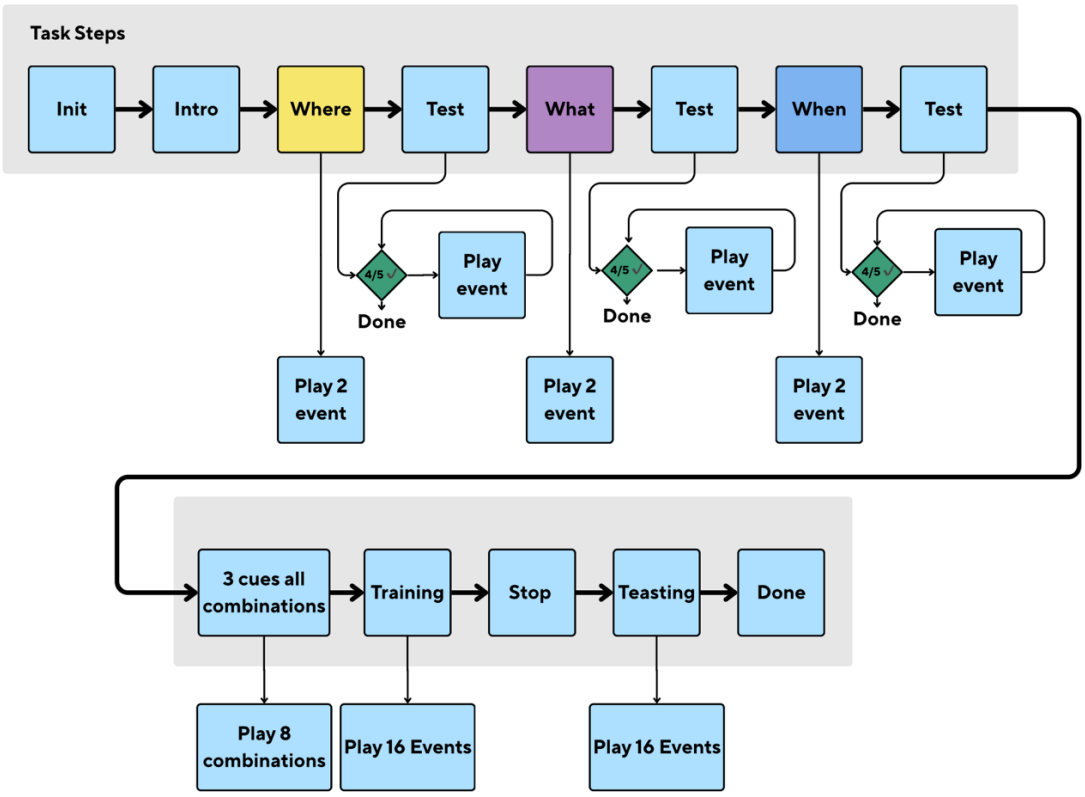

Figure 21 High-level task flow architecture

4.4 Event Execution Pipeline

Figure 22 presents the event-level execution pipeline that operates within each of the phases described above. Every event (whether part of the individual cue introduction, the combination phase, training, or testing) follows this same pipeline.

When an event is triggered, the system first records the presentation time and then delivers the appropriate cues based on the participant's assigned sensory group.

The system can deliver any of the cues without the other, that allows us to deliver one cue at the individual cue sessions and three cues together in the combined cues sessions, it’s worth mentioning here that when a cue is fired the execution of the program does not stop, the code continues allowing for simultaneous cue execution.

After cue delivery, the system enters a wait for input state. Upon the participant's response, the action and its time are recorded. The system then evaluates the participant's action against the expected correct response, and logs the event data, including presentation time, action time, presented cues, participant response. Finally, all event-related variables are reset before the next event begins, ensuring a clean state for each trial.

This event-driven pipeline provides several architectural advantages. First, the identical execution path across all events ensures that timing measurements (presentation time, reaction time) are captured consistently regardless of phase or task. Second, the modular cue delivery subsystem decouples the sensory channel assignment from the task logic, enabling the study's core experimental manipulation (varying which sensory channel delivers which cue) without modifying the underlying event structure. Third, the systematic logging at the event level produces a uniform and comprehensive dataset that supports the study's within-subject and between-group statistical analyses.

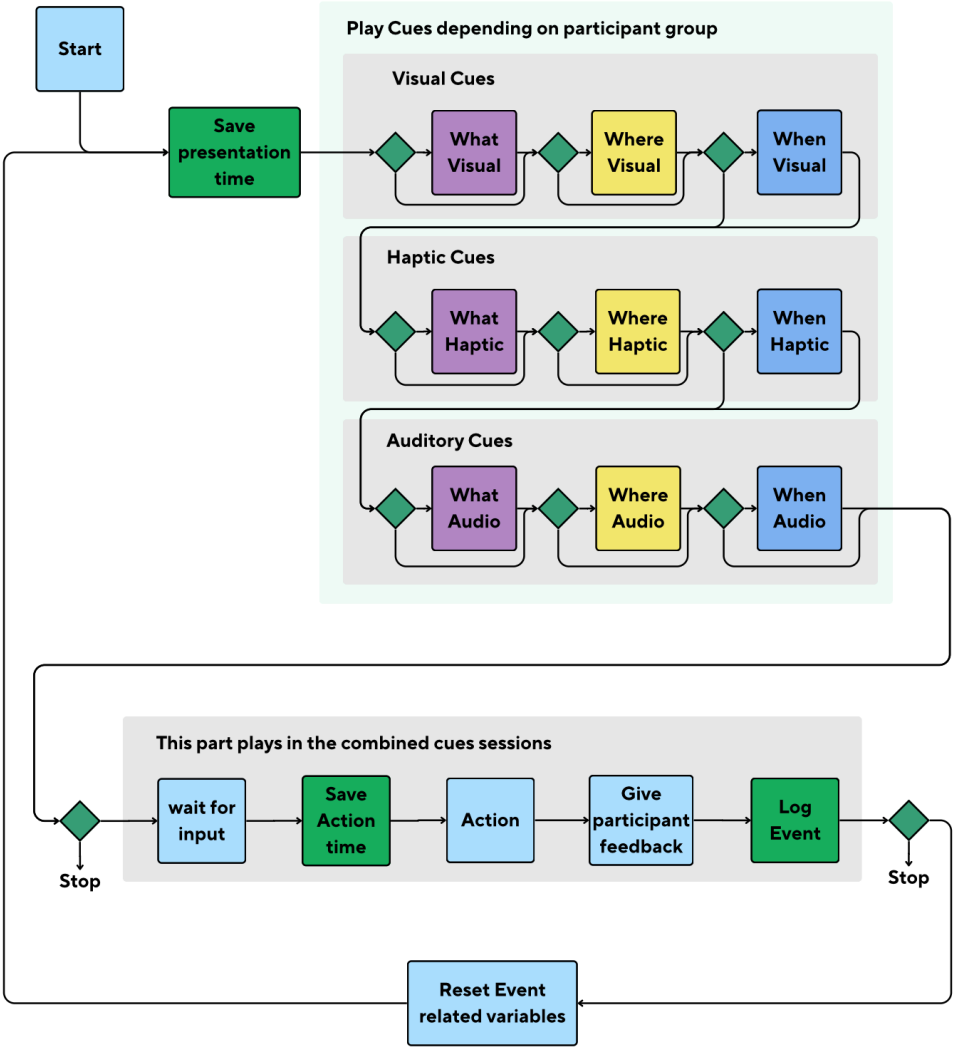

Figure 22 Event-level execution pipeline

CHAPTER FIVE

RESULTS AND FINDINGS

To answer our research question, we need to map the sensory modality to the information cues. We are doing a preliminary analysis on the results using 18 participants (2 per group), This configuration systematically provides N= 6 participants who engaged exclusively with each designated sensory modality across all cue typologies.

It is important to note the sequence in which accuracy values were obtained relative to participant training. For individual cue accuracy, values were derived from the initial single-cue testing phase, after the participants had been introduced to each cue's two possible states, they were tested by correctly identifying 4 out of 5 presentations. These values therefore reflect initial perceptual comprehension of each cue in isolation prior to any combined-cue exposure. The combined-cue training session accuracy reflects performance when participants first encountered all three cues simultaneously and received corrective feedback, while combined-cue testing session accuracy reflects performance after the training phase, without corrective feedback. This sequential structure allows for the examination of both baseline modality effectiveness and learning effects across phases.

We looked at the results in two ways. Firstly, scores for when each cue was presented in each task and modality, this gives us an idea of how well that specific cue performs. Secondly, accuracy scores were collapsed across all three tasks. This aggregation is beneficial because it reveals which sensory channel most reliably delivers each cue type regardless of the specific operational context, providing actionable guidance for

multimodal interface designers who need channel-to-cue mappings that are not task-dependent.

We will be highlighting the best and worst scores in each section, it is still worth it to consider the other results in the tables.

4.1 Accuracy

Individual Cues

We will start with the accuracy when the cues were presented alone, look at Table 4 and Table 5. The highest accuracy was 100% for when the "What" cues were presented visually in task 1, 2, and 3, for when the "Where" cues were presented visually in task 1 and 2, and for when the "Where" cues were presented haptically in task 1. The lowest accuracy was 66.67% and that was for when the "Where" cues were presented auditorily in task 3.

Table 4 Accuracy when the cue was presented alone

Cue Type	Task & Parameters			Visual	Audio	Haptic
What	Task 1	Supply	Civilian	100.00	87.50	83.33
	Task 2	Storm	LightRain	100.00	91.67	79.17
	Task 3	Asphalt	Gravel	100.00	91.67	95.83
Where	Task 1	Left	Right	100.00	95.83	100.00
	Task 2	Above	Below	100.00	75.00	75.00
	Task 3	Front	Behind	95.83	66.67	95.83
When	Task 1	Now	Delayed	95.83	83.33	75.00
	Task 2	FullBattery	LowBattery	83.33	95.83	91.67
	Task 3	Deflating	Flat	91.67	95.83	87.50

When looking at the result of the accuracy when the cues were presented alone, regardless of the task. The best accuracy was the "What" cues visually with 100% score,

and the worst accuracy was for the "Where" cues auditorily with a score of 79.17%. Overall, when we present one cue at a time, visual presentation performs best with 96.3% accuracy.

Table 5 Accuracy when the cue was presented alone, summary.

Modality	What	Where	When	Average
Visual	100.00	98.61	90.28	96.30
Audio	90.28	79.17	91.67	87.04
Haptic	86.11	90.28	84.72	87.04

Combined Cues (Training Session)

In the 3 cues combined training session, when the cue was presented with 2 other cues, Table 6 and Table 7. The highest accuracy was for when the "Where" cues were presented auditorily with a score of 98.96% in task 2. On the other hand, the lowest accuracy was for when the "Where" cues were presented haptically with a score of 72.92% in task 3.

Table 6 Accuracy when the cue was presented with 2 other cues (training session).

Cue Type	Task & Parameters			Visual	Audio	Haptic
What	Task 1	Supply	Civilian	97.92	91.67	89.58
	Task 2	Storm	LightRain	90.63	82.29	84.38
	Task 3	Asphalt	Gravel	95.83	89.58	88.54
Where	Task 1	Left	Right	93.75	86.46	81.25
	Task 2	Above	Below	96.88	98.96	83.33
	Task 3	Front	Behind	86.46	86.46	72.92
When	Task 1	Now	Delayed	95.83	77.08	83.33
	Task 2	FullBattery	LowBattery	91.67	86.46	84.38
	Task 3	Deflating	Flat	89.58	84.38	77.08

Looking at the results of when the cues were presented with 2 other cues in the training sessions regardless of the task. The best accuracy becomes for when the "What"

cues were presented visually with an accuracy of 94.79%. The lowest accuracy we got was for when the "Where" cues were presented haptically 79.17%. Overall, when 3 cues are presented together in the training session, the sensory channel with best accuracy was the visual channel with an accuracy of 93.17%.

Table 7 Accuracy when the cue was presented with 2 other cues (training session), summary.

Modality	What	Where	When	Average
Visual	94.79	92.36	92.36	93.17
Audio	87.85	90.63	82.64	87.04
Haptic	87.50	79.17	81.60	82.75

Combined Cues (Testing Session)

In the 3 cues combined testing session after the participants have learned the system for a little bit, when the cue was presented with 2 other cues, Table 8 and Table 9. The highest accuracy was for when the "Where" cues were presented auditorily with an accuracy of 100% in task 2. The lowest accuracy was for when the "When" cues were presented haptically with an accuracy of 70.83% in task 3.

Table 8 Accuracy when the cue was presented with 2 other cues (testing session).

Cue Type	Task & Parameters			Visual	Audio	Haptic
What	Task 1	Supply	Civilian	98.96	89.58	88.54
	Task 2	Storm	LightRain	97.92	80.21	78.13
	Task 3	Asphalt	Gravel	96.88	87.50	87.50
Where	Task 1	Left	Right	91.67	85.42	81.25
	Task 2	Above	Below	95.83	100.00	93.75
	Task 3	Front	Behind	90.63	89.58	75.00
When	Task 1	Now	Delayed	95.83	84.38	80.21
	Task 2	FullBattery	LowBattery	91.67	93.75	89.58
	Task 3	Deflating	Flat	91.67	79.17	70.83

When we look at the testing results regardless of the tasks, the best accuracy becomes for when the "What" cues were presented visually with a score of 97.92%. And the lowest accuracy score was for when the "When" cues were presented haptically with a score of 80.21%. The overall best sensory channel in the testing session was the visual channel with an accuracy score of 94.56%.

Table 9 Accuracy when the cue was presented with 2 other cues (testing session), summary.

Modality	What	Where	When	Average
Visual	97.92	92.71	93.06	94.56
Audio	85.76	91.67	85.76	87.73
Haptic	84.72	83.33	80.21	82.75

Graph: Information Cues vs. Sensory Modality Accuracy

Figure 23 presents the accuracy distributions for each information cue (What, Where, When) across the three sensory modalities (Visual, Auditory, Haptic). For the "What" cues, the visual modality exhibited the highest and most consistent accuracy, with scores clustering between 95.83% and 100%. The auditory modality showed a slightly wider spread (82.29%-91.67%), while the haptic modality had the lowest median and the greatest variability, ranging from 78.13% to 95.83%. For the "Where" cues, the visual modality again achieved the highest scores (88.46%-100%), while the auditory modality demonstrated a competitive performance with scores reaching up to 100% in certain tasks. Notably, the haptic modality showed the widest interquartile range for "Where" cues, with scores spanning from 66.67% to 100%, indicating inconsistent participant performance. The "When" cues proved most challenging across all modalities. While visual scores remained relatively high (77.08%-95.83%), both the auditory (77.08%-95.83%) and haptic

(70.83%-91.67%) modalities showed greater downward spread, reflecting the inherent difficulty of conveying temporal information through non-visual sensory channels.

Figure 23 Information Cues vs. Sensory Modality Accuracy

Graph: Accuracy by Modality Across Sessions

Figure 24 presents the accuracy for each combination of sensory modality (Visual, Auditory, Haptic), information cue (What, Where, When), and session type (Individual, Training, Testing). Across all session types, the visual modality maintained the highest accuracy, with "What" cues achieving 100% in individual presentation and remaining above 94.8% through training and 97.9% in testing. The visual "Where" cues showed a similar pattern, starting at 98.6% in individual sessions and stabilizing at 92.7% in testing. The auditory modality showed a notable strength in "Where" cues, achieving 91.7% accuracy in both training and testing, approaching visual-level performance (92.7%) on the same cue dimension. However, auditory accuracy for "What" and "When" cues was lower, particularly in testing where "When" cues dropped to 85.8%. The haptic modality consistently recorded the lowest accuracy across all cue types and sessions. While haptic

"What" cues reached 90.3% in individual presentation, accuracy declined progressively through training (87.5%) and testing (84.7%). Haptic "When" cues showed the weakest performance overall, dropping from 83.3% in individual presentation to 80.2% in testing. Importantly, the visual and auditory modalities showed stable or slightly improving accuracy from training to testing, suggesting effective learning transfer, whereas the haptic modality exhibited a consistent decline across sessions.

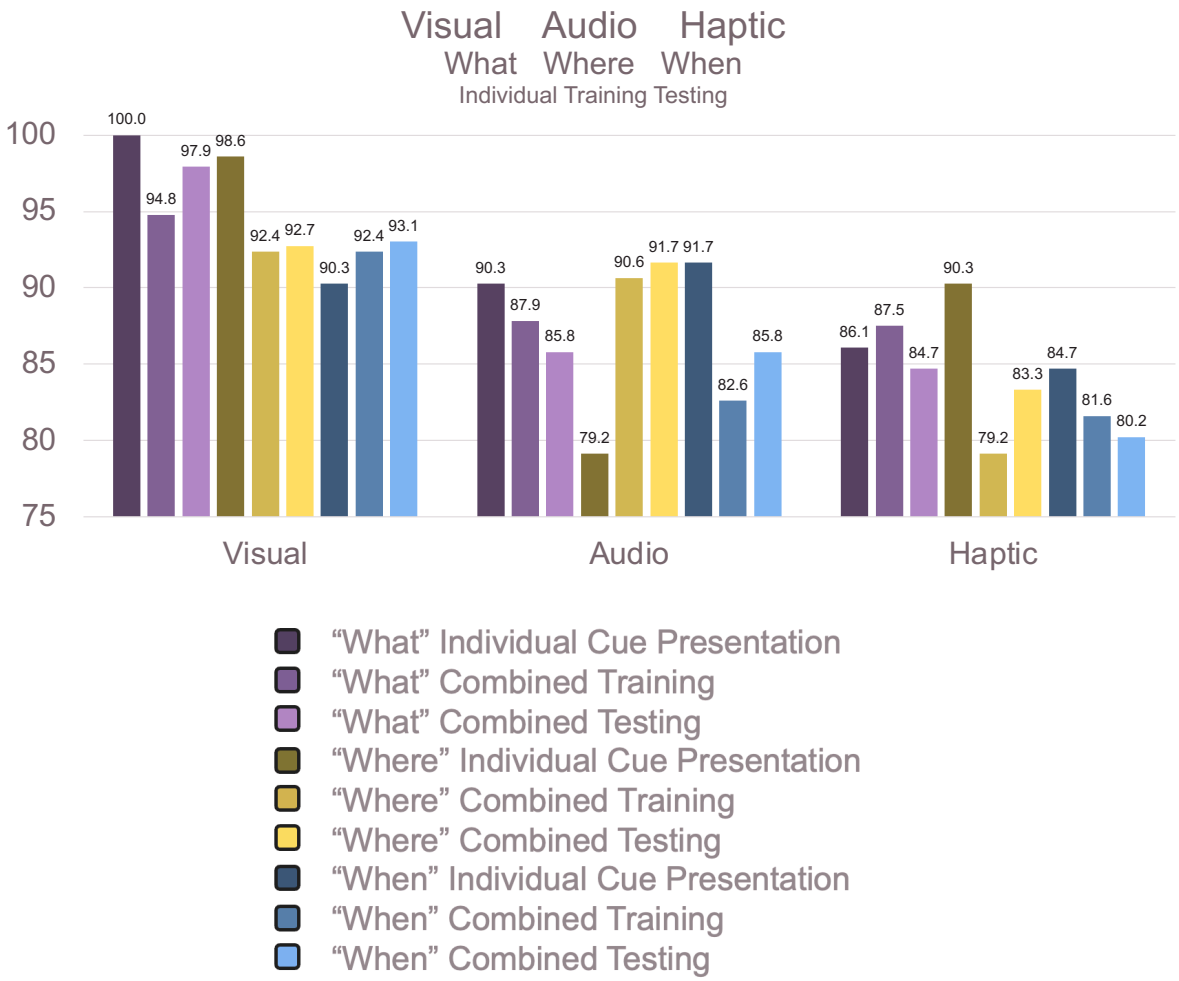

Figure 24 Accuracy by Modality Across Sessions

4.2 Reaction time

We got reaction time for when the cue was presented alone, Table 10 and Table 11. But we cannot get reaction times for when the information cue was presented with other cues, because the time recorded includes comprehending all three cues. The fastest reaction time was for when the "Where" cues were presented haptically in task 1 with a score of 0.77 seconds. The slowest reaction time we got was for when the "When" cues were presented auditorily with a score of 3.04 seconds.

Table 10 Accuracy and Reaction time when the cue was presented alone.

Cue Type	Task & Parameters			Visual		Audio		Haptic	
				RT (s)	Accuracy	RT (s)	Accuracy	RT (s)	Accuracy
What	Task 1	Supply	Civilian	1.40	100.00	2.11	87.50	1.36	83.33
	Task 2	Storm	LightRain	0.96	100.00	1.56	91.67	1.53	79.17
	Task 3	Asphalt	Gravel	1.07	100.00	1.85	91.67	1.85	95.83
Where	Task 1	Left	Right	0.92	100.00	1.12	95.83	0.77	100.00
	Task 2	Above	Below	0.83	100.00	2.61	75.00	1.97	75.00
	Task 3	Front	Behind	0.88	95.83	2.43	66.67	1.06	95.83
When	Task 1	Now	Delayed	1.65	95.83	3.04	83.33	1.64	75.00
	Task 2	FullBattery	LowBattery	1.97	83.33	1.50	95.83	1.84	91.67
	Task 3	Deflating	Flat	1.21	91.67	1.68	95.83	1.59	87.50

When looking at the reaction times regardless of the task, the fastest reaction time becomes when the "Where" cues were presented visually with a score of 0.88 seconds. When looking at the scores overall regardless of information cue type, the best reaction time was 1.21 seconds for the visual channel.

Table 11 Accuracy and Reaction time when the cue was presented alone, summary.

Modality	What	Where	When	Average

Visual	**Accuracy**	100.00	98.61	90.28	96.30
	RT (s)	1.14	0.88	1.61	1.21
Audio	**Accuracy**	90.28	79.17	91.67	87.04
	RT (s)	1.84	2.05	2.07	1.99
Haptic	**Accuracy**	86.11	90.28	84.72	87.04
	RT (s)	1.58	1.26	1.69	1.51

Graph: Reaction Time for Individual Cue Presentation

Figure 25 presents the reaction time distributions for individual cue presentation across the three sensory modalities (Visual, Haptic, Auditory) and three information cues (What, Where, When). For the "What" cues, the visual modality produced the fastest and most tightly clustered reaction times (median ≈ 1.1s), followed by the haptic modality (median ≈ 1.5s) and the auditory modality (median ≈ 1.9s) which showed the widest spread. For the "Where" cues, the visual modality achieved the fastest overall reaction times in the study (median ≈ 0.9s) with minimal variance, while the haptic modality showed a moderate range (median ≈ 1.3s). The auditory "Where" cues had the slowest and most variable reaction times (median ≈ 2.5s), with considerable spread in response latency. For the "When" cues, reaction times converged across modalities more than in other cue types, with visual (median ≈ 1.6s), haptic (median ≈ 1.6s), and auditory (median ≈ 2.0s) showing overlapping distributions. This convergence suggests that temporal information is inherently more demanding to process regardless of the delivery channel. Across all cue types, the visual modality consistently yielded the fastest reaction times, averaging 1.21 seconds overall, while the auditory modality was the slowest, particularly for spatially encoded "Where" cues.

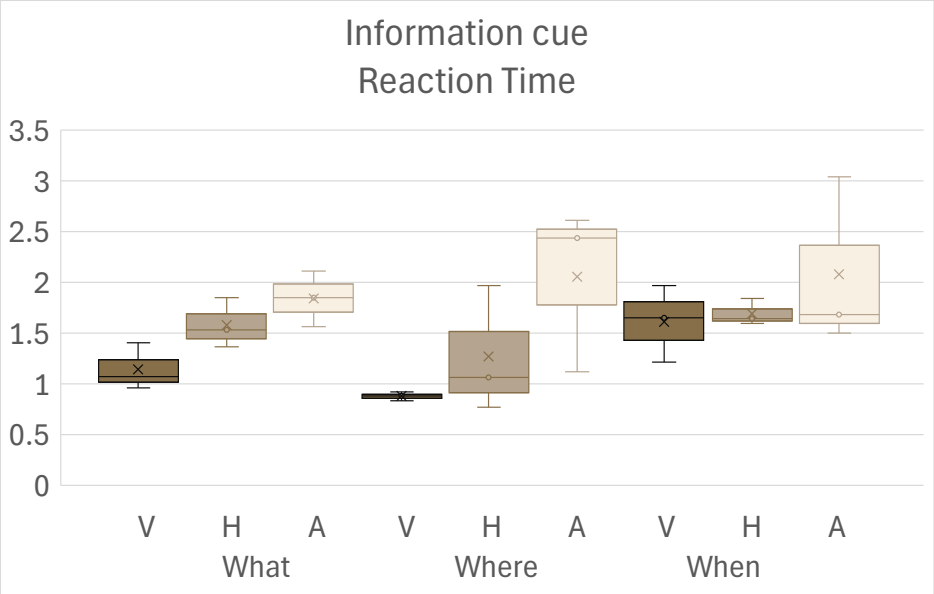

Figure 25 Reaction Time for Individual Cue Presentation

4.3 Summary

The preliminary results reveal a clear performance hierarchy across the three sensory modalities. The visual modality consistently outperformed the other channels, achieving the highest average accuracy of 94.7% and the fastest average reaction time of 1.21 seconds across all cue types and session phases. This dominance was most pronounced for "What" cues, where visual accuracy reached 100% in individual presentation and 97.9% in the final testing session.

The auditory modality ranked second in overall accuracy, with a notable strength in spatial information: auditory "Where" cues achieved 91.7% accuracy in the testing session, approaching visual-level performance (92.7%) on the same cue dimension. This competitive performance on "Where" cues suggests that spatialized audio can serve as an effective alternative to visual presentation for conveying directional information.

The haptic modality consistently recorded the lowest accuracy of the three channels and showed a declining trajectory across sessions, particularly for "When" cues, which dropped to as low as 80.2% in the testing session. This suggests that participants had greater difficulty learning and retaining temporal information delivered through haptic stimulation.

The overall performance ranking for accuracy was Visual > Auditory > Haptic. Interestingly, the reaction time ranking followed a different order: Visual < Haptic < Auditory, with the visual modality being the fastest (1.21s average), followed by haptic (1.51s), and auditory being the slowest (1.99s). This inversion (where auditory was more accurate than haptic but slower in reaction time) suggests a potential speed-accuracy trade-off in auditory cue processing, where participants may have taken longer to respond but did so with greater precision than when using haptic feedback.

Across all modalities, "When" cues proved the most challenging, consistently yielding the lowest accuracy scores and showing the least differentiation in reaction times across channels. This finding indicates that temporal information is inherently more difficult to encode and perceive through any single sensory modality, highlighting a potential design constraint for multimodal interface systems that must communicate time-sensitive information.

Figure 23, Figure 24, and Figure 25 visually summarize these patterns and complement the detailed per-task results presented in Tables 4 through 11.

CHAPTER SIX

DISCUSSION

The primary objective of this thesis was to determine if sensory modality (Visual, Haptic, Auditory) had an effect on the accuracy of comprehending information cues (What, Where, When) across different tasks. The preliminary results demonstrate that the visual presentation of the cues had a higher accuracy average and a faster reaction time average than the auditory and haptic presentations. When looking at the results through the lens of Multiple Resource Theory [2], the visual dominance could be due to the nature of human beings as the human cerebral cortex mathematically dedicates a massive percentage of its neural processing bandwidth exclusively to visual calculation and object tracking. The visual dominance could be also due to the visual cues being more familiar or intuitive to the participants. So, if a designer wants to deliver an information cue and maintain high level 1 SA [1], the best way to deliver it is through the visual sensory channel.

But what if the cue was not in the direct field of view of the user, or what if the visual sense of the user was busy doing a primary task and they had another secondary task they were also doing. Then the designer must consider using the other sensory channels, haptic and auditory. We can see from the results that the "Where" cues auditory performed much better than haptically. Theoretically, humans natively utilize auditory latency vectors (time-of-arrival differentials between ears) to pinpoint geospatial anomalies, making spatial mapping transfer exceptionally well to auditory delivery architectures [14]. That could mean that if you had an information cue of that type, the best way to deliver it after visually is auditorily. By offloading spatial telemetry parameters to localized spatial audio,

engineers can leverage this evolutionary heuristic, immediately granting the operator omnidirectional spatial awareness without compounding visual load. When looking at the "What" and "When" information cues, both auditory and haptic sensory modalities performed similarly giving the designer the choice to use either for information delivery.

The comparison between training and testing sessions provides evidence for a learning effect in multimodal cue comprehension. For the visual and auditory modalities, accuracy remained stable or improved slightly from training to testing (visual: 93.2% to 94.6%; auditory: 87.0% to 87.7%). The haptic modality showed stable but unchanging accuracy (82.75% in both sessions).

These findings have direct implications for the design of future remote drone operation systems for example. Imagine a system where the operator's primary visual attention is focused on the drone's live camera feed. Rather than adding more visual overlays that could cause information overload, designers could offload spatial cues, such as the location of nearby obstacles or threats, to a spatial audio system that leverages the operator's natural ability to localize sound. Meanwhile, haptic feedback through a controller or wearable device could deliver timing-related alerts. Such a system would distribute information across sensory channels based on empirically supported mappings, maximizing situation awareness while minimizing cognitive overload.

Limitations

While these preliminary findings are compelling, there are limitations to acknowledge. The current analysis is based on a sample size of 18 participants; a full analysis of the intended 108 participants will be required to confirm these trends with

statistical significance. Additionally, because the study was designed to compare only how the different modality combinations affect the accuracy and reaction times of participants and was not designed to study how each modality affected each information cue, we could not eliminate the learning effect from the individual cues presentation since it was the first thing the participant did. In addition, we could not get a reaction time for each information cue when 3 cues were presented simultaneously.

Finally, this current analysis is restricted to performance metrics (accuracy and reaction time). The study also collected data on perceived workload, system usability, and presence, which were not analyzed in this preliminary report. These subjective measures may reveal important differences between modality conditions that are not captured by performance data alone.

CHAPTER SEVEN

CONCLUSION AND FUTURE WORK

6.1 Conclusion

The thesis investigated a fundamental question in multimodal human-machine interface design: how the mapping of information cues (Where, What, When) to sensory channels (Visual, Auditory, Haptic) impacts user accuracy in XR environments. That was achieved through a preliminary analysis on data from a novel 3 x 3 x 3 human subject research study designed to test how different information to sensory mapping combinations affect user performance. By testing these mappings across three distinct tasks, vehicle operation, remote drone operation, and vehicle damage assessment, each requiring participants to perceive, interpret, and act upon three simultaneously delivered information cues.

The preliminary findings of 18 participants provided great insights that helps in designing systems with task critical information cues:

First, the visual modality demonstrated consistent superiority across all information cue types and experimental conditions. Visual cue delivery achieved the highest overall accuracy (96.3% for individual cues, 93.2% in combined-cue training, 94.6% in combined-cue testing) and the fastest average reaction time (1.21 seconds). This result confirms the well-established visual dominance effect in human perception and underscores the reliability of the visual channel as the primary pathway for delivering task-critical information when it is available and not saturated by competing demands.

Second, the nature of the cue type affects its comprehension accuracy, for example the “What” cue proved to be the most robust and easiest to comprehend across all sensory channels, unlike the “Where” cues that proved to perform better auditory comparing to haptically.

Third, when alternative modalities are required, for instance, when the visual channel is occupied by a primary task, the auditory modality emerged as a strong candidate for spatial (Where) information, achieving accuracy levels comparable to visual delivery (91.67% vs. 92.71% in combined-cues testing). This advantage is attributable to the human auditory system's innate spatial processing capabilities.

Fourth, the data confirmed the presence of a learning effect, with accuracy generally improving from training to testing sessions for the visual and auditory modalities. This finding affirms the importance of structured training in multimodal interface systems and suggests that dedicated onboarding protocols could improve operator performance, particularly for the less intuitive haptic and auditory channels.

Collectively, these findings tell us that effective multimodal interface design is not simply achieved by distributing information cues randomly across all available sensory channels. Instead, designers must strategically map information cues to the best available sensory channels depending on cue type.

6.2 Contributions

This thesis makes the following contributions to the fields of human-computer interaction and human factors engineering:

1. Development of Three Virtual Task Environments in Unreal Engine 5.1: The primary technical contribution of this thesis is the design and full implementation of three distinct virtual reality task environments (a vehicle operation scenario, a remote drone operation scenario, and a vehicle damage assessment scenario)using Unreal Engine 5.1. This development included creating all 3D environments, designing and implementing the visual, auditory, and haptic cue delivery systems, building the event sequencing and randomization logic, and developing the real-time data logging infrastructure. These environments are fully functional experimental platforms capable of running human-subject studies with no additional development.
2. Contributing to Information Cue and Task Design: This thesis contributes to the theoretical design of the information cue framework, specifically the process of translating the abstract What-Where-When cue categories into concrete, perceivable stimuli across three modalities. This includes the design of stackable cue sets that can be delivered simultaneously on a single sensory channel without masking, and the systematic variation of spatial dimensions across tasks (lateral, vertical, longitudinal) to ensure generalizability of findings.
3. Experimental Framework Development: The development of a novel 3 × 3 × 3 experimental paradigm that systematically crosses information cue types with sensory modalities across multiple task contexts, providing a replicable methodology for investigating information-to-modality mappings. This framework, including the counterbalanced task ordering, the progressive training protocol (individual cues → combined training with feedback → combined testing

without feedback), and the binary cue design, constitutes a methodological contribution that other researchers can adopt for similar multimodal studies.

4. Preliminary Empirical Evidence for Modality-Cue Interactions: Initial data from 18 participants demonstrating that modality effectiveness is not uniform across information types, with spatial ("Where") information showing particular compatibility with auditory delivery and categorical ("What") information showing the strongest resilience across all modalities. These preliminary findings provide the first empirical data points from this experimental paradigm and lay the groundwork for the full N = 108 analysis.
5. Data Collection: I contributed to running participants through the experimental protocol, including hardware setup, Teslasuit calibration, and session management, enabling the collection of the behavioral data analyzed in this thesis.

6.3 Future work

Complete the statistical analysis for the total N = 108 participant. Additionally, looking into the findings in relation to other collected data including workload, presence, system usability and see if there are some correlations between them.

Execute a structured qualitative analysis on the interview questions asked at the end of each task providing a different Lens to the one quantitative data provides. Future work could explore individual cue presentation without having other cues on the other channels by making a study focused on individual cue presentation using different sensory channels.

REFERENCES

[1] Endsley, Mica. (1995). Endsley, M.R.: Toward a Theory of Situation Awareness in Dynamic Systems. Human Factors Journal 37(1), 32-64. Human Factors: The Journal of the Human Factors and Ergonomics Society. 37. 32-64. 10.1518/001872095779049543.

[2] Wickens, Christopher. (2008). Multiple Resources and Mental Workload. Human factors. 50. 449-55. 10.1518/001872008X288394.

[3] Elor, A., Kurniawan, S., & Teodorescu, M. (2022). Towards Multimodal Affordances of Extended Reality for Inclusive Learning. IEEE Transactions on Learning Technologies, 15(3), 342–355. doi:10.1109/TLT.2022.3144357

[4] Bresciani, J.-P., Dammeier, F., & Ernst, M. O. (2006). Vision and touch are automatically integrated for the perception of sequences of events. Journal of Vision, 6(5), 554–564. doi:10.1167/6.5.2

[5] Sheridan, T. B. (2016). Human–Robot Interaction: Status and Challenges. Human Factors, 58(4), 525–532. doi:10.1177/0018720816644364

[6] Van Erp, J. B. F., & Van Veen, H. A. H. C. (2004). Vibrotactile in-vehicle navigation system. Transportation Research Part F: Traffic Psychology and Behaviour, 7(4–5), 247–256. doi:10.1016/j.trf.2004.09.003

[7] Lu, S. A., Wickens, C. D., Prinet, J. C., Hutchins, S. D., Sarter, N., & Sebok, A. (2013). Supporting interruption management and multimodal interface design: Three

meta-analyses of task performance as a function of interrupting task modality. Human Factors, 55(4), 697–724. doi:10.1177/0018720813476298

[8] Grassini, S., & Laumann, K. (2020). Questionnaire Measures and Physiological Correlates of Presence: A Systematic Review. Frontiers in Psychology, 11, 349. doi:10.3389/fpsyg.2020.00349

[9] Slater, M. (2009). Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. Philosophical Transactions of the Royal Society B: Biological Sciences, 364(1535), 3549–3557. doi:10.1098/rstb.2009.0138

[10] Cummings, M. L., & Mitchell, P. J. (2008). Predicting controller capacity in supervisory control of multiple UAVs. IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans, 38(2), 451–460. doi:10.1109/TSMCA.2007.914757

[11] Shu, Y., & Furuta, K. (2005). An inference method of team situation awareness based on mutual awareness. Cognition, Technology & Work, 7(4), 272–287. doi:10.1007/s10111-005-0012-x

[12] Wickens, C. D. (2002). Multiple resources and performance prediction. Theoretical Issues in Ergonomics Science, 3(2), 159–177. doi:10.1080/14639220210123806

[13] Politis, I., Brewster, S., & Pollick, F. (2015). To Beep or Not to Beep? Comparing Abstract versus Language-Based Multimodal Driver Displays. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15), 3971–3980. ACM. doi:10.1145/2702123.2702167

[14] Carlini A, Bordeau C and Ambard M (2024) Auditory localization: a comprehensive practical review. Front. Psychol. 15:1408073. doi: 10.3389/fpsyg.2024.1408073

[15] Towards Size of Scene in Auditory Scene Analysis: A Systematic Review J Audiol Otol. 2020;24(1):1-9. Published online November 22, 2019 DOI: https://doi.org/10.7874/jao.2019.00248

[16] Thomas Muender, Michael Bonfert, Anke Verena Reinschluessel, Rainer Malaka, and Tanja Döring. 2022. Haptic Fidelity Framework: Defining the Factors of Realistic Haptic Feedback for Virtual Reality. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22). Association for Computing Machinery, New York, NY, USA, Article 431, 1–17. https://doi.org/10.1145/3491102.3501953