AUDIO-VISUAL SALIENCY

IN OMNIDIRECTIONAL PANORAMIC SCENES

HALİT ÖZSOY

BOĞAZİÇİ UNIVERSITY

2024

AUDIO-VISUAL SALIENCY

IN OMNIDIRECTIONAL PANORAMIC SCENES

Thesis submitted to the

Institute for Graduate Studies in Social Sciences

in partial fulfillment of the requirements for the degree of

Master of Arts

in

Cognitive Science

by

Halit Özsoy

Boğaziçi University

2024

Audio-Visual Saliency in Omnidirectional Panoramic Scenes

The thesis of Halit Özsoy

has been approved by:

Assoc. Prof. İnci Ayhan
(Thesis Advisor) _____

Assoc. Prof. İbrahim Aykut Erdem
(Thesis Co-Advisor) _____

Assoc. Prof. Emre Uğur _____

Assoc. Prof. Güneş Ünal _____

Assist. Prof. Nihan Alp
(External Member) _____

May 2024

DECLARATION OF ORIGINALITY

I, Halit Özsoy, certify that

- I am the sole author of this thesis and that I have fully acknowledged and documented in my thesis all sources of ideas and words, including digital resources, which have been produced or published by another person or institution;
- this thesis contains no material that has been submitted or accepted for a degree or diploma in any other educational institution;
- this is a true copy of the thesis approved by my advisor and thesis committee at Boğaziçi University, including final revisions required by them.

Signature ……

Date ……

ABSTRACT

Audio-Visual Saliency in Omnidirectional Panoramic Scenes

As virtual reality devices become more widespread, investigating how humans explore omnidirectional panoramic scenes has become a topic of interest. Several computational models now exist for predicting visual saliency in such environments. However, the existing literature largely ignores auditory cues while estimating saliency maps. To pave the road for more capable models, we investigate how scene attributes and sound categories influence visual exploration. In that manner, how several semantic categories for audio and places influence observer saliency and how such effects differ across different color and audio modulations are investigated by conducting an experiment and creating a novel eye-tracking dataset. Analyses revealed that when viewing 360° videos freely, observer exploration behavior completely changed across different audio categories. The presence of audio, irrespective of whether it contained spatial features, strongly guided this exploration behavior. Color information did not necessarily have a similar effect. Furthermore, the effects of task type over eye fixations were explored by conducting a psychophysics experiment via presenting participants with simple three-dimensional audio-visual stimuli across free viewing, search, and memory recall tasks where the stimuli differed by colors. Overall, analyses revealed that neither of the modulated features had any effect except color contrast levels, hinting that only the task-relevant features play a role in eye fixations in the presence of a task.

ÖZET

Çok Yönlü Panoramik Sahnelerde Görsel-İşitsel Belirginlik

Sanal gerçeklik cihazlarının daha yaygın hale gelmesiyle, insanların panoramik sahneleri nasıl keşfettiklerini araştırmak ilgi çekici bir konu haline gelmiştir. Bu tür ortamlarda görsel dikkati tahmin etmek için çeşitli hesaplamalı modeller mevcuttur. Ancak, mevcut literatür, belirginlik haritalarını tahmin ederken işitsel ipuçlarını büyük ölçüde göz ardı etmektedir. Daha yetkin modellerin yolunu açmak için, sahne özelliklerinin ve ses kategorilerinin görsel keşfi nasıl etkilediğini araştırdık. Bu bağlamda, çeşitli ses ve mekanlara ait anlamsal kategorilerin gözlemci dikkatini nasıl etkilediğini ve bu tür etkilerin farklı renk ve ses modülasyonları arasında nasıl farklılaştığını bir deney yaparak ve yeni bir göz hareketi izleme veri seti oluşturarak araştırdık. Analizler, gözlemcinin 360° videoları serbestçe izlerkenki keşif davranışının, farklı ses kategorileri arasında tamamen değiştiğini ortaya koydu. Sesin mekansal özellikler içerip içermediğine bakılmaksızın, sesin varlığı bu keşif davranışını güçlü bir şekilde yönlendirdi. Renk içeriği ise benzer bir etkiye sahip değildi. Ayrıca, katılımcılara serbest izleme, arama ve hafıza geri çağırma görevlerinde farklı renklerde basit üç boyutlu görsel-işitsel uyaranlar göstererek gerçekleştirilen bir psikofizik deneyi ile farklı görev türlerinin göz sabitlemeleri üzerindeki etkileri araştırıldı. Analiz sonuçları, renk kontrast seviyeleri dışında modüle edilen özelliklerin hiçbirinin bir etkisi olmadığını, dolayısıyla bir görev mevcut olduğunda yalnızca görevle ilişkili özelliklerin göz sabitlemleri üzerinde rol oynadığına işaret etmektedir.

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my supervisor, İnci Ayhan, and my co-supervisor, Aykut Erdem, for their invaluable guidance, support, and encouragement throughout this project. Their expertise and insights were instrumental in the completion of this work. I also wish to thank Erkut Erdem for their additional instruction and assistance, which greatly contributed to my understanding and progress.

I am also thankful to Mert Çökelek for collaborating with me on the research topic and co-authoring a paper. Their contributions and insights were invaluable.

I am immensely grateful to the members of Visionlab and Buviarlab for their collaborative spirit and support. In particular, I would like to acknowledge Batuhan Koçak for his significant help with the analyses, which was crucial to the success of this project. Additionally, I would like to thank my undergraduate assistants, Sevban Cander, Zeynep Çekinmez, and Çağlar Kurt.

This work was supported by The Scientific and Technological Research Council of Turkey (TUBITAK) ARDEB 1001 [Project number: 120E501] program.

DEDICATION

To all my friends, family, instructors and my university. Each of you played a significant role in shaping the way I am and will be.

Thank you for being part of my journey.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION 1

1.1 Saliency estimation models 3

1.2 Exploratory research on saliency 7

1.3 The scope and the purpose of this thesis 8

CHAPTER 2: PRELIMINARY EXPERIMENT 10

2.1 Objectives 10

2.2 Methodology 10

2.3 Results 16

2.4 Discussions and conclusions 24

CHAPTER 3: 360° VIDEO WATCHING EXPERIMENT 26

3.1 Research questions 26

3.2 Methodology 28

3.3 Data processing 41

3.4 Results 52

3.5 Discussion 59

CHAPTER 4: LOW LEVEL PSYCHOPHYSICS EXPERIMENT 68

4.1 Research questions 69

4.2 Experimental design 69

4.3 Data collection 80

4.4 Results 82

4.5 Discussion 107

CHAPTER 5: CONCLUDING REMARKS 110

APPENDIX A: VIDEO CLIPS IN ASOD60K DATASET 113

APPENDIX B: VIDEO FRAME MATCHING METHOD 114
APPENDIX C: SIMILARITY METRIC CALCULATIONS 115
APPENDIX D: CUSTOM VIDEO ANNOTATION WEBSITE TOOL SCREENSHOTS 116
APPENDIX E: CURATED VIDEO CLIPS FOR NEW YT360-EYE-TRACKING DATASET 122
APPENDIX F: QUESTIONNAIRE FORM (ENGLISH) 124
APPENDIX G: QUESTIONNAIRE FORM (TURKISH) 125
APPENDIX H: ETHICS COMMITTEE APPROVAL FOR 360° VIDEO EXPERIMENT 126
APPENDIX I: EXPERIMENT SETUP AND ENVIRONMENT 127
APPENDIX J: EXPERIMENTER PROMPT FOR 360° VIDEO EXPERIMENT (ENGLISH) 128
APPENDIX K: EXPERIMENTER PROMPT FOR 360° VIDEO EXPERIMENT (TURKISH) 129
APPENDIX L: VIDEO THUMBNAILS WITH HEATMAP OVERLAY OF FIXATION MAPS 130
APPENDIX M: ADDITIONAL PLOTS OF CONSISTENCY SCORES 139
APPENDIX N: ETHICS COMMITTEE APPROVAL FOR PSYCHOPHYSICS EXPERIMENT 144
REFERENCES 145

LIST OF TABLES

Table 1. Top 6 VGGSound Labels of the Sample Video Clip 12

Table 2. Top 6 AudioSet Labels of the Sample Video Clip 13

Table 3. Top 6 Places365 Labels of the Sample Video Clip 15

Table 4. Similarity Scores of the Sample Video Clip 16

Table 5. Principal Component Analysis Loadings Table 48

Table 6. Exploratory Factor Analysis Loadings and Communality Values 50

Table 7. Confirmatory Factor Analysis Loadings and R^2 Values 51

LIST OF FIGURES

Figure 1. Histogram of VGGSound high-level audio-type categories across videos 12

Figure 2. Histogram of AudioSet high-level audio-type categories across videos ... 13

Figure 3. A collage of all 15 extracted view shots from a single frame 14

Figure 4. Histogram of Places365 high-level place-type categories across videos .. 15

Figure 5. Human eye fixation density map visualizations on the sample video 17

Figure 6. Boxplots of similarity scores for Places365 place categories..... 20

Figure 7. Boxplots of similarity scores for VGGSound audio categories..... 22

Figure 8. Boxplots of similarity scores for AudioSet audio categories 24

Figure 9. Video clips in YT360-EyeTracking by categories 32

Figure 10. A researcher testing the experiment setup via watching a demo video ... 36

Figure 11. Outline of the experiment blocks and trials..... 40

Figure 12. Distribution of all raw and standardized consistency scores over time ... 46

Figure 13. Correlation scores between similarity metrics for consistency 48

Figure 14. Scree plot of the principal component analysis on consistency scores ... 49

Figure 15. Ambisonics vs mute fixation similarities across place categories..... 52

Figure 16. Ambisonics vs mute fixation similarities across audio categories 54

Figure 17. Mean consistency scores and the fitted regression line plotted over time 55

Figure 18. Boxplots of consistency scores across audio categories and audio modes 58

Figure 19. Boxplots of consistency scores across audio and place categories..... 58

Figure 20. An equirectangular projection of the spotted sphere grating texture..... 70

Figure 21. Trial structure of the JND Calculation Block 73

Figure 22. Trial structure of the free view task block 78

Figure 23. Trial structure of the search task block 79
Figure 24. Trial structure of the memory recall task block 80
Figure 25. A boxplot showing average fixation durations across tasks 84
Figure 26. A boxplot showing target fixation rates across tasks 84
Figure 27. Boxplot showing response durations across tasks 86
Figure 28. A boxplot showing correct response rate across tasks 86
Figure 29. A boxplot showing target fixation rate across target positions per task .. 88
Figure 30. A boxplot showing target fixation rate task types per Z position 89
Figure 31. A boxplot showing response durations across target positions per task .. 90
Figure 32. A boxplot showing durations across task types per Z position 90
Figure 33. A boxplot showing correct response rate across target positions per task 91
Figure 34. Average fixation durations across luminance and color contrast per task 92
Figure 35. Target fixation rate across luminance and color contrast per task 93
Figure 36. Target fixation rate across different levels of color contrast 93
Figure 37. Response durations across luminance and color contrast per task 95
Figure 38. Response durations across different levels of color contrast per task 95
Figure 39. Correct response rate across luminance and color contrast per task 96
Figure 40. Average fixation durations across audio source groups per task 97
Figure 41. Target fixation rate across audio source groups per task 98
Figure 42. Response durations across audio source groups per task 99
Figure 43. Correct response rate across audio source groups per task 99
Figure 44. Average fixation duration across different loudness levels per task 100
Figure 45. Target fixation rate across different loudness levels per task 101

Figure 46. Response duration across different loudness levels per task 102
Figure 47. Correct response rate across different loudness levels per task type..... 103
Figure 48. Audio element fixation rate across different loudness levels per task type
..... 103
Figure 49. Average fixation duration across loudness and audio source per task ... 104
Figure 50. Target fixation rate across loudness and audio source per task..... 105
Figure 51. Response duration across loudness and audio source per task..... 106
Figure 52. Correct response rate across loudness and audio source per task..... 107
Figure 53. Login page screenshot of the custom video annotation website..... 116
Figure 54. Video annotation page screenshot of the custom website 116
Figure 55. Question list page screenshot of the custom website..... 117
Figure 56. Edit question details modal screenshot of the custom website..... 117
Figure 57. Video list page screenshot of the custom website 118
Figure 58. Edit video details modal screenshot of the custom website 118
Figure 59. Application programming interface documentation for the website 119
Figure 60. Mongo model definition for responses in the custom website 120
Figure 61. Mongo model definition for videos in the custom website 120
Figure 62. Mongo model definition for questions in the custom website..... 121
Figure 63. Mongo model definition for users in the custom website..... 121
Figure 64. The first half of the video clips in the created dataset 122
Figure 65. The second half of the video clips in the created dataset 123
Figure 66. Heatmap overlays from human speech, indoors category 130
Figure 67. Heatmap overlays from human speech, outdoors (human-made) category
..... 131
Figure 68. Heatmap overlays from human speech, outdoors (natural) category 132

Figure 69. Heatmap overlays from music instrument, indoors category 133
Figure 70. Heatmap overlays from music instrument, outdoors (human-made) 134
Figure 71. Heatmap overlays from music instrument, outdoors (natural) 135
Figure 72. Heatmap overlays from vehicle sound, indoors category..... 136
Figure 73. Heatmap overlays from vehicle sound, outdoors (human-made)..... 137
Figure 74. Heatmap overlays from vehicle sound, outdoors (natural) category..... 138
Figure 75. Boxplot of consistency scores across audio categories 139
Figure 76. Boxplot of consistency scores across place categories..... 139
Figure 77. Boxplot of consistency scores across audio modes 140
Figure 78. Boxplot of consistency scores across color modes..... 140
Figure 79. Consistency scores line plot of audio categories per audio mode 141
Figure 80. Consistency scores line plot of audio categories per color mode..... 141
Figure 81. Consistency scores line plot of place categories per audio mode..... 142
Figure 82. Consistency scores line plot of place categories per color mode 142
Figure 83. Consistency scores of audio categories per audio and color mode 143
Figure 84. Consistency scores of place categories per color and audio mode..... 143

ABBREVIATIONS

2AFC	Two-Alternative Forced Choice
2D	Two Dimensions
3D	Three Dimensions
ANOVA	Analysis of Variance
AUC	Area Under the Curve
CC	Correlation Coefficient
CFA	Confirmatory Factor Analysis
EEG	Electroencephalography
I-DT	Identification by Dispersion-Threshold
ID	Identifier
JND	Just Noticeable Difference
KLD	Kullback-Leibler Divergence
KMO	Kaiser-Meyer-Olkin
MANOVA	Multivariate Analysis of Variance
MEG	Magnetoencephalography
ML	Maximum Likelihood
NSS	Normalized Scanpath Saliency
RMSEA	Root Mean Square Error of Approximation
SIM	Similarity
SRMR	Standardized Root Mean Square Residual
TLI	Tucker-Lewis Index
UI	User Interface
VR	Virtual Reality

CHAPTER 1

INTRODUCTION

Throughout history, many researchers have investigated several kinds of human attention to better understand the wondrous system, the brain. Its high complexity requires lots of analysis in order to grasp a thorough model, but the combined effort of all brain-enthusiast researchers allows us to take little steps toward that goal. With a broad division, numerous research showed that attentional mechanisms and exploration behavior are influenced by both bottom-up processes (Treisman & Gelade, 1980; Van der Burg, Olivers, Bronkhorst, & Theeuwes, 2008; Yantis & Hillstrom, 1994) and top-down processes (Bruce & Tsotsos, 2005; A. Coutrot & Guyader, 2014; Itti & Koch, 2001; Judd, Ehinger, Durand, & Torralba, 2009; Wolfe, 1994), albeit both findings presume eye movement data as the indicator of the focal attention due to psychophysical evidence (Kaldas, 2022; Parkhurst, Law, & Niebur, 2002), consequently, investigating external focused attention. Highly salient features (Parkhurst et al., 2002; Treisman & Gelade, 1980) and temporal changes (Van der Burg et al., 2008; Wada, Kitagawa, & Noguchi, 2003; Yantis & Hillstrom, 1994) are example factors in involuntary bottom-up attention, whereas example top-down factors include informativeness, tendency to look at human faces, and task instructions (Cerf, Frady, & Koch, 2009; A. Coutrot & Guyader, 2014; Fan et al., 2018; Parkhurst et al., 2002). Based on information theory (Shannon, 1948), feature integration theory (Treisman & Gelade, 1980), and theory of visual attention (Bundesen, 1990), several visual attention models have developed, such as the Itti-Koch visual saliency model (L. Itti, Koch, & Niebur, 1998) with a focus on stimulus-driven selective visual attention, guided search model (Wolfe, 1994) with a focus on

search task, attention based on information maximation (Bruce & Tsotsos, 2005) with a focus on informativeness of the stimuli content. Furthermore, it has been shown that even without any spatial localization, auditory stimuli also take a role in guiding audio-visual attention both via its temporal features (Coutrot, Guyader, Ionescu, & Caplier, 2012; Van der Burg et al., 2008) and its semantic content (Coutrot & Guyader, 2014; Min, Zhai, Gao, Hu, & Yang, 2014). Combined efforts of computer vision, psychophysics, and many other disciplines yielded many models and observations regarding integrated audio-visual attention mechanisms (Bylinskii et al., 2016; Kaldas, 2022); however, such research is mostly limited to two-dimensional visual content and unlocalized auditory stimuli (Elwardy, Zepernick, & Sundstedt, 2019; Xu, Li, Zhang, & Callet, 2020). Therefore, the focus of this thesis is to compensate for the lack of research regarding attentional mechanisms (as a combination of both bottom-up and top-down processes) on 360° virtual reality (VR) environments and spatial audio via creating a controlled eye-tracking dataset on panoramic videos with omnidirectional audio and employing psychophysics methodology to study attentional mechanisms in three-dimensional VR environment.

According to the taxonomy Kaldas (2022) suggested, attention can be regarded as being composed of multiple mechanisms. Hence, it is important to clarify which mechanisms are to be studied. Within the scope of the thesis and the reviewed literature, the terms saliency and attention are usually used interchangeably. That is because the literature is focused on modeling the salience phenomenon of external focal attention measured via behavioral markers as per the definitions within that taxonomy. It is important to note that such a phenomenon can be guided by the combination of multiple mechanistic operations at the same time, so

it would be wise to keep that in mind when discussing the reasons behind any observed effects. Interested readers can check out their full taxonomy.

Commonly characterized via low-level features of signals, saliency is regarded as a critical driving force of bottom-up attention (Koch & Ullman, 1987; Ruesch et al., 2008; Treisman & Gelade, 1980). In a broader sense, however, it becomes tricky to draw a line between lower and higher-level features as well as to segregate bottom-up attention from top-down effects (Bylinskii et al., 2016; Xu, Jiang, Wang, Kankanhalli, & Zhao, 2014). Hence, a more appropriate definition would be to consider it as an "attend to me" signal. In other words, it can be regarded as the phenomenon of drawing attention to a target due to the properties of the target that make it stand out (Kaldas, 2022).

1.1 Saliency estimation models

Modeling saliency has many beneficial and practical applications in real life. Firstly, custom video compression methods that compress salient regions with higher quality and other regions with lower quality can be employed (Xu et al., 2020). Next, video quality assessment tools can leverage saliency models to prioritize quality at potential human fixation locations (Chiariotti, 2021). Furthermore, sophisticated frameworks for advertisement can be developed, such as generating heat maps of estimated view amounts for different locations in advertisements or even suggesting where to place critical headlines. Finally, saliency models can shed some light on how some of the human attention mechanisms are guided, allowing us to create better theories regarding the working principles of brain functions. Overall, saliency models have many applications, and better models can provide more benefits.

Due to the possible economic and psychological applications of saliency models, saliency estimation of scenes is a popular task in computer vision. Many researchers have developed several techniques to estimate the salient regions of scenes from visual or audio-visual stimuli. The widely accepted approach assumes that human observers are more likely to look at regions that are more salient in a scene, and hence, saliency can be measured by a behavior marker such as eye-tracking data (Judd, Durand, & Torralba, 2012; Kaldas, 2022). To that end, researchers commonly create datasets with eye-tracking data of several human observers when they are presented with several scenes. Then, the raw eye-tracking data (gaze directions per timestamp) obtained from those observers are processed to separate fixations, which occur when the gaze remains relatively stable on a particular point of interest for a certain duration, from saccades, which are rapid ballistic eye movements which help direct the gaze quickly to another region (Carter & Luke, 2020). Subsequently, the fixations are aggregated per scene to obtain fixation maps (heatmaps or probability distributions). Finally, these fixation maps are used as ground-truth saliency maps when benchmarking saliency prediction techniques (Bylinskii, Judd, Oliva, Torralba, & Durand, 2019; Judd et al., 2012; Kaldas, 2022). Whereas early techniques of saliency estimation consisted of signal processing methods, more recent techniques are dominated by deep learning-based machine-learning models (Borji, Cheng, Hou, Jiang, & Xu et al, 2020; Judd et al., 2012).

In those earlier models, traditional methods in signal processing were employed such that some low-level features assumed relevant to saliency were extracted from a target stimulus, and then, these features were integrated to obtain predicted saliency maps. Commonly used features can be grouped as follows: Static

visual features include color and intensity. Spatial visual features include orientation. Spatio-temporal visual features include flicker, motion, and optical flow. Temporal auditory features include intensity, temporal contrast, frequency, and frequency contrast. The most significant advantage of these models is that they are straightforward, defining the exact features used and how they are integrated, so understanding their behavior is easier compared to deep learning-based models (de Santana Correia & Colombini, 2022; Itti et al., 1998; Treisman & Gelade, 1980).

As a typical pattern, whenever the accuracies of predicted saliency estimates for a specific type of stimuli get very close to human performance (measured as how well observers' fixations can be predicted by the saliency map obtained from other observers), researchers start developing saliency estimation techniques for more complex stimuli building upon previous research. More specifically, the stimuli observed in contemporary research have evolved from images to silent videos to mono audio videos to multi-channel audio videos and from two dimensional (2D) visual content to 360° panorama visual content (Borji et al., 2019; Chiariotti, 2021; Elwardy et al., 2019) As the modalities of stimuli expanded, new challenges emerged from the complexity of additional modalities, and more sophisticated techniques have been developed to overcome them. Consequently, saliency estimation models have progressed from modeling effects of spatial-only to spatiotemporal to the fusion of temporal audio and spatiotemporal visual to the fusion of spatiotemporal audio and spatiotemporal visual stimuli. In addition to previous challenges, 360° panorama visual content requires handling viewport effects as well.

As of now, there are several saliency models developed for 360° videos, but most of them only consider visual modality such that they model visual saliency (Dahou, Tliba, McGuinness, & O'Connor, 2021; Hu, Bulling, Li, & Wang, 2021; Hu

et al., 2020; Qiao, Xu, Wang, & Borji, 2021; Zhang et al., 2020; Zhu, Chen, Zhao, Zhou, & Yang, 2021; Zhu, Zhai, Min, & Zhou, 2020), whereas there are only a handful of studies modeling multi-modal audio-visual saliency (Chao et al., 2020). In standard two-dimensional videos, however; there is already a lot of research on modeling audio-visual saliency (Min, Zhai, Gu, & Yang, 2016; Min et al., 2020; Tavakoli, Borji, Rahtu, & Kannala, 2019; Tsiami, Katsamanis, Maragos, & Vatakis, 2016; Tsiami, Koutras, Katsamanis, Vatakis, & Maragos, 2019; Tsiami, Koutras, & Maragos, 2020).

Unlike visual saliency, auditory salience by itself is hard to measure, as there is no type of ground-truth data that is agreed on by all researchers. Kaya and Elhilali (2017) created a detailed survey about auditory saliency and explained different methods of measurement for auditory saliency. An interesting novel technique is to collect brain activity data via magnetoencephalography (MEG) or electroencephalography (EEG) and try to reproduce the attended sounds from that data (Ding & Simon, 2012; Mesgarani & Chang, 2012; O'Sullivan et al., 2014). More commonly, researchers directly model audio-visual saliency instead and rely solely on eye-tracking data as the measurement. Nevertheless, combining data across different modalities presents another challenge. Many researchers approached this challenge by developing methods with different fusion approaches (Min et al., 2020; Tsiami et al., 2019). Such methods are common for videos with single-channel audio, considering only the temporal effects of audio. Instead of conventional mono and stereo audio formats, VR devices can leverage the ambisonics audio format (Gerzon, 1973; Zotter & Frank, 2019), which represents the audio in three dimensions (3D) and is preferred in 360° videos so that both visual and auditory signals are presented according to the viewer's head position and rotation (Martin,

Malpica, Gutierrez, Masia, & Serrano, 2022). This type of stimuli allows for the consideration of both the temporal and spatial effects of audio when studying audio-visual saliency and integration.

There are many datasets designed for saliency modeling purposes in two-dimensional environments, including both visual and audio-visual stimuli and eye fixation data (Judd et al., 2012). Likewise, benchmarks and similarity metrics for evaluating models are well established in two-dimensional environments (Bylinskii et al., 2019; Judd et al., 2012). Recently, similar datasets have been created by different researchers in the omnidirectional domain as well (David, Gutiérrez, Coutrot, Da Silva, & Callet, 2018; Zhang, 2021; Zhang, Zhang, Hamidouche, & Deforges, 2020). Moreover, early benchmarks and modified similarity metrics are already available for evaluating models in the omnidirectional domain (Gutierrez, David, Coutrot, Da Silva, & Callet, 2018). Latterly, Yuxin Zhu et al. (2023) created a new eye-tracking omnidirectional video dataset of several content categories across three audio modalities (mute, mono, and ambisonics). However, the literature is not yet wholly established as there are only a handful of datasets, and the preparation of additional datasets is required to develop more sophisticated models.

1.2 Exploratory research on saliency

Along with modeling efforts, many studies also focus on analyzing human behavior to draw inferences about the mechanisms behind the saliency phenomenon. There are several observational studies in two-dimensional environments that might be interesting to repeat in omnidirectional videos. Ciortan, Dinet, and Trémeau (2014) revealed that the color modality of short videos has only a limited influence on visual attention. Xu et al. (2014) and Bylinskii et al. (2016) studied the effects of specific

types of objects in scenes and their semantic attributes on fixations. Fan et al. (2018) analyzed objects' emotional valence and how that affects attention. Such studies not only improve our understanding of saliency but also help researchers to develop better models. Working with 360° content also allows observing human behavior in a more naturalistic environment since they are more ecologically valid than their 2D alternatives. Moreover, VR devices allow the use of spatial audio components, hence allowing to study audio-visual integration effects with spatial audio. Thus, such content allows richer observations and analyses. In that manner, as the audio-visual saliency literature for omnidirectional videos is being developed, interesting observations about new high-level features are made. Pillai & Verma (2019) studied perceptual cues in VR cinema. Haskins, Mentch, Botch, & Robertson (2020) analyzed the effects of being able to move the head and the corresponding changes in the field of view on eye fixations. Maranes, Gutierrez, & Serrano (2020) used head-motion data from people watching VR movies and analyzed user behavior across different movie cuts. Hu et al. (2021) and Jiang, Chen, Yang, & Zhao (2020) studied how task-oriented behavior affects fixations. In order to build a comprehensive understanding of saliency and to develop better models, even more such studies are required.

1.3 The scope and the purpose of this thesis

All experiments aim to explore several mechanisms of human saliency to help develop a more comprehensive understanding of the saliency. To begin with, several exploratory analyses regarding the effects of audio and place categories were conducted in available datasets. Next, a new dataset consisting of an equal number of panoramic videos with spatial audio across audio and place categories was created to

conduct a more controlled experiment by collecting and analyzing eye-tracking data from participants while they watched these videos via a VR headset across different audio (mute, mono, spatial) and color (grayscale, colored) modulation conditions. Finally, the effects of task type over salience were investigated by conducting a psychophysics experiment in which participants were presented with several simple stimuli with varying chromaticity contrast, luminance, and loudness levels in different tasks (free viewing, search, memory recall).

CHAPTER 2

PRELIMINARY EXPERIMENT

2.1 Objectives

In this experiment, the aim is to explore the effects of audio on saliency in omnidirectional videos and how it varies according to several high-level features of the videos. Throughout this study, it is assumed that the ground truth for saliency is human fixation locations collected through eye trackers. We suggest that the role and effects of audio might change depending on the video scene type, whether indoors or outdoors, and the audio content type. The findings of the study help improve the current understanding of human attention mechanisms and allow researchers to develop better saliency models that incorporate observations from the study.

2.2 Methodology

To conduct planned analyses, instead of collecting new data, it was decided to use an existing dataset, ASOD60K, which is a recently published dataset. It contains eye fixations and head positions from 40 participants collected via HTC Vive VR headset with an embedded Tobii eye tracker device while free viewing 67 video clips with no instruction (Zhang, 2021). More importantly, during the dataset preparation, the participants were divided into two groups. The first group viewed the videos with their audio muted, whereas the second group viewed the videos with mono audio. The videos are collected from YouTube and cropped at specific time intervals to create the video clips. All videos had resolutions varying between high definition (HD) and ultra-high definition (4K UHD), the equirectangular image format, and ambisonics audio format. Appendix A shows a collage of all videos in the dataset.

2.2.1 Collecting original videos

In order to conduct analyses regarding the study's predictions, it was necessary to label scene types and audio types for each video. Unfortunately, mute clips with reduced framerate and the links of full YouTube videos were published in ASOD60K instead of the original clips. Thus, a method was developed to match frames of muted clips with the original uncropped videos to deduce crop time boundaries. See Appendix B for details of that method. Using that method, original video clips were extracted from the full videos. Then, these clips were manually inspected by comparing them to low frame-rate video clips. Doing so revealed that original video clips were reproduced with at least 0.2 seconds precision.

2.2.2 Labeling audio categories

To analyze audio-type-based effects, video clips needed to be associated with some semantic categories. To do so, the audio of each clip was extracted. Then, a machine learning model trained on the VGGSound dataset was used to assign audio type labels for each audio (Chen, Xie, Vedaldi, & Zisserman, 2020). For simplicity and as per the input type of the machine learning model, all audio data was converted to mono audio before labeling. After running the model for each video clip, each video clip was assigned the audio label with the highest prediction score. When the assigned labels were inspected, some labels did not seem reasonable. As an example, a very noisy video clip was assigned dinosaur sounds under animal sounds. As an alternative, another machine-learning model trained on the AudioSet dataset was used to assign a second audio-type label for each video clip (Gemmeke et al., 2017; Hershey et al., 2017; Tensorflow, Plakal, & Ellis, 2020). Unlike the previous model, where the labeling was based on the whole video clip, this one labeled each

timeframe. For each video clip, prediction scores for each label across timeframes were averaged, and the label with the highest average prediction score was assigned. Table 1 and Table 2 show the labels predicted by running these models for an example video clip. Figure 1 and Figure 2 show the number of video clips assigned to high-level audio type labels for each model.

Table 1. Top 6 VGGSound Labels of the Sample Video Clip

Prediction Score (Mean)	VGGSound Label	High-Level Label
0.4705	Playing double bass	Music
0.2037	Female singing	Music
0.0974	Playing acoustic guitar	Music
0.0704	Playing bass guitar	Music
0.0328	Male singing	Music
0.0228	Playing harp	Music

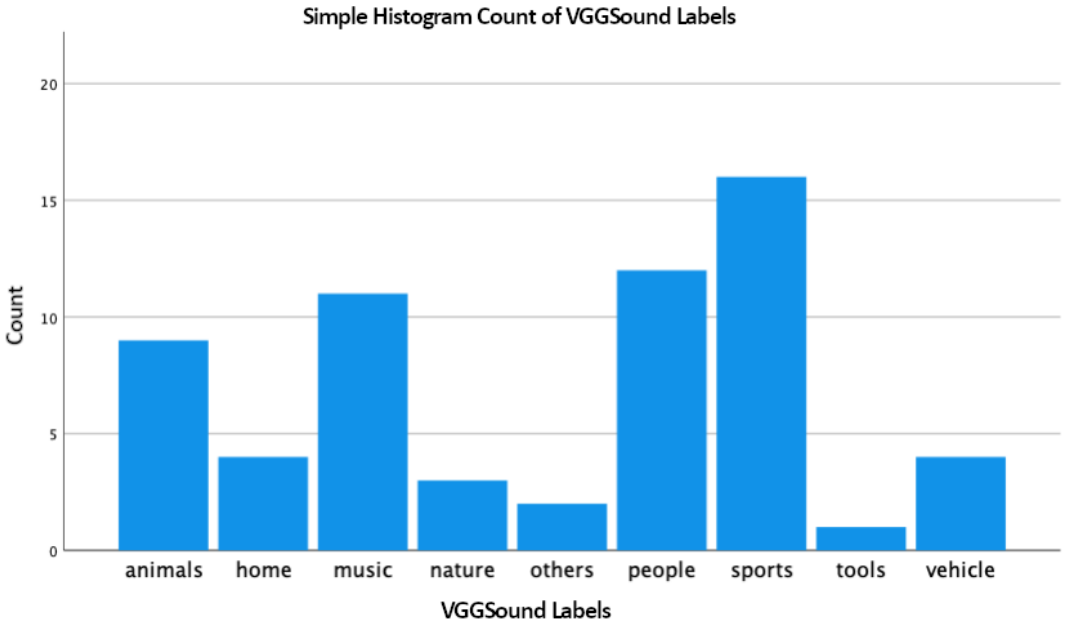


Figure 1. Histogram of VGGSound high-level audio-type categories across videos

Table 2. Top 6 AudioSet Labels of the Sample Video Clip

Prediction Score (Mean)	AudioSet Label	High-Level Label
0.7240	Music	Music
0.0826	Lullaby	Music
0.0222	Musical instrument	Music
0.0202	Singing	Music
0.0192	Plucked string instrument	Music
0.0171	Silence	Music

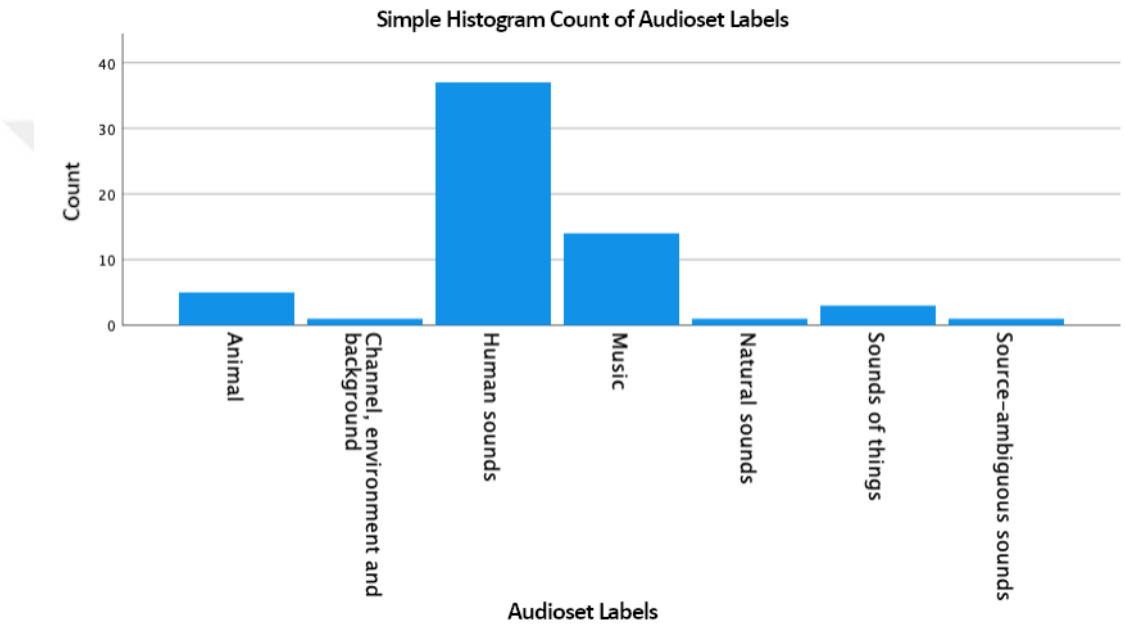


Figure 2. Histogram of AudioSet high-level audio-type categories across videos

2.2.3 Labeling place categories

Lastly, a similar procedure was applied to analyze scene-type-based effects. However, existing scene detection machine learning models were based on two-dimensional images rather than equirectangular images representing a spherical image. To deal with this problem, the frame rate of each video was reduced to one-fifth, and 15 two-dimensional images were extracted from each of these frames. More specifically, images centered at 0, 72, 144, 216, and 288 degrees in the horizontal scale and -18, 0, and 18 degrees in the vertical scale were extracted.

Figure 3 shows a collage of such images extracted from a single frame of an example video. That way, each adjacent extracted image greatly overlapped with the other, so no scene detail was missed. The extremes on the vertical scales were not included to prevent using any image consisting of sky or ground only, which could be misleading for the scene classification. Next up, all these extracted two-dimensional images were classified by running a machine learning model trained on the Places365 dataset (Zhou, Lapedriza, Khosla, Oliva, & Torralba, 2018). The resulting prediction scores for each label across different extracted images and frames were averaged per video clip. Finally, the label with the highest average prediction score was assigned to each video clip. Table 3 shows the labels predicted by running the model for an example video clip. Figure 4 shows the number of video clips assigned to high-level place type labels.

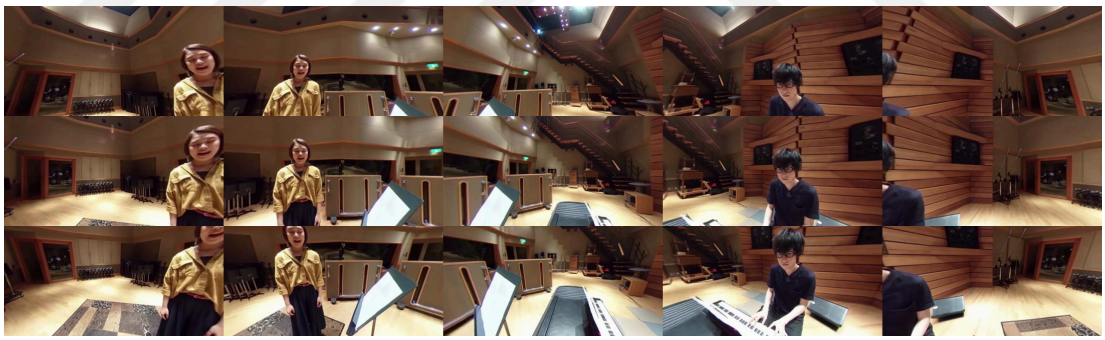


Figure 3. A collage of all 15 extracted view shots from a single frame

The frame is taken from the middle of the sample video clip, “gTB1nfK-0Ac”. The same video clip is referred whenever the sample video clip is mentioned in other figures and tables.

Table 3. Top 6 Places365 Labels of the Sample Video Clip

Prediction Score (Mean)	Places365 Label	High-Level Label
0.7240	Sauna	Indoor
0.0826	Attic	Indoor
0.0222	Music Studio	Indoor
0.0202	Television Studio	Indoor
0.0192	Auditorium	Indoor
0.0171	Mezzanine	Indoor

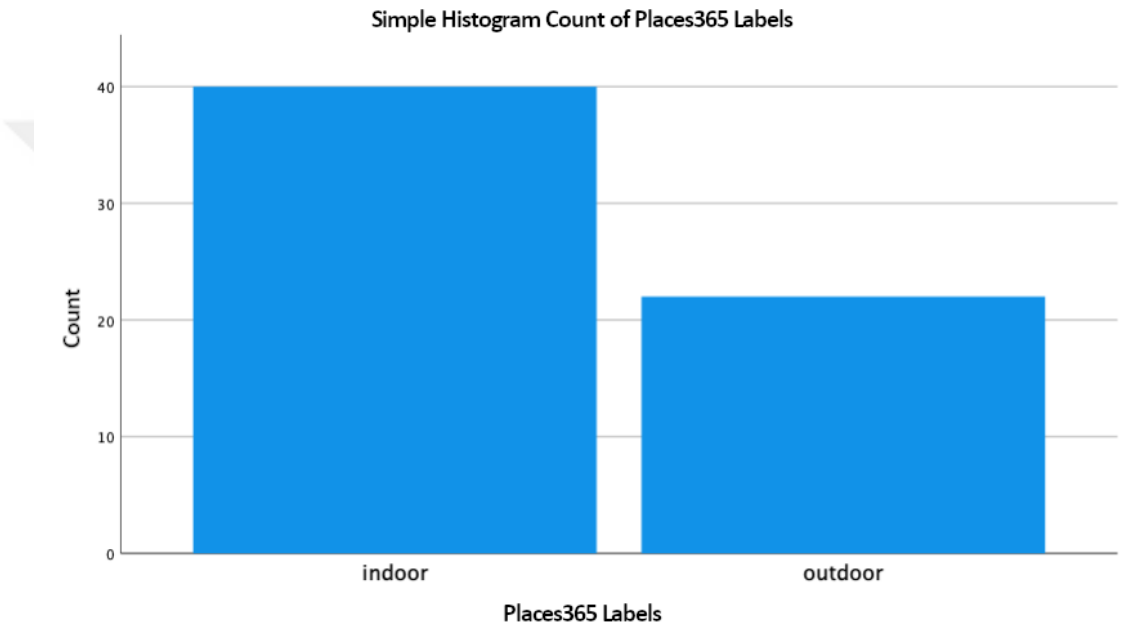


Figure 4. Histogram of Places365 high-level place-type categories across videos

Thus, in addition to eye fixations from observers when observing each video clip (grouped by whether the clip was observed with audio or in mute), the dataset was augmented by the most dormant VGGSound and AudioSet audio categories and Places365 scene categories assigned per video clip.

2.3 Results

2.3.1 Similarity scores

There are different metrics and methods for comparing how similar a group of fixations is to a given human eye fixation density map. Area under the curve (AUC), normalized scanpath saliency (NSS), correlation coefficient (CC), similarity (SIM), and Kullback-Leibler divergence (KLD) scores are the most popular and well-studied metrics, each differing slightly due to differences in their mechanisms and can be interpreted separately (Bylinskii et al., 2019). Thus, these similarity scores for each video were calculated by comparing fixations from the with-sound case to that of the without-sound (muted) case to use in the analyses. One can refer to Appendix C for details regarding these calculations. Table 4 shows such calculated similarity scores for an example video.

Table 4. Similarity Scores of the Sample Video Clip

Time (s)	AUC Borji	AUC Judd	NSS	CC	SIM	KLD
…	…	…	…	…	…	…
10.6	0.723	0.801	3.505	0.677	0.474	3.855
10.7	0.675	0.754	2.956	0.615	0.409	5.080
10.8	0.676	0.676	2.334	0.546	0.398	6.021
10.9	0.666	0.744	2.068	0.533	0.406	4.589
11.0	0.682	0.767	2.436	0.529	0.404	4.546
11.1	0.689	0.764	2.424	0.530	0.392	4.530
11.2	0.676	0.726	2.913	0.586	0.405	4.929
11.3	0.661	0.713	3.105	0.627	0.409	5.091
11.4	0.656	0.758	3.208	0.720	0.451	3.718
11.5	0.661	0.771	3.811	0.742	0.446	4.444
…	…	…	…	…	…	…
Mean	0.719	0.732	3.626	0.674	0.502	4.904

Note: Similarity scores across the 11th second are listed for the sample video clip. Each row corresponds to scores for individual 200ms chunks centered at given time points. The mean row corresponds to the average of all rows for the whole video clip.

To be able to interpret the results better, heatmap visualizations of density maps for each video clip were generated and overlayed on top of the video clips. Density maps for cases with sound were visualized as shades of red, and density maps for cases without sound were visualized as shades of green; hence, the overlapping regions became shades of yellow. An example heatmap visualization is shown in Figure 5. Broadly, these scores correspond to the amount of overlapping regions of heatmaps of the cases with and without sound. Thus, when comparing across different groups, if a group has a higher similarity score than another, fixations collected when video clips are watched with and without sound are more similar to each other than that of the other group. So, the effect of auditory signals on fixations is lower in groups with higher similarity scores than those with lower similarity scores.



Figure 5. Human eye fixation density map visualizations on the sample video

The heatmap calculated at the 11th second of the sample video clip from observer eye fixations is overlayed on top of the video content. Red shades of heatmaps correspond to density maps for the case with the sound, green shades of heatmaps correspond to density maps for the case without the sound, and yellow shades of heatmaps correspond to overlapping regions of those two.

Our initial expectation was that the fixations from the with-sound conditions would be concentrated on locations relevant to the audio, and the fixations from the without-sound conditions would be scattered more randomly. However, manual inspection of heatmaps showed that participants from both with- and without sound conditions fixated on locations irrelevant to audio from time to time. Therefore, it was expected for biased metrics, AUC and KLD, to capture less of the overall variance across conditions than other metrics since AUC focuses more on false positives (unexpected fixations or fixations in low probability regions of estimated saliency map distributions) and KLD focuses more on false negatives (no fixations present in high probability regions of estimated saliency map distributions) but both conditions contain many random false positives and negatives due to random viewing behavior. As opposed to those, unbiased metrics CC, NSS, and SIM measure similarity via symmetric comparisons of the whole probability densities, and hence, it was expected for these metrics to capture the variance of overall behavior across conditions (Bylinskii et al., 2019). This expectation suggests that although both metrics measure similarity, the analyses based on their outputs could yield slightly different results. Consequently, along with separate univariate tests, we also conducted multivariate analysis of variance (MANOVA) tests. However, correlations between all similarity metrics were usually high (Pearson's $r > 0.8$) and sometimes very high ($r > 0.9$) in all analyses, leading to multicollinearity. To deal with this issue, separate univariate analysis of variance (ANOVA) tests were done as follow-ups to the MANOVA analysis instead of conducting discriminant analyses. Nevertheless, the most exciting results came from NSS, CC, and SIM scores rather than AUC and KLD scores, validating the expectation that NSS, CC, and SIM might have the highest differences across different cases. All analyses in the current

experiment were conducted using SPSS, a popular statistical analysis software (IBM Corp, 2020).

2.3.2 Effects of place category

Places365 provides hierarchical labels according to scene type (Zhou et al., 2018). In order to get enough samples from all groups, the highest-level labels were used. These labels are (a) indoors, and (b) outdoors. A MANOVA analysis was conducted to investigate the effects of scene types, indoors ($n = 40$) and outdoors ($n = 22$), on similarity metrics: AUC Borji, AUC Judd, NSS, CC, SIM, and KLD. Using Pillai's trace, a significant effect of place category on the similarity of with and without sound fixations was detected, $V = 0.25$, $F(6, 55) = 3.11$, $p = 0.011$, $\eta^2 = 0.253$. Separate univariate tests revealed non-significant effects of scene type on NSS, $F(1,60) = 1.18$, $p = 0.281$, $\eta^2 = 0.019$, and CC, $F(1,60) = 2.80$, $p = 0.100$, $\eta^2 = 0.045$; but significant effects on AUC Borji, $F(1,60) = 4.44$, $p = 0.039$, $\eta^2 = 0.069$, and KLD, $F(1,60) = 6.85$, $p = 0.011$, $\eta^2 = 0.102$; and strong significant effects on SIM, $F(1,60) = 6.99$, $p = 0.010$, $\eta^2 = 0.104$, and AUC Judd, $F(1,60) = 7.40$, $p = 0.009$, $\eta^2 = 0.110$. In order to control Type 1 error inflation for multiple comparisons, only the strong effects on SIM and AUC Judd are considered. Follow-up two-tailed t-tests revealed that AUC Judd scores in the indoors case ($M = 0.82$, $SD = 0.05$) were significantly lower than in the outdoors case ($M = 0.85$, $SD = 0.04$, $p = 0.009$). Likewise, SIM scores in the indoors case ($M = 0.60$, $SD = 0.06$) were significantly lower than in the outdoors case ($M = 0.64$, $SD = 0.06$, $p = 0.010$).

Mean similarity scores (and negative KLD scores), were always lower in the indoors case than in the outdoors case (see Figure 6). This difference was significant

for AUC Judd and SIM scores. From this finding, it can be inferred that the fixations are regulated by auditory signals more in indoors than in outdoors.

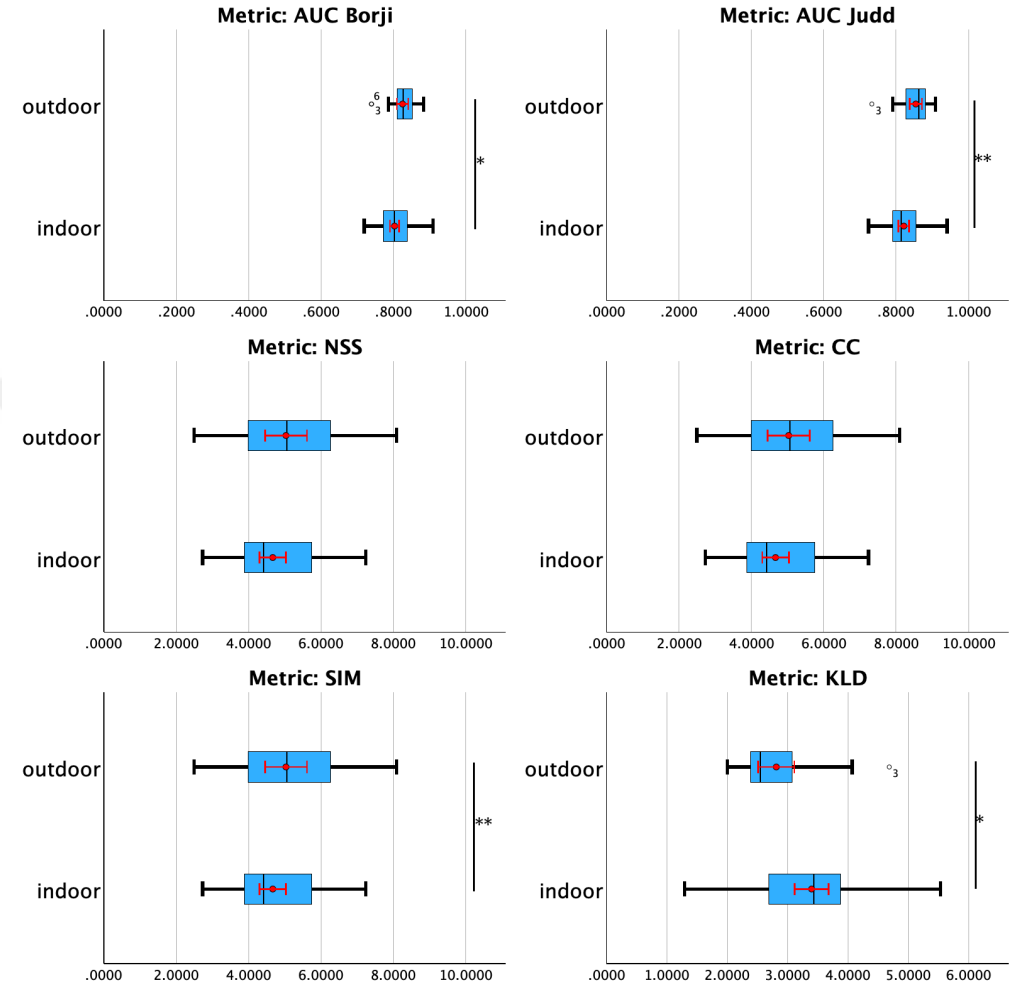


Figure 6. Boxplots of similarity scores for Places365 place categories

All similarity metrics correspond to similarity of fixation maps from mute and mono conditions.

2.3.3 Effects of audio category

VGGSound provides hierarchical labels according to audio type (Chen et al., 2020). In order to get enough samples from all groups, the highest-level labels were used, and the cases with fewer than five samples were merged into the “others” group. The resulting labels are (a) animals, (b) people, (c) music, (d) sports, and (e) others. A

MANOVA analysis was conducted to investigate the effects of audio types; animals ($n = 9$), people ($n = 12$), music ($n = 11$), sports ($n = 16$), and others ($n = 14$) on similarity metrics: AUC Borji, AUC Judd, NSS, CC, SIM, and KLD. Using Pillai's trace, no significant effect of audio type on the similarity of with and without sound fixations was detected, $V = 0.44$, $F(24{,}220) = 1.14$, $p = 0.303$, $\eta^2 = 0.111$. Since the multicollinearity assumption for most variables failed due to Pearson's correlations higher than 0.9 ($r > 0.9$), the MANOVA results were ignored, and separate univariate tests were conducted. These tests revealed non-significant effects of audio type on AUC Judd, $F(4{,}57) = 1.86$, $p = 0.130$, $\eta^2 = 0.116$, and KLD, $F(4{,}57) = 1.92$, $p = 0.120$, $\eta^2 = 0.119$, and CC, $F(4{,}57) = 2.43$, $p = 0.058$, $\eta^2 = 0.146$, and NSS, $F(4{,}57) = 2.52$, $p = 0.051$, $\eta^2 = 0.150$; but significant effects on AUC Borji, $F(4{,}57) = 2.69$, $p = 0.040$, $\eta^2 = 0.159$, and strong significant effects on SIM, $F(4{,}57) = 3.68$, $p = 0.010$, $\eta^2 = 0.205$. In order to control Type 1 error inflation for multiple comparisons, only the strong effects on SIM are considered. Bonferroni corrected post-hoc tests revealed that SIM scores in the music sounds case ($M = 0.57$, $SD = 0.04$) were very significantly lower than the animal sounds case ($M = 0.67$, $SD = 0.05$, $p = 0.008$). There were no significant differences between other pairwise comparisons (all $p > 0.05$).

All mean similarity scores (and negative KLD scores) for sports and other sounds were lower than animal sounds and higher than people sounds. Moreover, all similarity scores (and negative KLD scores) were lowest for music sounds, resulting in a non-significant ordering of $music < people < sports, others < animals$. SIM scores for the music sounds case were significantly lower than the animal sounds case (see Figure 7). From these findings, it can be concluded that the fixations are regulated more by music-related sounds than by animal-related sounds.

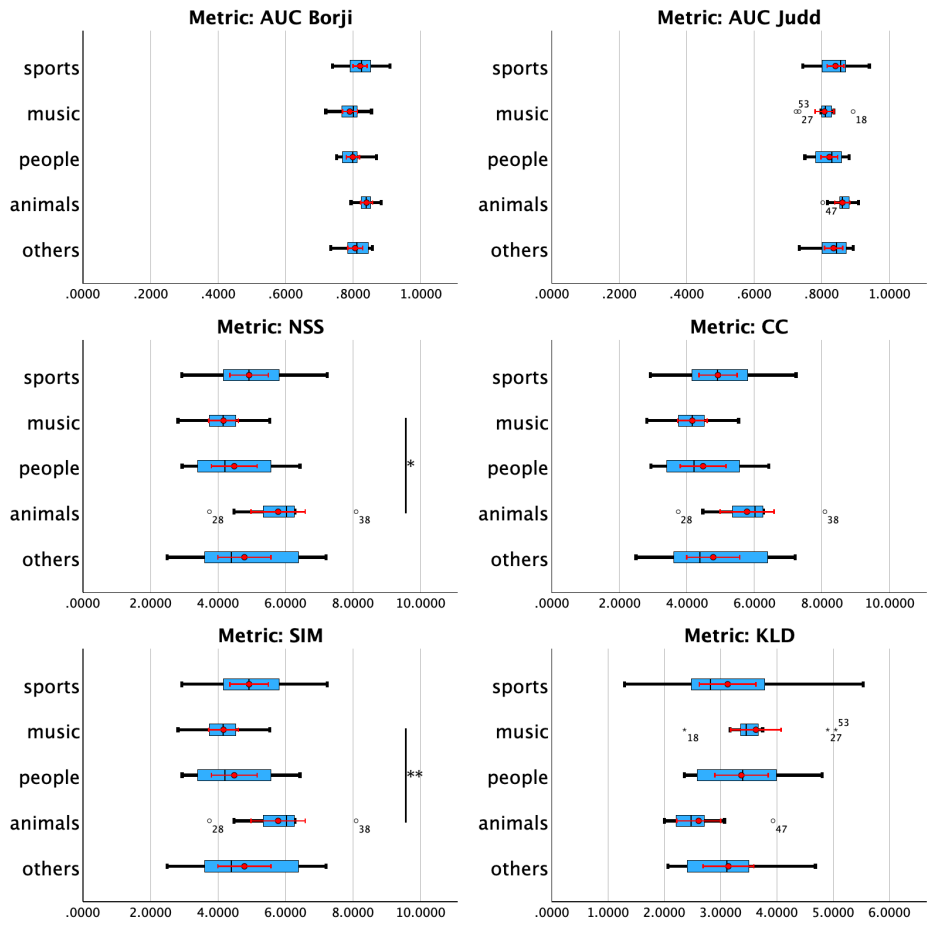


Figure 7. Boxplots of similarity scores for VGGSound audio categories

All similarity metrics correspond to similarity of fixation maps from mute and mono conditions.

AudioSet also provides hierarchical labels according to audio type (Gemmeke et al., 2017). In order to get enough samples from all groups, the highest-level labels were used, and the cases with less than five samples were merged into the "other" group. The resulting labels are (a) animal, (b) human, (c) music, and (d) other. A MANOVA analysis was conducted to investigate the effects of audio types; animal ($n = 5$), human ($n = 37$), music ($n = 14$), and other ($n = 6$) on similarity metrics: AUC Borji, AUC Judd, NSS, CC, SIM, and KLD. Using Pillai's trace, a significant effect of audio type on the similarity of with and without sound fixations was detected, $V = 0.52$, $F(18,165) = 1.91$, $p = 0.018$, $\eta^2 = 0.173$. Separate univariate

tests revealed significant effects of audio type on AUC Judd, $F(3,58) = 3.06$, $p = 0.035$, $\eta^2 = 0.136$, and KLD, $F(3,58) = 3.18$, $p = 0.030$, $\eta^2 = 0.141$, and AUC Borji, $F(3,58) = 4.01$, $p = 0.012$, $\eta^2 = 0.172$; very strong significant effects on CC, $F(3,58) = 6.66$, $p = 0.001$, $\eta^2 = 0.256$, and NSS, $F(3,58) = 6.99$, $p < 0.001$, $\eta^2 = 0.265$, and SIM, $F(3,58) = 7.70$, $p < 0.001$, $\eta^2 = 0.285$. In order to control Type 1 error inflation for multiple comparisons, only the strong effects on CC, NSS, and SIM are considered. Bonferroni corrected post-hoc tests revealed that CC scores for the music sounds case ($M = 0.72$, $SD = 0.05$) were significantly lower than the animal sounds case ($M = 0.85$, $SD = 0.07$, $p = 0.013$) and the human sounds case ($M = 0.81$, $SD = 0.09$, $p = 0.002$). Likewise, NSS scores for the music sounds case ($M = 3.85$, $SD = 0.59$) were significantly lower than the animal sounds case ($M = 5.73$, $SD = 1.70$, $p = 0.011$) and the human sounds case ($M = 5.16$, $SD = 1.23$, $p = 0.002$). Additionally, SIM scores for the music sounds case ($M = 0.56$, $SD = 0.04$) were very significantly lower than the animal sounds case ($M = 0.67$, $SD = 0.05$, $p = 0.002$) and the human sounds case ($M = 0.63$, $SD = 0.06$, $p = 0.002$). Finally, SIM scores for the other sounds case ($M = 0.58$, $SD = 0.05$) were significantly lower than the animal sounds case but only slightly ($p = 0.040$). There were no significant differences between other pairwise comparisons (all $p > 0.05$).

All mean similarity scores (and negative KLD scores) were ordered from highest to lowest as animals, humans, others, and music sounds, resulting in a non-significant ordering of $music < other < human < animal$. NSS, CC, and SIM scores for the music sounds case were significantly lower than both the human and the animal sounds case (see Figure 8). Thus, it can be concluded that the fixations are regulated more by music-related sounds than by animal or human-related sounds.

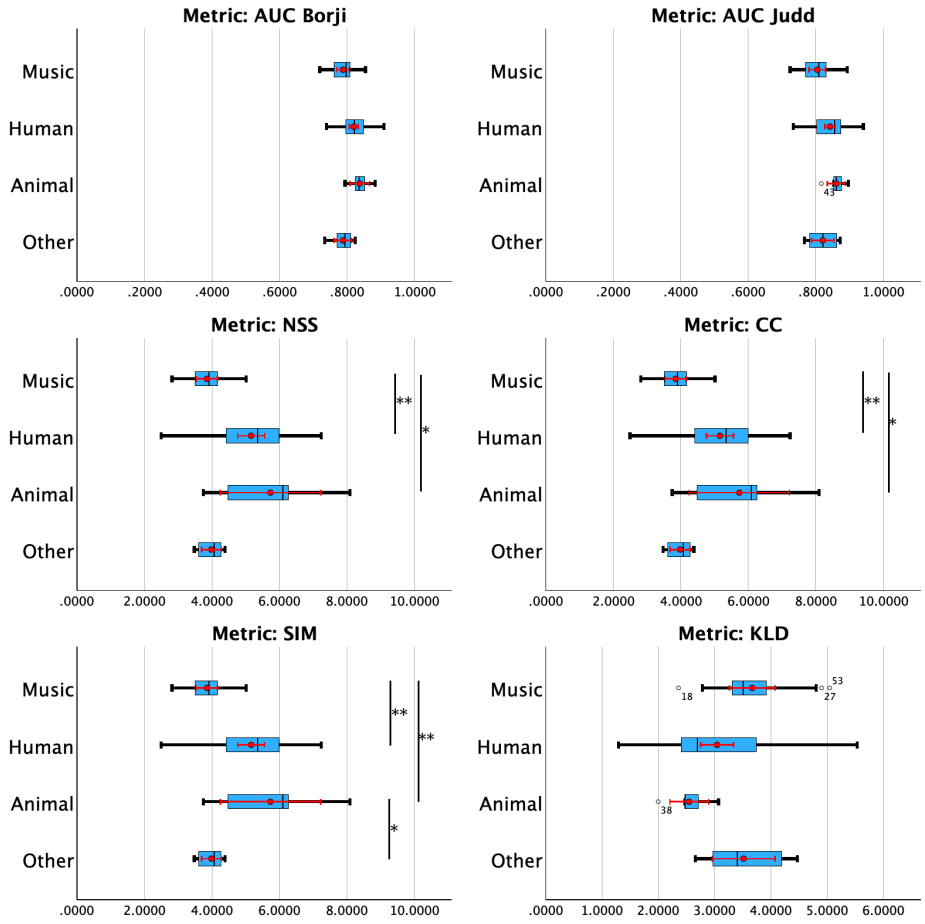


Figure 8. Boxplots of similarity scores for AudioSet audio categories

All similarity metrics correspond to similarity of fixation maps from mute and mono conditions.

2.4 Discussions and conclusions

The phenomena observed in regulatory effects of audio in human fixations were that these effects increased in indoor scenes or if the audio is of music but decreased in outdoor scenes or if the audio is of animal or human sounds. This phenomenon can possibly be explained by speculating that there usually are many inanimate objects indoors, leading humans to depend more on auditory cues when deciding which one to focus on, whereas there are more actions and many dynamic visual cues outdoors, making the auditory cues a second priority on where to focus. Within observational studies in two-dimensional environments with visual-only stimuli, it has already been

shown that human faces are considered very salient (Bylinskii et al., 2016; Fan et al., 2018). In accordance with that, the phenomenon regarding audio type might be explained by saying that humans already direct their attention towards humans and animals by visual stimuli, but their attention can shift towards instruments or music-generating inanimate objects via music sounds.

All strongly significant results found in the analyses had large effect sizes ($\eta^2 > 0.14$), and only those results were considered accurate and discussed further. Since the analyses were conducted on an existing dataset prepared for a different purpose, there were unequal sizes of samples of each scene type. Thus, conducting a further study by collecting fixation samples both with sound and without sound observations on carefully selected video samples that include many and preferably equal numbers of samples from different contexts would be more reliable. Moreover, repeating the analyses for different types of contextual information would yield more observations about how high-level contextual features affect human fixations.

Results support that building saliency models would be improved by extracting contextual information from the content first, such as audio type and scene type, and using it within the model. Moreover, instead of developing separate models for each potential context, converting context information into a weighting scheme and incorporating it into the saliency models might be more feasible. Therefore, as a simple example scheme, weighting auditory saliency higher when the context is indoors or consists of music sounds before combining auditory and visual saliency models could be useful when designing an audio-visual saliency model.

CHAPTER 3

360° VIDEO WATCHING EXPERIMENT

The purpose of this experiment is to collect eye-tracking data from participants in a more controlled method via equal sampling of videos across several audio and place categories to investigate further the findings shown in the preliminary experiment. Additionally, several audio and color modulations of videos allow the creation of a new dataset valuable for both computer science and psychophysics research. To our knowledge, during our experiment, there were neither any available datasets to investigate spatial effects of audio by comparing mono to ambisonics, nor any datasets to investigate the effects of color on eye fixations when users were watching 360° videos. Therefore, the main objective is to create and publish a general-purpose dataset, the YT360-EyeTracking dataset, to allow future research and analyses.

3.1 Research questions

The initial question is to see how reliable the results regarding audio and place categories from the preliminary experiment are. Therefore, they were investigated using a more controlled design, curating equal-sized samples of videos across each category combination, and conduct an exploratory research.

3.1.1 Effects of time

In general, when a new video starts, we expect users to first make a quick exploration of their surroundings to get adjusted to the new environment, and then switch to exploring based on semantic content. Thus, the consistency scores should be lower in the early stages than later. However, as we start all users from the same initial

position, their eye fixations would be very close initially, then scatter around during a quick exploration and adaptation behavior, and slowly converge into salient regions of the stimuli.

3.1.2 Effects of color and audio modalities

Due to information theory, the information of content is reduced as audio and color channels are removed as a consequence of reduced amounts of signals in stimuli (Itti & Koch, 2000; Krasovskaya & MacInnes, 2019; Treisman & Gelade, 1980). When the color channels are removed, many visual stimuli might be challenging to differentiate and become indistinguishable from a noise signal. Consequently, the meaningfulness of the content, along with the overall saliency, might be reduced. Moreover, as Tsiami et al. (2016) mentioned, the sound can cause a pop-out effect irrespective of the number of distractors. Hence, irrespective of whether it is spatial or mono, the audio can govern a pop-out effect, preventing any effect regarding the number of objects. Therefore, we hypothesize that eye fixations might be more random and unguided in mute and grayscale conditions, leading to lower inter-group consistency scores.

3.1.3 Effects of audio category

The preliminary experiment showed a significant effect of music sounds on fixations. It is predicted that even if a music instrument is being played, observers only look at it when it is audible. Therefore, it is expected that consistency scores should be higher when those videos are watched with increasing audio modalities. Next up, human faces are already considered very salient by their visual features alone (Bylinskii et al., 2016; Fan et al., 2018). It is predicted that when a human speech is

audible, they would be even more salient, causing the viewers to look around less randomly. Therefore, it is expected that consistency scores should be high for videos with human speech but even higher when watched with increasing audio modalities. Finally, the vehicle sounds are considered similar to a noise signal. Therefore, we expect consistency scores to be the same or lower with increasing audio modalities.

3.1.4 Effects of place category

The preliminary experiment showed a significant effect of place category on fixations. To better understand this effect, we will explore it in more detail and check its interaction with color and audio.

3.2 Methodology

3.2.1 Dataset preparation

In order to conduct our research, we needed 360° videos with first-order ambisonics sounds but with equal samples of semantic categories of audio types and place types. Even though such videos are not as common as their 2D counterparts, the YouTube platform still contains a lot of them. To train a semantic segmentation machine learning model for 360° videos with first-order ambisonics, Morgado, Li, & Vasconcelos (2020) created a dataset, Youtube-360, which contains 88733 clips of 10 seconds each and calculated semantic segmentation data via a state-of-the-art model for 2D. They did so by first projecting random viewpoints to 2D, then calculating segmentation maps, and then re-projecting back to the 360° domain. In summary, their dataset is not only filtered for erroneous content (such as still images, invalid ambisonics format, and significant post-production sounds) but also semantically segmented with 32 types of objects and background classes, which

might come in handy for alternative analyses. The only downside was that the clips were only 10 seconds each.

For the current research, A similar process of downloading videos, filtering for erroneous content, labeling them with audio and place types, and then choosing appropriate samples for each category was needed. An a priori power analysis was conducted using G*Power version 3.1 (Faul, Erdfelder, Buchner, & Lang, 2009). Results indicated that to achieve 95% power for detecting a medium effect ($f = 0.5$) at a significance criterion of $\alpha = .05$, the required sample size was $N = 63$ for a mixed model ANOVA (3x3 between factors of sound and scene types, 3x2 within factors of sound and color modalities) corresponding to seven videos per category. To be safer, a sample size of $n = 9$ per category ($N = 81$ overall) was decided for our tests. Also, it was decided to use clips of 30 seconds so that users could have enough time to explore a scene. After realizing that many clips in YT-360 were actually consecutive sequences within a longer video, those sequences were merged to obtain 2309 clips of 30 seconds or more. Only 2171 were still accessible on YouTube. Additionally, it was decided only to use high-quality video with resolutions of at least 3840x1920, leaving 1867 clips to be usable. Since there seemed to be enough remaining videos, it was decided to choose a subset of those clips to simplify the data filtering process and include semantic segmentation data along with the proposed dataset.

All 1867 such clips of YT-360 were downloaded via YT-DLP (yt-dlp, pukkandan, & shirt-dev, 2021) and labeled with models trained on Places365 (Zhou et al., 2018), VGGSound (Chen et al., 2020), and AudioSet (Gemmeke et al., 2017; Hershey et al., 2017; Tensorflow et al., 2020) datasets, similar to the preliminary research. Upon manual inspection of several videos, it was seen that many videos

contain erroneous content, namely, black screens, scene jumps, and post-production audio. Moreover, the labels assigned via machine-learning models were sometimes not appropriate due to models' accuracies and sometimes ambiguous due to videos containing multiple matching labels when the content includes sounds from multiple labels or when the places could not be classified directly via common sense.

To overcome these issues, it was decided to select clips with the highest match scores of target labels and manually label them for both audio and place categories and erroneous content. In order to do so, a website with a simple user interface (UI) for annotating videos was built. The website allowed master users to enter custom questions and answer options. Also, it allowed the rest of the users to answer those questions per video clip. The question "What is the dominant audio type in the video?" was added with the options "vehicle sound," "music instrument," "human speech," "human vocal," "other," and "noise" to decide the dominant audio type label. The question "What kinds of audio types are in the video?" was added with the same options to check for cases with multiple audio types. Since music instrument and human speech types can intersect when the audio contains human vocal sounds, "human vocal" was also included in options to remove such ambiguous cases. The question "Is the video recorded indoors or outdoors? If outdoors, is the environment natural or human-made?" was added with the options "indoors," "outdoors (natural)," and "outdoors (human-made)" to decide the place types for the video clips. The questions "Does the video contain any black screen?" and "Is there any scene jump in the video?" with options "yes" and "no" were added to filter out erroneous content. Likewise, the question "Does the audio come within the scene of the video, or was it added in post-production?" with options "From scene" and "Post-production" was added to filter out post-production sounds.

Additionally, whether it is easy to localize sound-generating objects and whether there is any text within the clips could come in handy in future analyses since there is some relevant research pointing out that those features play a role in saliency as well (Cerf et al., 2009; Fan et al., 2018; J. Xu et al., 2014). So, the questions "Can you detect locations of the audio sources within the video?" and "Is there any text overlay or an object containing text?" with options "yes" and "no" were added to collect such information along with the annotation process. The architecture details and the design of the website are included in Appendix D, Figures 53-63.

Next up, 20 clips with the highest machine-learning label scores for each category combination, totaling 180 video clips, were added to the system. Then, five researchers labeled videos by answering the questions, and each video was labeled by at least three researchers. Finally, responses were exported, merged, analyzed, and eventually used to filter out videos according to the research needs specified above. Results included at least nine good and one weak match per target category. Of those with more than nine good matches, the nine most seemingly distinct videos were selected to maximize the representativeness of the final choice. In the end, a total of 81 videos, nine good matches from each category, were chosen as the target stimuli of the dataset (see Figure 9). Additionally, one weak or good match per category from the remaining videos was chosen to build a demo dataset of nine videos for testing the experiment setup.

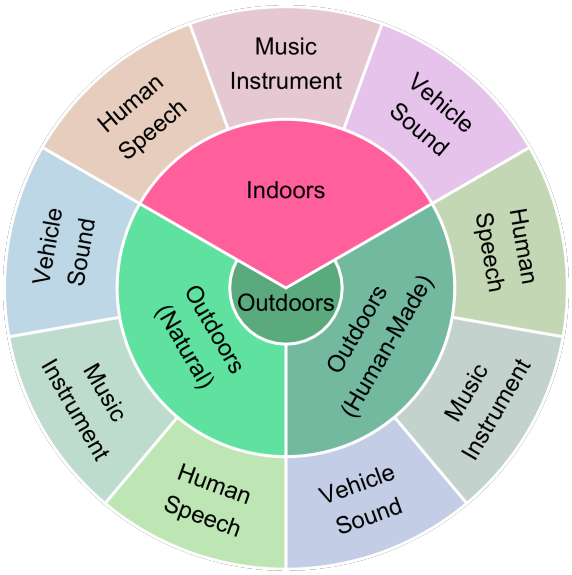


Figure 9. Video clips in YT360-EyeTracking by categories

The newly curated dataset includes nine clips per scene-audio category combination.

All those videos were watched via the VR headset for a final check to see if any of them contained any problems. Doing so revealed that two video clips contained differently formatted ambisonics audio, and one contained a deafening sound. The audio formats of these clips were manually corrected by mapping the audio channels to the correct order, and the amplitude of the deafening one was halved using the FFMPEG program (Tomar, 2006). Then, each chosen video was manipulated via FFMPEG to the needs and encoded in h264 format in one run. These manipulations can be listed as follows: The higher-resolution videos were reduced to 3840x1920 pixels to make them all the same. The videos' frames per second (FPS) ranged from 24 to 30, except for one with 60 FPS. As there is no good way of changing fps without affecting output originality except for selecting one of every x frames, only the framerate of the video with 60 FPS was halved by skipping one of every two frames. That way, all videos had similar fps ranging from 24 to 30. Also, the videos with higher resolutions were reduced to 3840x1920 pixels resolution to

make them all the same. Finally, all six combinations of audio options ambisonics, mono, mute, and video options colored and grayscale were obtained by removing audio and color channels. After all those manipulations were done, all combinations of each video were again watched via the same headset to validate them. To sum up, 486 different stimuli from 81 clips for the main dataset and 54 stimuli from nine clips for the demo dataset were prepared. The thumbnails of all video clips are shown in Appendix E, Figures 64-65.

It is useful to collect additional info per participant in order to both check for the representativeness of the attended participants and investigate potential cofactors in different analyses. Previous VR experiences can influence user behavior during the experiment by reducing the novelty factor of the experience and causing scene exploration strategies to differ. Also, previous video gaming experience may influence users' fixation or exploration patterns as well. Thus, a questionnaire form was created to collect demographic information along with previous VR experience and video gaming experience (Celikcan, Askin, Albayrak, & Capin, 2020; Krüger et al., 2021; Schubert et al., 2015; Shinkle, 2008). This form, both in English and Turkish, can be found in Appendix F and Appendix G.

Along with quantitative data, qualitative data can also provide valuable insights and observations in better understanding user exploration behaviors of scenes (Mishra, Nikam, Vidhyadharan, & Cheruvalath, 2022; Pillai & Verma, 2019). Requesting written text from participants explaining their experiences could lead them to share only a brief summary, whereas they could provide broader explanations verbally. To achieve that goal, it was decided to ask the participants to share their experiences freely and verbally so that the researcher could transcribe them in real-time to be read later.

3.2.2 Participants

At least 10-15 observers are usually recommended for similar eye-tracking studies (Judd et al., 2012; Qiao et al., 2021). Hence, aiming to collect data from 15 participants per condition of each video clip for all six conditions, 90 participants were required. Additionally, it was decided to collect data from 100 participants to be able to deal with potential outliers. In order for the participants to distinguish colors, color-blind people were not targeted. Also, in order for the eye-tracking mechanism to work properly, cross-eyed people were not targeted. Excluding a few outliers where they either mismatched the target criteria or when experimental procedure errors occurred during their data collection, but including participants with partial data collection, a total of 102 participants entered the experiment. There were 47 participants (24 female, 23 male) with prior VR experience and 55 participants (29 female, 26 male) without any prior VR experience. Data was partially collected from 5 participants who either kept their eyes closed for long periods in a few videos or decided to leave the study in half due to motion sickness effects but wanted the data collected until that point to be used in the study. The weighted sample size was 100.6 when weighted with the number of videos watched by each participant. Most of the participants were Boğaziçi University undergraduate students who received 1.5 extra course credits, and the rest were voluntary university students or alumni. All of the observers had normal or corrected-to-normal vision either by wearing contact lenses or glasses. The VR headset has an adjustable space to fit worn glasses. All of the participants were naive to the experiment's purpose except for one undergraduate researcher**.** The study complied with the Boğaziçi University Ethics Coordinating Committee's requirements of research ethics and was approved by them, as shown in Appendix H. All data were collected via randomly assigned identifiers (IDs) to the

participants to assure data anonymity and confidentiality. Also, to avoid recording or storing audio, which could breach confidentiality, the researcher transcribed the verbal qualitative data in real time, and it was only matched with the randomly assigned IDs. Most participants were native Turkish speakers, except for three, who were only proficient in English. The consent form, the questionnaire form, and the experimental instructions were given in either Turkish or English to the participants' preference.

3.2.3 Stimulus and apparatus

In the experiment, participants wore an HTC Vive Pro Eye VR headset, allowing them to watch and hear the stimuli while their eyes were tracked. The VR headset was connected to a Dell Precision 5820 Tower Workstation Computer (Intel Xeon(R) W-2235 CPU, 16GB DDR4 RAM, NVIDIA Quadro RTX 5000 16GB graphics card) via a DisplayPort and a USB 3.0 cable. During the experiment, the participants did not use any controller; however, before each trial, the participants used the headset controller stick to navigate and complete the native eye-tracking calibration program of the headset. The headset's cable was attached to an extensible and adjustable mechanism fixed to the ceiling with a hook, allowing the wearers to walk and rotate freely without any entanglement or restraint of the cable. Appendix I shows photos of the experiment environment. Figure 10 shows an example researcher testing the experiment environment.



Figure 10. A researcher testing the experiment setup via watching a demo video The photo is published with prior consent of the researcher.

The experiment environment, essentially as a panoramic video player with eye-tracking data recording, was coded in Unity 2020 3d Game Engine with C# programming language. Randomization of stimuli within trials was achieved by creating precomputed lists via a Python code, and then both the prepared video clips, audio files, and lists were fed into the Unity program. Stimuli (video clips and audio files) used were prepared beforehand as explained in detail in the previous chapter.

3.2.4 Design and procedure

In order to decide the exact trial duration and procedure, four researchers entered a demo experiment and watched the nine demo videos. That way, the feeling of motion sickness, common in VR, was experienced beforehand, as well as finding out that at least 5 seconds of a black screen was needed between each video clip to feel fully ready for the next clip, leading to a decision of including 10 seconds of black-screen

stimuli between each video clip to be safe. Also, 15-20 minutes per trial was decided to prevent participants from getting too tired of the VR environment over prolonged use. Then, a detailed guideline of the exact procedure, including the prompts of the experiment, was prepared to be followed by the researcher through the procedure. Regarding the actual experiment procedure, all participants were asked about their auditory and visual problems and how they were corrected to normal. Then, all participants gave informed consent to participate in the study after reading the consent form, which explained the procedure and potential side effects of the experiment. The main unwanted side-effect was motion sickness, which is common in VR experiments and was observed by researchers to be more apparent when the stimuli included a video from an unfixed camera. After the participants read the consent form, the researcher also verbally explained how motion sickness could happen to ensure it was understood as clearly as possible, reminding them that the experiment could be stopped anytime they felt uncomfortable or for any reason. Next, the researcher verbally explained the procedure following the prompt of the experiment guidelines as shown in both English and Turkish in Appendix J and Appendix K. The participants were told that they were going to watch 360° panorama videos while their eyes were tracked, and they would be standing up while rotating freely during the experiments. Additionally, the participants were informed that some videos were in grayscale, but some were in color, and some videos had audio, but some were mute to prevent them from questioning whether there was any device malfunction and to minimize any additional surprise effects that could interfere with the participant's attention during the experiment. The researcher also explained the durations of each trial, the video clip, and the black screen breaks between each video clip. The content of the videos was not explicitly explained prior

to the experiment, but the participants were informed that there were no clips designed to be scary, and none of them included jump-scare content (such as objects or entities appearing out of nowhere to scare the viewer), but the sudden start of each next video scene after each black screen could be slightly unsettling. Such a brief was included to avoid any confusion and ensure the participants felt safe to explore the content without questioning whether any horror content could pop up. The rest of the prompt follows the same reasonings: always minimize any surprise events, make the experiment procedure as clear as possible to the user, and preserve both the safety and the comfort of the participants. Finally, the participant was shown which buttons of the controller they would use during the calibration process.

The experiment followed a cross-design such that each participant watched all 81 videos in random order and under one random condition per video. The randomization process also controlled each participant to watch videos with each condition as equal as possible from each semantic video category. This design allowed the inclusion of partial data collection in the analysis as it was thought that partial data collection could be common if many participants decided to stop the experiment due to experiencing motion sickness effects. Interestingly, the motion sickness effects were much less common in participants than the researchers in the nine-video demo experiment, probably due to the purposefully designed prompts minimizing surprise effects or clear explanations of the motion sickness experience. Alternatively, it may be due to the small sample size of the researchers. Nevertheless, the few partially collected data were still considered valid and used in the analysis as per the original design.

The experiment was conducted in three blocks, with an average of a five-minute break between each block. In each trial, participants were presented with 27

stimuli (video clips and the corresponding audio) of 30 seconds with a 10-second black-screen environment in between each stimulus (see Figure 11). Also, each participant started watching each stimulus from the same starting angle to control any effects of starting position on exploration (This behavior was implemented in the Unity environment programmatically). Prior to each trial, the researcher put on the headset to the participant, and then the native calibration process of the headset was followed to make sure the headset fit properly, and its eye-tracking feature was calibrated correctly. Then, the participant was moved to the center of the room and stood up while being allowed to rotate freely during the trial. After a beep sound test, the trial block was started, and the target stimuli were presented to the participant, as their eye-tracking data was recorded in a file. The researcher avoided talking or making sounds to prevent disturbing the participant during that phase. One exception is if a participant asked whether there was any problem with the device when they saw a grayscale or mute video for the first time, the researcher reminded the participant that some videos are grayscale, some are colored, some have audio, some are mute in the next black screen. Another is if the participant moved very close to room boundaries, the researcher warned the participants to take a few steps back in the next black screen. Finally, if a participant wanted the experiment to be stopped, the researcher immediately stopped the experiment and noted down the stopped video index. A few participants who stopped the experiment wanted to continue the experiment after resting, and they were allowed to do so by starting from the next video after the stopped one. After a trial block, the researcher helped remove the headset, and the participant sat down and rested until they felt comfortable to continue the experiment. During these resting periods, they were asked to optionally describe their experiences freely, and the researcher noted those down.

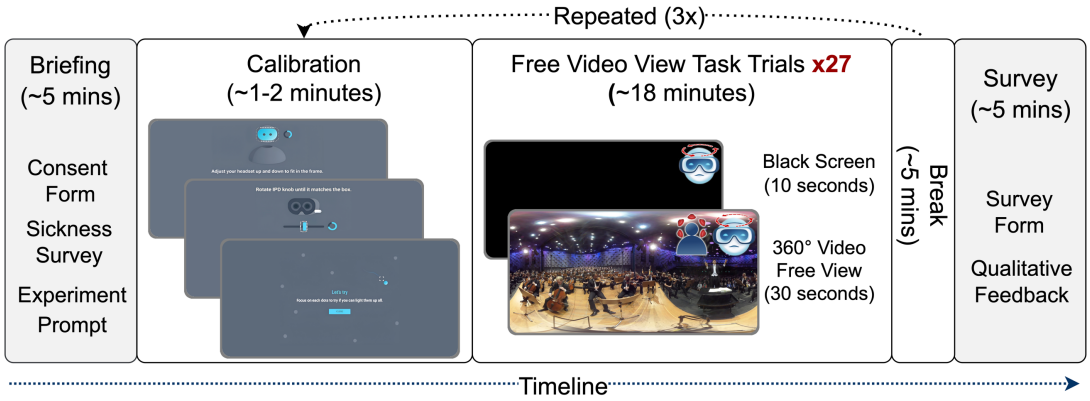


Figure 11. Outline of the experiment blocks and trials

Participants completed a free-viewing task across three sessions with 81 videos, incorporating a five-minute break between sessions to lessen fatigue. Sessions started with eye calibration, followed by 27 audio-visual stimuli, each being a 30-second video clip in one of three pseudo-randomized audio (mute, mono, ambisonics) and color (grayscale, colored) conditions, separated by 10-second black screen intervals to reduce carry-over effects.

After all trials were complete, the participants were again asked to optionally provide feedback or describe their experiences freely, and the researcher again noted those down. This time, after all the stimuli had been watched and the eye-tracking data had already been collected, and after noting down the free explanations, the researcher asked several questions of interest about the participants' experiences, reminding them that this part was also optional, and the responses were noted down. Finally, they were asked to fill out the questionnaire form.

The tracking data which includes the gaze directions and pupil diameters of each eye as well as the combined gaze direction and head orientation was saved to a JSON formatted (JSON Org, 1999) text file per trial per participant in real time along with the timestamps of each sampled data. Also, the exact starting and ending timestamps of each stimulus were saved in comma-separated values format to a text file per trial per participant in real time. That way, the contents of both files could be

matched post-experiment while avoiding any unexpected data-collection or corruption problems due to matching in real time.

3.3 Data processing

The initial data cleaning and pre-processing procedures were implemented in the Python (Python Software Foundation, 2001) programming language, consisting of fixation extraction, fixation map generation, and similarity metric calculations. This was followed by an investigation in R software (R Core Team, 2023), including a factor analysis to reduce multiple similarity metric scores into a single factor for further analysis.

3.3.1 Extracting fixations

First, using the saved data from the experiments, the start and stop times of each stimulus per trial per participant were matched with the corresponding tracking data to obtain gaze and head direction data for each stimulus per participant. To extract fixations from 360° eye-tracking data, the Identification by Dispersion-Threshold algorithm (I-DT), as explained by Agtzidis and Dorr (2019), was implemented. Using that algorithm, consecutive sequences of raw gaze data with a maximum dispersion of 1.5 visual angles and with a minimum duration of 0.1 seconds were marked as fixations, saving the start, end, duration, and mean direction of each fixation to comma-separated values formatted text files per stimulus per participant. The maximum dispersion value was chosen by considering both the suggested thresholds for I-DT in various research and the fact that the accuracy of the eye-tracking device is in the range of 0.5-1.1 degrees such that the chosen threshold was

higher than the device accuracy but not too high (Blignaut, 2009; E. David et al., 2023; Salvucci & Goldberg, 2000; Schuetz & Fiehler, 2022).

3.3.2 Fixation map generation

All further calculations are done using extracted fixations per person per stimuli. First, in order to be able to work with algorithms defined for equirectangular images, a Python script was created to convert a list of fixations to frames of fixation maps in equirectangular format. Doing so is tricky for 360° content as the algorithm also needs to account for the dispersion on poles in the equirectangular format. The algorithm was implemented by following the modified gaussian algorithm suggested by John, Raiturkar, Le Meur, and Jain (2018) for that purpose but running it in full resolution via several mathematical simplifications that yield the exact same results but much faster. In the original suggestion, the author did not optimize the proposed method; hence, in order to make it computationally feasible, they suggested downscaling before calculation and upscaling afterward, which only yielded an approximation of the full resolution result. Like the original algorithm, the implemented version also takes a sigma parameter for the Gaussian kernel, which is commonly chosen in the range of 1-2 visual degrees (E. David et al., 2023; Gutierrez et al., 2018; John et al., 2018; Judd et al., 2012; Yi Zhang et al., 2020). Next, another Python script that converts these fixation maps to heatmaps overlayed over videos using OpenCV (Bradski, 2000) is written and optimized to handle large data streams. Using it, heatmap-overlaid videos are generated per stimuli so that the researchers can visually inspect the fixation behavior of observers. During these visualizations, the Gaussian kernel was chosen as 1.6 visual degrees considering eye-tracker accuracy and average fixation dispersion, and the resolution was set to 3840x1920,

matching the original resolution of the videos to be overlaid, as well as aggregating fixations within 0.1 seconds forward and backward of each frame, i.e., a window duration of 0.2 seconds. To ease visual inspection, three color channels (red, green, blue) were used to denote observers watching the same video under different audio conditions. Likewise, separate heatmap overlaid videos were generated for colored and grayscale video conditions. Sample frames from the middle of all these heatmap overlayed videos can be seen in Appendix L, Figure 66-74.

3.3.3 Similarity metric calculations

Subsequently, the similarity metric calculation scripts of the Salient360 toolbox (E. David et al., 2023; Gutierrez et al., 2018) were modified to work with our codebase and optimized to handle large data streams via GPU computing. The similarity scores calculated are AUC Borji, AUC Judd, CC, NSS, SIM, and negative KLD, each having slightly different properties but all used for measuring similarity (Djilali, McGuinness, & O'Connor, 2021; Bylinskii et al., 2019). These are commonly used in similar research and supported in the Salient360 toolbox. During the calculation, when converting fixations to fixation maps, the sigma parameter of our Gaussian kernel was chosen as 2.0 visual degrees, the intermediary maps are generated in 1000x500 resolution, and the fixations are aggregated within window durations of 2.0 seconds during calculations, the same values as in the Salient360 toolbox to comply with other research. Next, two separate approaches were taken to prepare the data to be analyzed. Firstly, in order to reproduce the results of the preliminary research, a similar methodology of comparing aggregated fixations of each stimulus was compared across different conditions. A Python script to calculate pairwise similarity scores for each combination of conditions per video per two-second time

block was implemented and run, saving the calculation results in comma-separated values format to a text file for the actual statistical analysis. Secondly, inter-observer consistency scores were calculated by comparing the fixations of each participant per stimuli per two-second time block with the fixation map obtained from the rest of the participants observing the same stimuli, as described in Bylinskii et al. (2019) commonly called the leave-one-out approach. Again, the calculation results were saved in comma-separated values format to a text file for the actual statistical analysis.

3.3.4 Data inspection

The files obtained in previous chapters consisting of pairwise comparison scores of each condition and inter-group consistency scores, along with manually entered questionnaire answers of each participant and labels of each video, were used in the analysis. All these data are imported and combined in R language. (R Core Team, 2018) Also, all the following statistical analyses are conducted in R language. Demographic information and the questionnaire were only used to check whether the participants had any unexpected distribution, revealing that the participants' age, weight, and height were distributed normally in each gender. The comparison scores where the condition, colored and with spatial audio, was compared with the condition, colored with mute audio, were analyzed with a 3 x 3 (Place type: indoors, outdoors natural, outdoors human-made. Audio type: music instrument, human speech, vehicle sound) repeated measures MANOVA analysis to reproduce results from the preliminary research. The rest of the analyses were done by using inter-observer consistency scores. Although inter-observer consistency scores contain data per participant per video, which could be modeled as crossed random effects in linear

mixed model-based analyses, doing so invalidated the independence of observations assumption due to the nature of how those scores are calculated. These scores were calculated by comparing each observer with the rest of the observers, thus making the scores dependent on other observers. To overcome that, inter-observer consistency scores per stimuli are averaged across observers, and only the random effects of different video clips and nested random effects of time blocks are modeled in the analyses. Audio (mute, mono, or spatial) and video (grayscale or colored) conditions are used as repeated measure factors of each video clip time block, and audio type (music instrument, human speech, and vehicle sound) and place type (indoors, outdoors natural, outdoors human-made) are used as independent measure factors of each video clip time block.

3.3.5 Cut-off initial data

As all participants start observing from the same position, and they need some time to adapt to their new environment in the video, it is common to cut off a fixed portion of the collected data from the beginning. (David et al., 2018; Xu & Tian, 2020) A visual assessment of data revealed that it took six seconds for observer fixations to diverge in all similarity metrics as shown in Figure 12. Therefore, it was decided to filter out data from the first six seconds of each video.

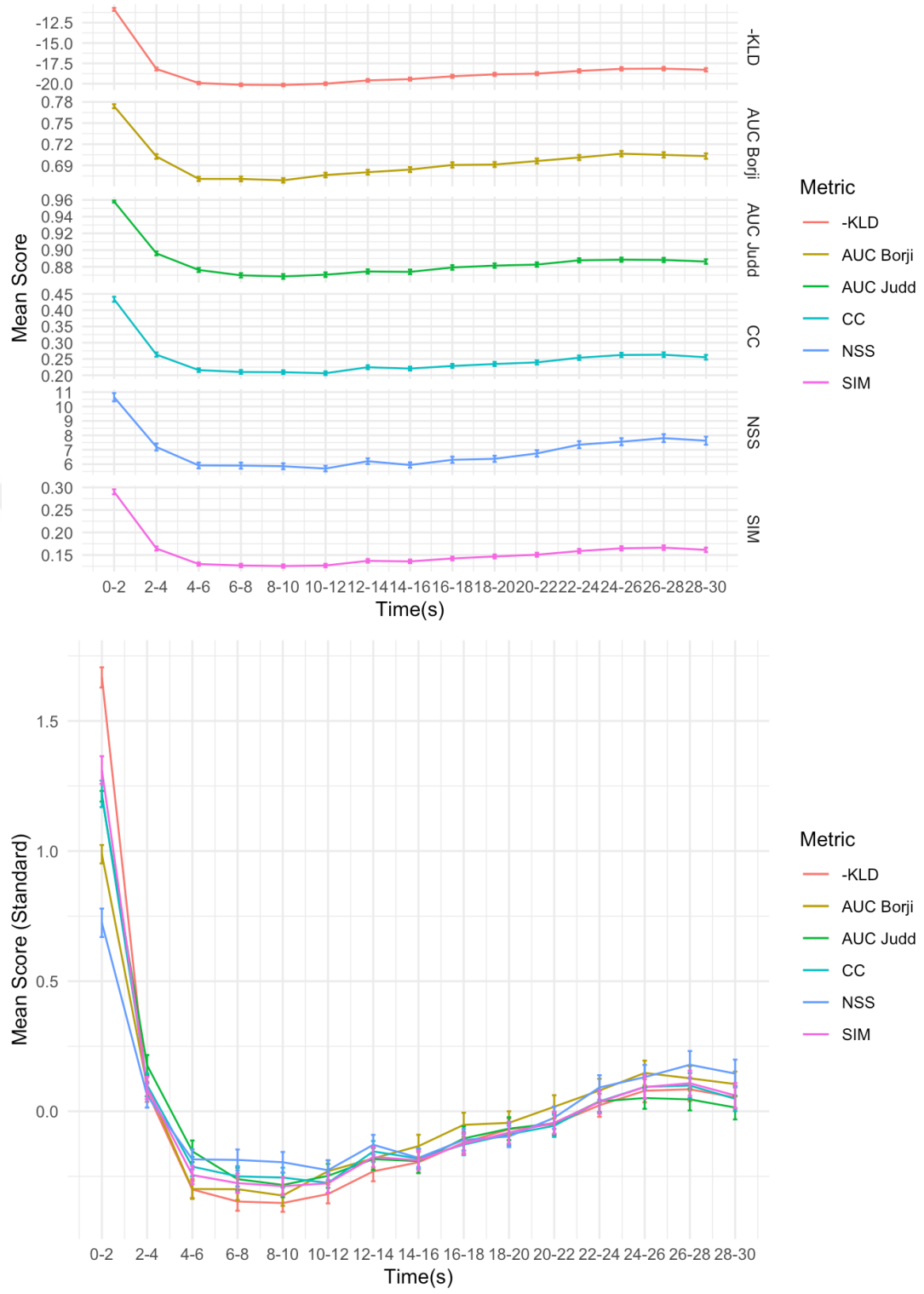


Figure 12. Distribution of all raw and standardized consistency scores over time

Both plots shows a sharp decrease in all metrics until sixth second.

3.3.6 Factor analysis

To reduce six similarity metrics corresponding to the inter-observer consistency scores of fixation location of observers on videos to a single variable, a factor analysis approach was employed. Since KLD measures dissimilarity instead of similarity, its scores are negated for it to comply with other metrics. Furthermore, standardized scores for each metric were calculated.

To begin, a correlation analysis between these metrics was done, revealing high correlations amongst all metrics (> 0.30) along with high multicollinearity in several ($> .90$). All correlations among metrics were positive and statistically significant, ranging from .46 to .98. (See Figure 13) The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy and Bartlett's test of sphericity were used to evaluate the appropriateness of the correlation matrix for factor analysis. The KMO measure of sampling adequacy was .83, exceeding the recommended value of .60 (Kaiser, 1974) and indicating that the correlation matrix was suitable for factor analysis. Bartlett's test of sphericity was significant, $\chi^2(5) = 146050$, $p < .001$, supporting the presence of meaningful associations among variables. These results confirm that the metrics can be combined to create a single factor.

To start an exploratory factor analysis, a principal component analysis was conducted using the psych (Revelle, 2024) package in R and revealed a single-component solution accounting for 81% of the total variance. All loadings are presented in Table 5. The component ($eigenvalue = 4.9$) was labeled "consistency," and it had loadings from all items. The second potential component had less than 1.0 eigenvalue ($eigenvalue = .68$, 11% of variance) and only had considerable loadings (> .30) from AUC Judd and NSS. Therefore, the scree plot in Figure 14 suggested retaining only the first component.

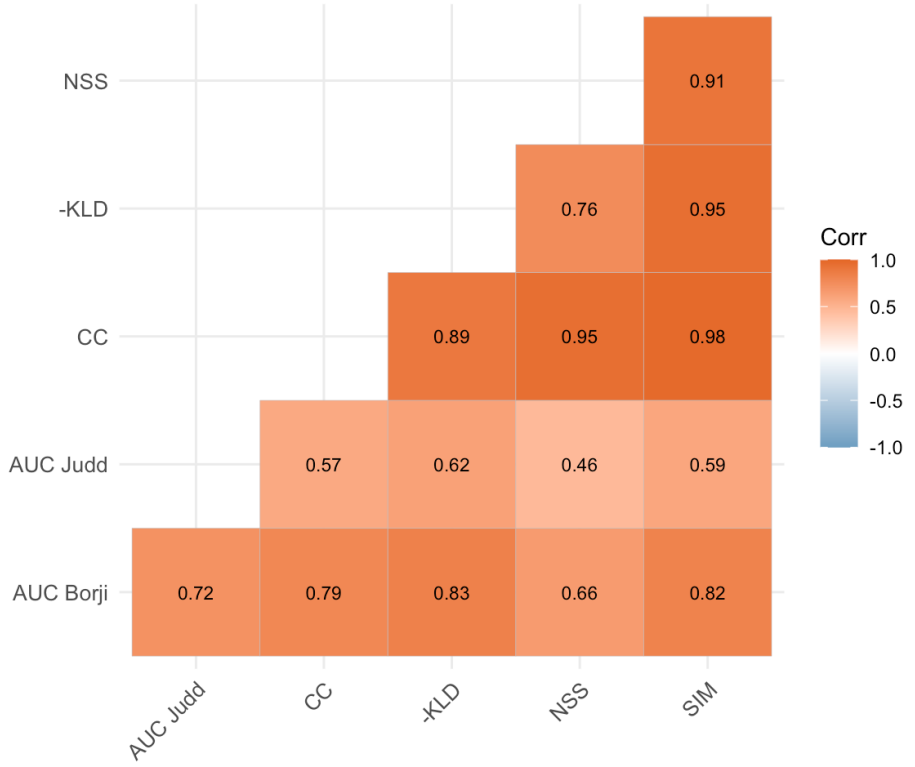


Figure 13. Correlation scores between similarity metrics for consistency

The correlation matrix shows Pearson correlation. Every similarity metric had a high correlation (> .30) with each other. SIM, CC, and NSS showed multicollinearity between each other (> .90). Additionally, SIM and -KLD showed a separate multicollinearity with each other (> .90).

Table 5. Principal Component Analysis Loadings Table

	PC1	PC2	PC3	PC4	PC5	PC6
AUC Borji	0.89071	0.25799	-0.26373	0.26540	0.00821	0.00080
AUC Judd	0.71341	0.64852	0.25828	-0.06136	0.00007	-0.00162
CC	0.96789	-0.21400	0.06970	0.00595	-0.09898	0.052024
-KLD	0.94265	-0.00801	-0.22226	-0.24045	0.05597	0.03149
NSS	0.88635	-0.36021	0.25488	0.11542	0.07955	0.00436
SIM	0.97897	-0.16191	-0.03393	-0.07563	-0.03558	-0.08525

Note: It can be seen that all variables are strongly loaded on the first component. AUC Judd and NSS cross-loaded on the second component as well (> .30). No other items loaded strongly in any other component.

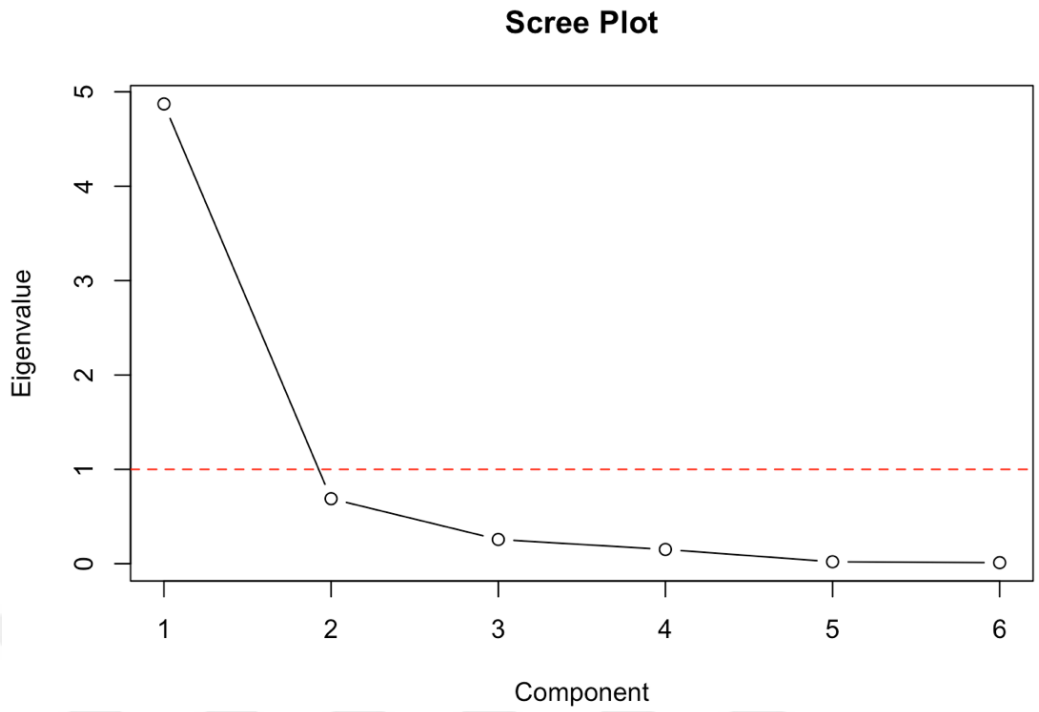


Figure 14. Scree plot of the principal component analysis on consistency scores

It can be seen that only the first component had an eigenvalue greater than 1.0. Therefore, the analysis suggested retaining only the first component.

Following that, an exploratory factor analysis using maximum likelihood (ML) extraction for a single factor was conducted again using the psych package in R. No rotation was applied as only one factor was extracted. The single factor accounted for 78% of the total variance in the measures. The loadings are presented in Table 6. All six measures loaded strongly on the single factor, with standardized loadings ranging from .64 to 1.00. Communalities, representing the proportion of each measure's variance explained by the factor, were high (range: .41 to .99), indicating that the factor captured a substantial amount of the measures' variance. The goodness-of-fit indices suggested mixed results for the one-factor model. The Tucker-Lewis index (TLI) was .684, below the recommended threshold of .90 for an acceptable fit. The root mean square error of approximation (RMSEA) was .468, higher than the cut-off of .06 for good fit. This result suggests that although it was

able to find a good factor, the data cannot be fit properly via factor analysis without dealing with multicollinearity.

Table 6. Exploratory Factor Analysis Loadings and Communality Values

Measure	Factor Loading	Communality
AUC Borji	.86	.73
AUC Judd	.64	.41
CC	.98	.96
-KLD	.94	.88
NSS	.86	.73
SIM	1.00	.99

Note: Results indicate that all variables are strongly loaded on the single factor.

In order to deal with multicollinearity prior to a confirmatory factor analysis (CFA), variables with high multicollinearity, CC, NSS, and SIM (those with correlation values $> .90$) were merged by taking their average. Afterward, a single-factor CFA model was specified to assess the unidimensional structure of the consistency measures. A latent factor, consistency score, was hypothesized to cause four observed variables: AUC Borji, AUC Judd, -KLD, and the average of the rest (CC, NSS, SIM). The CFA model was estimated using ML estimation in the lavaan (Rosseel, 2012) package in R. The single-factor CFA model showed a mixed fit to the data. The comparative fit index was .943, indicating acceptable fit ($CFI > .90$). However, the TLI was .829, slightly below the recommended threshold of .90. The RMSEA was .313 (90% CI [.298, .329]), higher than the cut-off of .06 for good fit. The standardized root mean square residual (SRMR) was .048, below the threshold of .08 for an acceptable fit. All four observed variables loaded significantly on the consistency factor ($p < .001$). The standardized factor loadings were high, ranging from .664 to .966 (see Table 7). The consistency factor explained a large proportion of variance in each variable.

Table 7. Confirmatory Factor Analysis Loadings and R^2 Values

Variable	Standardized Factor Loading	R^2
CC+NSS+SIM	.905	.820
AUC Borji	.869	.755
AUC Judd	.664	.441
-KLD	.966	.932

Note: All four variables are strongly loaded on the factor and explain a huge proportion of the variance.

CFA resulted in a good single factor, but as an alternative approach, Cronbach's alpha was calculated to assess the internal consistency reliability of the six measures: AUC Borji, AUC Judd, CC, -KLD, NSS, and SIM. Cronbach's alpha provides an estimate of the average inter-item correlation and the extent to which the items measure a common underlying construct. The overall Cronbach's alpha for the six measures was .95, indicating excellent internal consistency reliability. This suggests that the measures are assessing a common underlying construct and can be reliably combined into a composite score. Therefore, taking the mean of these standardized metric scores is also a valid approach.

Finally, estimated consistency factor results and the mean standardized metric scores are compared via a correlation analysis to see if they differ. Pearson correlation coefficients were calculated between these two scores. The correlation analysis revealed a very strong positive relationship between the estimated consistency factor scores and the mean standardized metric scores, $r = .98$, $p < .001$. This suggests that both approaches are capturing the same underlying construct. Therefore, for easier interpretability, it was decided to use mean standardized metric scores as the composite consistency scores in the rest of the analyses.

3.4 Results

3.4.1 Reproducing preliminary experiment results

In order to check if the results from the preliminary experiment can be reproduced a similar analysis is conducted over comparison scores of ambisonics and mute audio modes. Also, outdoor natural and outdoor human-made place categories were merged under outdoors, along with filtering out grayscale videos to comply with the preliminary experiment data structure. Finally, different similarity metric scores were averaged over time in a manner similar to the preliminary experiment analyses.

Due to high multicollinearity, MANOVA analysis was skipped, and since the place category contained two levels, two-tailed t-tests were employed to see if the similarity of fixated locations in the mute and ambisonics condition varied across indoor and outdoor places. No results were significant (all $p > 0.05$), indicating no such effect was shown, unlike the marginal significant effect in the preliminary experiment. Moreover, when box plots in Figure 15 were assessed visually, the mean similarity scores were slightly lower in outdoors compared to indoors according to all metrics, opposite to the trend shown in the preliminary experiment.

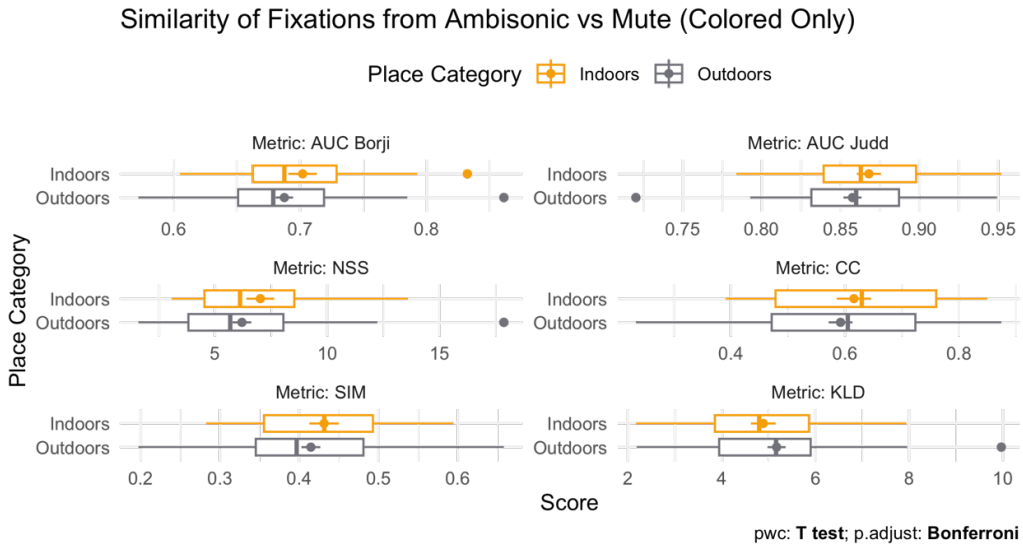


Figure 15. Ambisonics vs mute fixation similarities across place categories

Consequently, separate one-way ANOVA analyses are conducted per similarity metric to investigate the effects of audio categories. Results revealed a significant effect of audio category on AUC Borji, $F(2,78) = 3.87, p = 0.025, \eta^2 = 0.090$, AUC Judd, $F(2, 78) = 4.59, p = 0.013, \eta^2 = 0.105$, KLD, $F(2,78) = 4.10, p = 0.020, \eta^2 = 0.095$, SIM, $F(2,78) = 3.87, p = 0.025, \eta^2 = 0.090$, but not for CC, $F(2,78) = 1.15, p = 0.323, \eta^2 = 0.029$, and NSS, $F(2,78) = 1.08, p = 0.346, \eta^2 = 0.027$. Due to having considered SIM only in the preliminary experiment and the factor analyses on consistency scores showing that SIM was the metric with the highest loading on tested factors and components, post-hoc tests were only applied to SIM scores. Pairwise t-tests with Bonferroni correction showed a significant effect ($p = 0.038$) between music instrument ($M = 0.39, SD = 0.06$) and vehicle sound ($M = 0.46, SD = 0.11$) audio categories such that the similarity scores changed more profoundly in music instrument conditions when fixations of mute and ambisonics conditions were compared. For all metrics, by visual assessment of the box plots in Figure 16, it can be seen that a trend of $music\ instrument < human\ speech < vehicle\ sound$ occurs with only a significant difference between music instrument and vehicle sounds. Although the audio categories are not exactly the same in the preliminary experiment, both results follow the same trend: The fixation maps change the most across audio modes if the video contains music sounds rather than any other sounds. As a reminder, a trend of $music < human < others < animal$ for VGGSound labels and a trend of $music < others < human < animal$ for Audioset labels was apparent, with only significant results when animal and music categories were compared in the preliminary experiment.

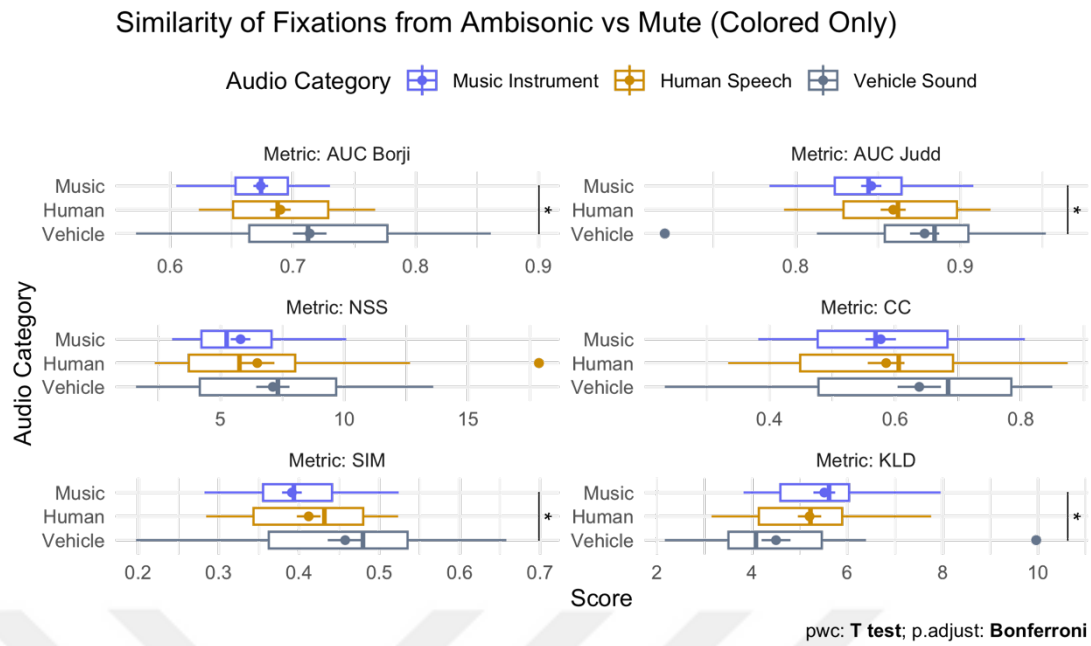


Figure 16. Ambisonics vs mute fixation similarities across audio categories

3.4.2 Simple linear regression of consistency scores over time

A linear regression analysis was conducted to assess the consistency scores over time. The results indicated that the video timestamp significantly predicted consistency, $F(1,5830) = 155.8$, $p < 0.0001$. The regression model explained 2.6% of the variance in consistency ($R^2 = 0.026$). The regression equation was determined as $consistency = -0.377 + 0.02096 \times timestamp$, suggesting inter-subject consistency of eye fixations increases by approximately 0.021 units per second in video clips. Figure 17 depicts mean consistency scores per timestamp and the fitted regression line. Since our research questions are not based on time but rather on the overall effects of video clips, and the distribution over time does not indicate any anomalies, the data was reduced by replacing consistency scores with mean scores over time for the rest of the analyses.

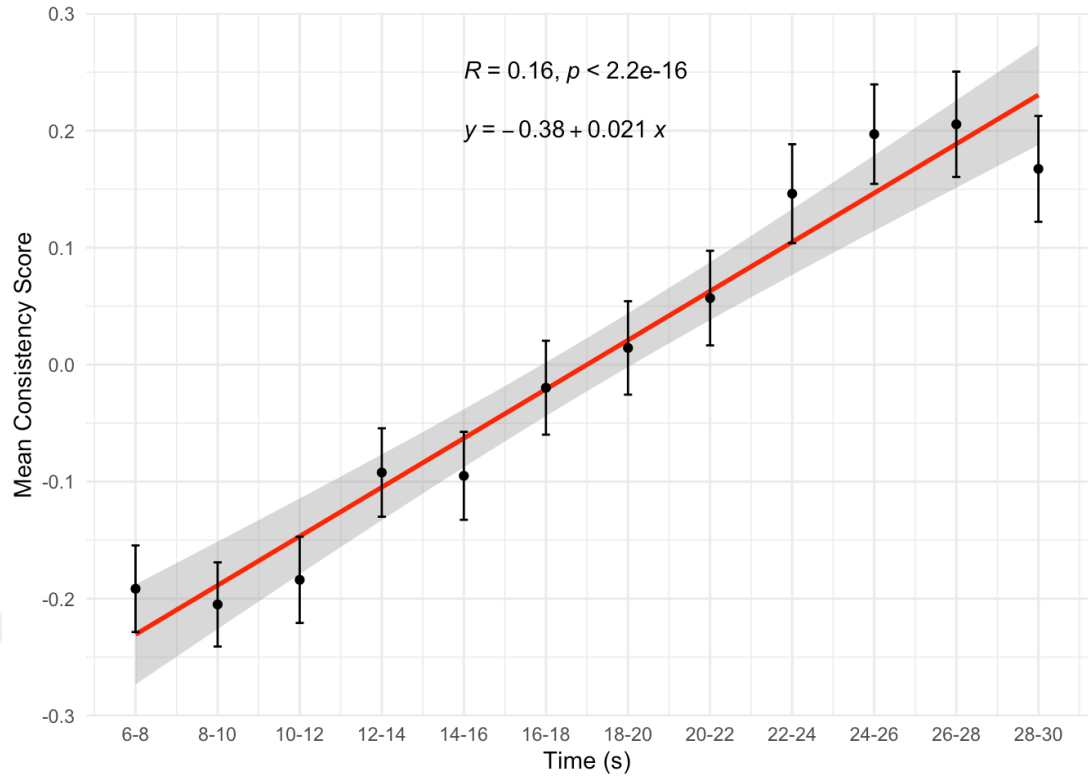


Figure 17. Mean consistency scores and the fitted regression line plotted over time

3.4.3 Mixed ANOVA analysis on consistency scores

A 3 x 3 x 3 x 2 mixed ANOVA was conducted to examine the effects of audio category, place category, audio modality, and color modality on consistency scores. The analysis revealed that the audio category had a significant main effect on consistency, $F(2, 72) = 4.56$, $p = .014$, $\eta^2 = 0.112$. This indicates that the audio content (human speech, music instrument, and vehicle sound) had a significant impact on consistency scores with a medium effect size. The main effect of the audio modality was also significant, $F(1.49, 107.51) = 15.78$, $p < .001$, $\eta^2 = 0.180$, using the Greenhouse-Geisser correction for sphericity. This suggests that the audio modality (ambisonics, mono, or mute) significantly influenced consistency scores with a large effect size. However, there was no significant main

effect of place category, $F(2, 72) = 0.46$, $p = .636$, $\eta^2 = 0.012$, or color modality, $F(1,\ 72) = 1.17$, $p = .283$, $\eta^2 = 0.016$.

The interaction of the audio category and the place category was significant, $F(4, 72) = 3.35$, $p = .014$, $\eta^2 = 0.157$, indicating that the effect of the audio category on consistency varied depending on the place category, with a large effect size. There was also a significant interaction effect between the audio category and the audio modality, $F(2.99, 107.51) = 2.83$, $p = .042$, $\eta^2 = 0.073$, using the Greenhouse-Geisser correction. This suggests that the impact of the audio category on consistency was moderated by the audio output option, with a medium effect size. No other interaction effects reached statistical significance (all $p > .05$).

However, Box's M-test indicated that the assumption of homogeneity of covariance matrices was severely violated ($p < .0001$). Thus, these results are not reliable by themselves. We suspected that this issue was raised from the relation of audio category and audio modes and visually confirmed that consistency scores were distributed differently across other factors per audio category. The analyses were continued by conducting separate ANOVA analyses per audio category. A variety of complementary plots are generated and put in Appendix M, Figures 75-84.

3.4.4 Mixed ANOVA analyses on consistency per audio category

Three 3 x 3 x 2 mixed ANOVA were conducted (once per audio category). Box's M-tests showed that the assumption of homogeneity of covariance matrices was not strongly violated (all $p > 0.001$) in either music instrument ($p = .060$), vehicle sound ($p = .087$), or human speech ($p = .015$). The results revealed a strongly significant main effect of audio mode on consistency scores for music instrument, $F(1.59, 38.3) = 12.90$, $p = .0004$, $\eta^2 = .35$, and human speech, $F(1.23, 29.5) =$

7.62, $p = .021$, $\eta^2 = .24$, after Bonferroni correction. There was a significant main effect of place category on consistency scores for music instrument, $F(2, 24) = 8.83$, $p = .003$, $\eta^2 = .42$, after Bonferroni correction. There were no other significant main or interaction effects in any of the audio categories. (all $p > 0.05$)

Follow-up post-hoc two-tailed paired t-tests with Bonferroni adjustment were conducted to investigate the effect of audio conditions. Within the music instrument audio category, the tests revealed that the consistency scores were significantly lower in the mute audio condition than in the mono ($p = .0003$) and ambisonics ($p = .001$) audio conditions. Within the human speech audio category, the tests revealed that the consistency scores were significantly lower in the mute audio condition than in the mono ($p = .001$) and ambisonics ($p = .003$) audio conditions. However, there was no significant difference in consistency scores between mono and ambisonics audio modes in either audio category (all $p > 0.05$). Since no significant effects were found in vehicle sounds, no follow-up tests were done on that audio category. Figure 18 shows plots corresponding to these tests.

Furthermore, the significant main effect of place type on consistency scores within the music instrument category was followed up via post-hoc two-tailed t-tests with Bonferroni adjustment. Figure 19 shows plots corresponding to these tests. The tests revealed that consistency scores in the indoors category were significantly lower than in the outdoors(human-made) ($p = .0009$) and outdoors(natural) ($p < .0001$) categories. Additionally, the consistency scores were significantly lower in the outdoors(human-made) category than in the outdoors(natural) category ($p < .0001$). Therefore, it is concluded that the consistency scores have a significant ordering of $indoors < outdoors(human\ made) < outdoors(natural)$ across place categories within the music instrument audio category,

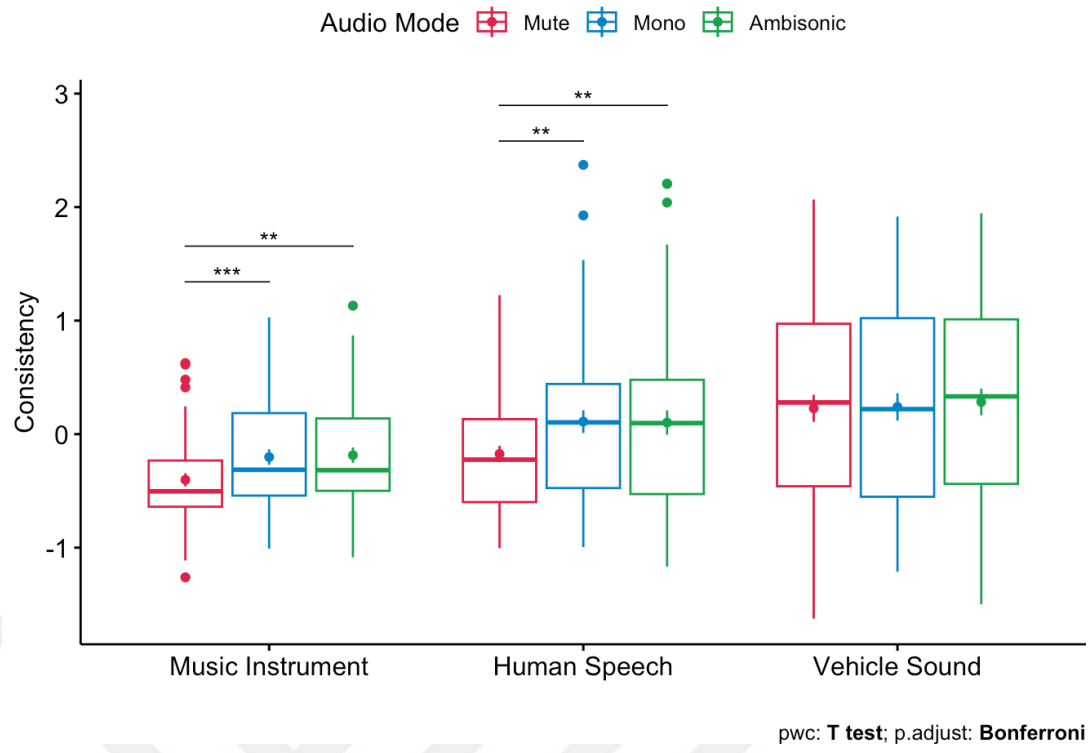


Figure 18. Boxplots of consistency scores across audio categories and audio modes

Significant differences between audio modes in each audio category are shown as "*" for $p < .05$, "**" for $p < .01$, "***" for $p < 0.001$, "****" for $p < 0.0001$.

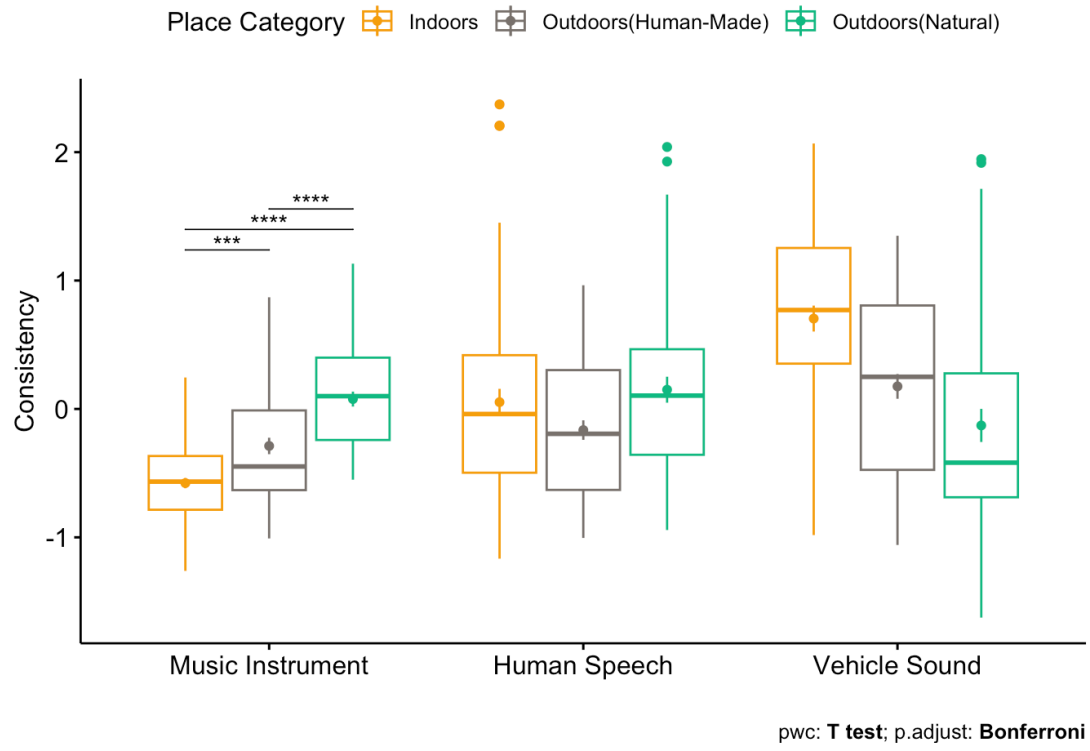


Figure 19. Boxplots of consistency scores across audio and place categories

3.5 Discussion

The results indicate that audio-visual saliency in videos is strongly regulated by audio content but does not depend much on color during a free-view task. Also, the findings hint that although effects arising from audio and place category combinations might be due to the correlation with other semantic content of the videos, the interaction of audio category and audio modulation results suggest that the semantic content of the audio itself is important in determining saliency.

3.5.1 Effect of time

As per expected, inter-subject consistency scores started very high (as all participants started from the same looking direction, their eye fixations were very close), then rapidly decreased as the fixated locations diverged somewhat randomly (initial adaptation behavior), then slowly increased as fixated regions slowly converged via the regulation of actual video content. As common with other research, it takes about 5-10 seconds for users to adapt to the current scene, which was also apparent in the first six seconds in the created dataset. Therefore, unless one studies explicitly the initial adaptation and exploration behavior, it is a good idea to remove data from this early adaptation stage when studying the effects of a scene content, as per our approach. Furthermore, 30 seconds per video clip was confirmed as a reasonable choice since it is longer than an initial cut-off of the early adaptation stage, and no anomalies were seen in the linear regression results after this period. Having no prior beliefs about the specific trends across fixations other than a slow convergence over time, it was interesting to see that this trend fits very nicely to a linear estimate, but longer content is needed to confirm how long it takes for the consistency scores to converge fully.

3.5.2 Comparison with the preliminary experiment results

During the data preparation stage, it was realized that the original uncurated YT360 dataset (Morgado et al., 2020) contained tons of similar content and was biased toward what kind of panoramic content was available as videos on the YouTube platform. A similar issue was observed in the dataset used in the preliminary experiment, the AOD60K dataset (Yi Zhang, 2021), as the videos were distributed equally across neither different place categories nor audio categories. That issue being the primary concern in the preliminary experiment results, the newly created YT360-EyeTracking dataset is curated via equal distributions across several categories in order to allow us to repeat our analysis in a controlled manner, as well as to provide higher generalizability to the dataset for any other research purpose.

Regarding the results of place category, similarity scores, resembling similarity of fixated regions in mute and spatial audio, were lower in indoors and higher in outdoors as a general trend. This trend was only significant for several similarity metrics. When the same analysis is conducted in the newly curated dataset, an opposite trend was shown but with no statistical significance. Therefore, the results of the preliminary experiment can be attributed to either a dataset bias or an unknown correlate of place categories that is not apparent in the new dataset.

On the other hand, when the effects of audio categories on similarity scores were analyzed, similar trends were apparent and included significant effects as well. As shown in both experiments, the music sound category yielded the lowest similarity (the highest difference) across audio conditions. Likewise, videos with human sounds yielded similarity scores higher than music but less than some other categories in both experiments. Other sound categories in both experiments differed and, therefore, were not compared due to the availability of videos and experimental

design. However, as it can be tricky to differentiate sound categories in randomly chosen videos when they contain multiple types of sounds or sounds that can fit into multiple categories (for example, should human singing be considered within the human sounds category or the music category), the current dataset was curated taking special care to choose videos such that each one only incorporated only one type of the studied audio categories. Pairwise comparison results only showed significant effects on similarity scores for several pairs in both experiments, indicating that all audio categories do not segregate fixation behavior well.

A similar issue of multicollinearity across calculated similarity scores was realized, similar to the preliminary experiment. After conducting a detailed factor analysis, it is decided that these scores are highly correlated when calculated for comparing different groups of human eye fixations. Although it might be beneficial to leverage them all when analyzing the performance of a prediction model, they can be safely merged into a single factor when analyzing solely human fixations.

3.5.3 Effects of audio category

Our initial design included selecting three audio categories with different potential behavior characteristics. Human speech conditions could regulate fixated regions both by their semantic content (such as looking at an apple tree when a talking person in the video mentions it) and by containing talking human faces, which are already considered salient by themselves. Therefore, it was hypothesized that consistency scores would be high for videos with human speech and even higher when it is audible). Music instrument conditions could regulate fixated regions by shifting attention towards a played instrument even though instruments themselves are not considered salient in mute conditions. Therefore, it was hypothesized that

consistency scores in the music instrument category would be low for mute but high in mono and even higher for spatial audio conditions. Finally, the vehicle sound category indicated a sound similar to a noise signal but still contained spatial properties and related to real objects but lacked any semantically meaningful information. Therefore, we hypothesized that overall consistency scores in the vehicle sound category would be low and would not vary across mute, mono, or spatial audio conditions. Consistency scores across each audio category were distributed very differently, as this caused very significant violations of the homogeneity of covariances and variances assumptions. The mean consistency scores were ordered as $music\ instrument < human\ speech < vehicle\ sound$; however, each varied completely differently across other factors. Therefore, it could be suggested that the effect of audio categories is complex and could be used to segment data into groups with different characteristics. Our initial design choices for these three audio categories are on par with this result. Music instrument and human speech categories had an overall consistency of low and high, as predicted. Interestingly, the vehicle sound category did not have the lowest consistency score but rather comprised a huge range of scores.

Feedback from participants collected as qualitative data indicated that many participants turned around less and commonly maintained their direction when the camera itself was moving when the video was recorded, suggesting that it felt more natural and moving against the camera movement was disturbing. As it turns out, vehicle sound contained many, although not all, videos in which the camera was placed on a moving vehicle (such as a car, a plane, or a pedestrian walking near a road), causing them to include not only movement but also the fast movement of the camera. Therefore, videos with high consistency scores in the vehicle sound category

could indicate participants consistently looking forward while maintaining their direction compliant to the movement of the camera as a new behavior, whereas videos with low consistency scores could indicate scenes with a fixed-positioned camera compliant to our initial expectation.

3.5.4 Effects of color and audio modes.

Our initial prediction was that reducing modalities would decrease the overall saliency and, therefore, exploration regulated by the video content and promote random exploration. When inter-subject consistency scores corresponding to the consistency of fixated locations within a condition group (grayscale vs. colored, and mute vs. mono vs. ambisonics audio) are analyzed, no significant effect of color condition (grayscale vs colored) by itself or as an interaction effect was revealed, contrary to our initial prediction. Therefore, even if the colors were removed, the remaining visual stimuli were sufficient to drive observers to fixate on consistent regions. A similar result, again contrary to our initial prediction, was revealed across spatial audio and mono audio conditions, such that consistency scores across these two conditions remained relatively the same. Again, it can be suggested that even if the spatial information of the audio was removed, the remaining auditory stimuli were sufficient to drive observers to fixate on consistent regions. However, consistency scores for the mute condition were significantly lower than both mono and spatial audio conditions in music instrument and human speech audio categories but not for vehicle sounds. Such a result is on par with the initial prediction as the stimuli modality is reduced when switched to the mute condition. A more comprehensive explanation of these results would be to theorize that stimuli with meaningful content (we initially designated vehicle sounds to be a category that still

contains spatial features but no semantic meaning, similar to a noise) continue driving fixated locations of observers even if its signal quality is reduced (in our cases; presenting only luminance signals with hue removed, and presenting audio only mono with its spatiality removed) until the signal is completely removed (in our case, the mute condition for audio). Alternatively, the effects of color and audio directionality can only be apparent when fixations are analyzed over time, or an analysis over scan paths are done.

3.5.5 Effects of place category

Place category was included to further investigate the effects seen in the preliminary experiment as an observation factor without any specific expectations. A significant ordering of consistency scores across different places categories was revealed as $indoors < outdoors(human\ made) < outdoors(natural)$ in videos with music instrument audio category, but no other effects in other audio categories. To investigate possible causes, the researcher inspected all videos by their event content. In each place category, none of the music instrument videos, three out of nine videos in human speech videos and two-to-three out of nine videos in vehicle sound videos indicated event segmentation and sequences of events. Therefore, videos with multiple event sequences are also distributed across place categories uniformly. Still, the place category by itself might not be directly relevant to consistency scores, but audio and place categories, when combined together, could indicate some resemblance within videos and segregate what can be expected from the content. For example, music instrument sound with outdoors natural place category combinations commonly included (but was not limited to) a single music performer, whereas

outdoors(human-made) included concerts with several music performers, and indoors included orchestra with many music performers simultaneously.

3.5.6 Notable observations from the qualitative data

The qualitative data was collected only for observatory purposes in an unstructured verbal feedback format from all 100 participants. The numbers explained below do not imply opposite results for other participants; they just did not mention the same statement, as the notes below are compiled from open-ended feedback sessions without any specific questionnaire. Although a qualitative analysis was not conducted, certain trends seem to emerge from the data that could be helpful in interpreting results and might be beneficial in future study designs. Nine participants indicated that they usually tried to understand their position (namely, where the camera was placed) within the virtual environment and get used to the new orientation whenever a new video started. The participants seemed to be guided by their individual preferences, habits, and what they considered interesting. Thirty participants explained their preferences, but interestingly, there were mixed preferences for video content, and each video content category was appealing to several participants while several participants loathed it. Twenty-five Participants found videos with sound more engaging and enjoyable than silent ones; some especially noted that it made the experience feel more immersive. Thirteen participants pointed out that the camera position, especially the elevation level, sometimes caused discomfort and made them feel less immersed. Five participants expressed a desire to interact with the objects inside videos rather than passively viewing.

3.5.7 Future Direction

As the dataset was collected with a mindset of allowing future research as well, it includes unused data that can be useful in several applications. First of all, different saliency prediction models can be evaluated in the dataset to determine their strengths and weaknesses over different conditions. That way, how well they behave in different audio modalities, modalities, place types, and audio types can be observed. Moreover, semantic segmentation data can also be leveraged to evaluate these models, indicating how accurately they predict the saliency of different kinds of objects. Furthermore, an audio-visual saliency prediction model with a complex architecture of extracting semantic features defined above and using these to make better saliency predictions can be built and trained in this dataset.

Secondly, the effects of previous VR experience and video game experience on fixation behavior can be analyzed by segregating participants and comparing their fixation locations. Also, a statistical analysis regarding which objects are looked at more in various conditions could also yield interesting insights. Likewise, the data can also be grouped as animate and inanimate objects. Sound generating moving objects can be marked or calculated via audiovisual correspondence. Moreover, effects of those might interact with which kinds of objects are fixated on by different participant groups and possibly vary between video conditions.

Thirdly, the qualitative data from all participants can be analyzed to better understand human behavior in VR and the saliency phenomenon, get new insight, and form new hypotheses. Moreover, the large sample size of observers can enable conducting a qualitative analysis via participants' open-ended explanations.

Furthermore, a future study might be conducted to investigate the effects of camera movement on fixated locations or consistency scores, since the results point

out that the camera movement might be a big factor in fixation locations, and the dataset contains many videos with different types of movement, although not labeled. Likewise, the effect of place type seen in music instrument videos can be further investigated by introducing new factors (such as the number of different objects).

Overall, the dataset contains many data that can be useful in both computer science and psychophysics research areas, and other researchers can freely augment and use the data for their research purposes.

CHAPTER 4

LOW LEVEL PSYCHOPHYSICS EXPERIMENT

Although the effects of several low-level factors on saliency are well-studied in 2D environments, their counterparts in 3D or panoramic environments are mostly missing (Haskins et al., 2020; Judd et al., 2012; Martin et al., 2022; Tsiami et al., 2016). Considering the research in the literature, it has been decided that studying the effects of luminance and color contrast as visual factors and sound amplitude as an auditory factor, along with their spatial directions, would be a good starting point (Herwig & Schneider, 2014; Itti et al., 1998; Itti & Koch, 2001; Meyer CA & Wuerger, 2001; Wada et al., 2003). Moreover, numerous researchers have proposed that the specific task significantly impacts how saliency is distributed (Corbetta & Shulman, 2002; Martin et al., 2022). As a result, it was decided that saliency under different tasks, specifically free view, search, and memory, could be investigated.

Moreover, as the literature for low-level psychophysics experiments in 3D is only recently growing, designing any controlled psychophysics experiment requires adaptation of stimuli and procedures from two dimensions to the new environment (de Gelder, Kätsyri, & de Borst, 2018; Kiepe, Kraus, & Hesselmann, 2023; Martin et al., 2022; Vasser & Aru, 2020). To this end, it was decided to use a uniformly spotted sphere grating as an alternative to a 2D directionless grating stimulus, such as a polar grating. Additionally, instead of forcing the head into a fixed position with a chin rest, which is common practice in 2D research (Lima & Ventura, 2023), it has been decided to keep the head free and present stimuli based on the last tracked position of the head. Finally, pure-tone auditory stimuli are employed similarly to

other research (Fastl & Zwicker, 2007) but are played with spatial direction according to the tracked head position of the listener.

4.1 Research questions

The main investigated question is how several low-level audio-visual factors affect saliency when observers are presented with 3D stimuli. To that goal, the experiment is designed to present stimuli with varying levels of the chosen factors and collect eye-tracking data from the participants. The dependent variables indicating saliency, number of fixations, percentage of fixations, and total fixated duration on each stimulus are calculated. Finally, the relationship between the factors and the dependent variables is analyzed.

The next investigated question is how the effects in the previous question vary for different types of tasks. To answer that question, the experiment was repeated with free viewing, search, and memory recall tasks. The saliency metrics are expected to be more goal-oriented during an active task than free viewing. In addition, for search and memory recall tasks, participants' performance through correct answer rates is also recorded and compared across different factor levels and stimuli positions.

4.2 Experimental design

4.2.1 Stimuli preparation

For calculating color and luminance levels, CIELUV color space was adopted as it offers a perceptually uniform space that is accurate for colored lights. The white point threshold is taken as d65, corresponding to approximately 6500 Kelvins, as it is recommended for colorimetric calculations as the standard illuminant for average

daylight (Fairchild, 2013). When the white point is taken as so, the luminance scale ranges from 0 to 100, and chromaticity scales, u^* and v^*, range from -200 to 200.

For presenting all visual stimuli, OpenGL was adopted due to its high performance and accuracy in generating per-eye camera projections of any 3D visual stimuli (Khronos Group, 2017). A combination of spherical harmonics was used to generate a spherical texture with uniform spotted gratings. More specifically, to create 12 uniformly distributed spots corresponding to a dodecahedron's face-centers or icosahedron's vertices, the formula $f(\theta, \phi) = Y_6^0(\theta, \phi) + \sqrt{\frac{28}{11}} Y_6^5(\theta, \phi)$ was used, followed by a min-max normalization to generate the final textures (Blais, 2011; Bertet, Daniel, & Moreau, 2006; GoLuckyRyan, 2017). Figure 20 shows a 2D equirectangular projection of the generated texture.

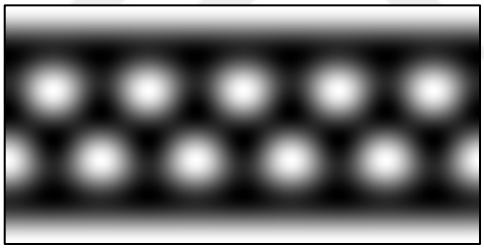


Figure 20. An equirectangular projection of the spotted sphere grating texture

This texture is multiplied by chromaticity contrast levels to prepare the stimuli as isoluminant color contrasted gratings and presented with a fixed luminance level. Finally, the texture defined in CIELUV color space is converted to RGB colors via transformations and gamma-correction at $\gamma = 2.4$ using Psychtoolbox (Kleiner, Brainard, & Pelli, 2007) functions and supplied to OpenGL with 16-bit depth for rendering.

Equal loudness levels across different frequencies of pure tones specified by ISO226 specification's equal loudness curves are used to calculate sound loudnesses. In that specification each phon level corresponds to the same dB SPL for 1-kHz pure tones while other frequencies are adjusted to match the same level of perceived loudness. On top of that, instead of scales defined on phons, a perceptually uniform space convertible to phon scale, sone scale is used. A sound loudness of 1 sone corresponds to 40 phons, and doubling the sone level corresponds to a perceptually twice as loud sound, corresponding to an increase of 10 phons approximately (Fastl & Zwicker, 2007). For calculating sound pitch, the mel scale corresponding to a perceptually uniform sound frequency is preferred. A sound frequency of 1000 Mel corresponds to a 1-kHz signal, and equal mel frequency spacings represent equal perceived sound pitch contrasts (Moore, 2012).

For presenting all auditory stimuli, OpenAL was adopted due to its high performance and accuracy in calculating binaural audio estimations of any 3D auditory stimuli (Hiebert, Gordon, Peacock, & Vogelsang, 2005). To design a controlled empty environment, no additional room reflection or reverb effects were activated when using OpenAL. For the task at hand, varying loudness levels of directed pure-tone sounds of 1000 Mel were generated with a sampling rate of 44100-Hz and 16-bit depth with the sine function in Matlab. Additionally, 0.05-second cosine-squared rise and decay times are implemented to prevent any rustle sound side-effect (Lentz & Townsend, 2022; Ryan-Warden, Ng, & Keating, 2019).

All trials take place on an empty gray screen with 50 luminance, zero color chromaticities, and no background sound. Prior to each experiment block, the headset's native calibration routine is run to both make sure the headset is worn correctly and to collect eye-tracking data correctly.

4.2.2 Just noticeable difference calculation

To calculate just noticeable difference (JND) levels for luminance, color contrast for each chromaticity, sound loudness, and sound pitch contrast per participant, 2 Answer Forced Choice tasks for each factor were designed. JND levels are calculated by fitting a Weibull psychometric function with a 50% guess rate and a 1% lapse rate to determine the level of 75% success rate, following the standard practice for two-alternative forced choice (2AFC) tasks (Nachmias, 1981; Pelli & Farell, 1995; Treutwein, 1995). In order to present trials more economically and calculate JND levels effectively, QUEST ZEST was used for psychometric function fitting (Farell & Pelli, 1998; King-Smith, Grigsby, Vingrys, Benes, & Supowit, 1994; Pelli & Farell, 1995; Watson & Pelli, 1983). Since it has been suggested that at least 40 trials per variable are enough for such a calculation, it has been decided to do 50 trials per variable for an accurate calculation. Making sure that the spherical stimuli presented both in the left and right reside within the field of view of both eyes while maximally separated in the currently used hardware, empirical testing of the shared visible field of view was done by the researcher, and the reported per-eye field of view angles by the hardware was used to determine the positions to place left and right visual stimuli. Specifically, it has been decided to place them such that each sphere center is 26cm away horizontally from the middle point when the viewer's head position is taken as the origin, and the spheres are placed 50cm in front of the viewer. After the block is complete, if the participant has misunderstood the task or felt that they could do better, the block is repeated once more. Figure 21 displays the trial structure and the design of the JND calculation block.

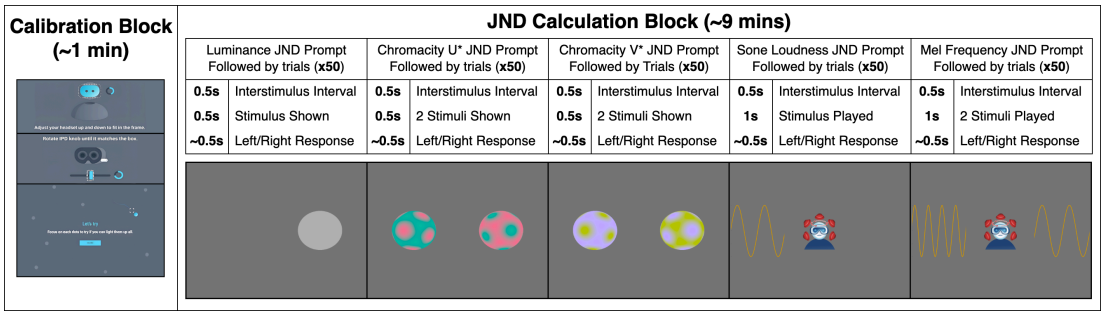


Figure 21. Trial structure of the JND Calculation Block

4.2.2.1 Calculation of luminance JND level

In each trial, after a 0.5-second interstimulus interval of an empty screen, a solid visual sphere stimulus of a different luminance level is presented in either the left or the right position in front of the participant for 0.5 seconds. Different luminance levels are decided by QUEST for optimal calculation. After the stimulus disappears, the participant presses either the left or the right controller button to denote the position of the stimuli. As the task is 2AFC, if the participant is unsure or cannot differentiate the stimuli from the background, they still need to press either left or right to continue. After all trials, QUEST computes the mean luminance JND level corresponding to the luminance level threshold needed for the participant to answer the task better than chance. The calculated value is used as the luminance JND level in the rest of the experiment.

4.2.2.2 Calculation of color chromaticity u* and v* contrast JND levels.

The second and third parts calculated u^* and v^* chromaticity contrast JND levels, respectively. In each trial, after a 0.5-second interstimulus interval of an empty screen, two uniformly but oppositely spotted visual sphere grating stimuli of a different color chromaticity contrast level u^* with zero v^* chromaticity contrast and luminance fixed to the previously calculated JND level are presented in left and right

positions in front of the participant for 0.5 seconds. Different color chromaticity contrast levels u^* are decided by QUEST for optimal calculation. After the stimulus disappears, the participant presses either the left or the right controller button to denote the position of the red-spotted cyan stimulus. As the task is 2AFC, if the participant is unsure or cannot differentiate the stimuli from each other, they still need to press either left or right to continue. Then, the same structured trials are applied to calculate chromaticity contrast level v^* by using different levels of v^* contrast and zero u^* contrast. This time, however, the participant tries to differentiate the yellow-spotted blue stimulus from the blue-spotted yellow stimulus. After all trials, QUEST computes the mean color chromaticity u^* and v^* JND levels corresponding to the color chromaticity contrast level thresholds needed for the participant to answer the task better than chance. The calculated value is used as the color chromaticity contrast u^* and v^* JND levels in the rest of the experiment.

4.2.2.3 Calculation of loudness JND level.

In each trial, after a 0.5-second interstimulus interval of a mute screen, a pure-tone sound of 1000 Mel frequent with a different level of sone loudness is played for 1.0 second from one of eight directions corresponding to corners of an invisible cube of one-meter side length in which the listener is placed at its center. More specifically, the sound stimulus comes from a combination of either front or back, top or bottom, left or right. Different levels of sone loudness are decided by QUEST for optimal calculation. After the stimulus stops, the participant presses either the left or the right controller button to denote whether the stimulus played from one of the four corners on the left or one of the four corners on the right. As the task is 2AFC, if the participant is unsure or cannot differentiate the stimuli from each other, they still

need to press either left or right to continue. After all trials, QUEST computes the mean sone loudness JND level corresponding to the loudness level threshold needed for the participant to answer the task better than chance. The calculated value is used to calculate the loudness of the JND level in the rest of the experiment.

4.2.2.4 Calculation of pitch contrast JND level.

In the final part, the pitch contrast JND level is calculated. In each trial, after a 0.5-second interstimulus interval of a mute screen. Two pure tone sounds with equal amounts of Mel frequency away from 1000 Mel frequencies are played at the same time at the previously calculated loudness JND threshold for 1.0 seconds such that one of them is played from one of the four corners on the left, the other one is played from one of the four corners on the right in the same setting as the previous task. Different levels of Mel frequency contrast are decided by QUEST for optimal calculation. After the stimulus stops, the participant presses either the left or the right controller button to denote whether the higher-pitched stimulus played from one of the four corners on the left or one of the four corners on the right. As the task is 2AFC, if the participant is unsure or cannot differentiate the stimuli from each other, they still need to press either left or right to continue. After all trials, QUEST computes the mean Mel frequency JND level corresponding to the pitch contrast level threshold needed for the participant to answer the task better than chance. The calculated value is used as the Mel frequency JND level in the rest of the experiment.

4.2.3 Experiment trials

The calculated values of JND per participant are used to prepare stimuli in the following blocks. They all follow the following rules:

- The head of the participant resides in the center of an invisible cube of a one-meter side length. Stimuli are presented from some of the corners of that cube; hence, there are eight different stimulus positions.
- There are four types of visual stimuli, namely: (i) pink-spotted green sphere, (ii) green-spotted pink sphere, (iii) orange-spotted blue sphere, (iv) blue-spotted orange sphere. These colors are generated through modulating chromaticities u and v. More specifically, +u,-v corresponds to pink, -u,+v corresponds to green, +u,+v corresponds to orange, -u,-v corresponds to blue.
- In each trial, all four types of visual stimuli are shown in four of eight different stimulus positions. The other four stimuli positions remain empty. Except for the free-view task, one of these four visual stimuli is the target stimulus, and the rest are distractors. Each trial can include a pure-tone auditory stimulus directed from either the target, a distractor, an empty position, or all positions simultaneously.
- There are two luminance level conditions: at the JND level (*barely visible*) and twice the JND level (*easily visible*). There are two color contrast level conditions: at the JND level (*barely visible*) and twice the JND level (*easily visible*). Positions of visual stimuli are randomized across trials, making sure that target stimuli occur in each of the eight positions an equal number of times across each condition. There are five types of audio conditions: mute, mono (audio coming from all eight positions), directed from the target stimuli, directed from a distractor stimulus, and directed from an empty position. There are two loudness conditions for those with audio: at the JND level (*barely audible*) and twice the JND level (*easily audible*).
- Each participant enters a total of 288 trials from each task type, divided into two blocks. Each participant enters trials from all types of conditions. Trial conditions

are randomized via balanced Latin squares to make sure all combinations of conditions are presented an equal number of times to each participant. Each participant enters each task block once for two days. The order in which each participant enters different task blocks is also randomized via Latin squares. So, the total number of trials is $luminance * color * audio * targetposition$, which corresponds to $2 * 2 * (5 * 2 - 1) * 8 = 288$ trials.

- Each block starts with a practice round of 12 trials. The participant is initially shown a prompt explaining the procedure. The participant starts the practice round by pressing the trigger button. Trials in the practice round contain random conditions for the task at hand. After completing the practice round, the participant is shown their task performance. If the participant misunderstood the task or it still felt unfamiliar to them, they can restart the practice round by double-pressing the trigger button. Otherwise, they press the trigger button once to continue. They are shown a prompt for the main round, again explaining the same procedure. They press the trigger button once to start. The main round consists of 144 trials, and upon completion, the participant is shown a final screen indicating the block is complete but is not notified about their task performance.
- To succeed, participants need to perceive the colors and locations of the stimuli and complete task-specific goals by making decisions about the correct answers. Tasks differ by cognitive functions such that the memory recall task relies on memorization and recalling locations of all stimuli via working memory while the search task relies on learning the color of the target stimulus and utilizing it for selective attention in visual search. Free view task is introduced as a control group in which the exploration behavior is not task-driven.

4.2.3.1 Free view task block

The free view task block starts with a prompt explaining the procedure. Each trial starts with an empty gray screen. The participant presses the trigger button to continue. All four visual stimuli are shown in four of eight positions for 6.0 seconds, along with an accompanying sound (except for the mute conditions). After these 6.0 seconds are complete, the stimuli disappear, and the trial is finished. Meanwhile, the participant's eye-tracking data is recorded. Figure 22 shows the trial design for the free view task block.

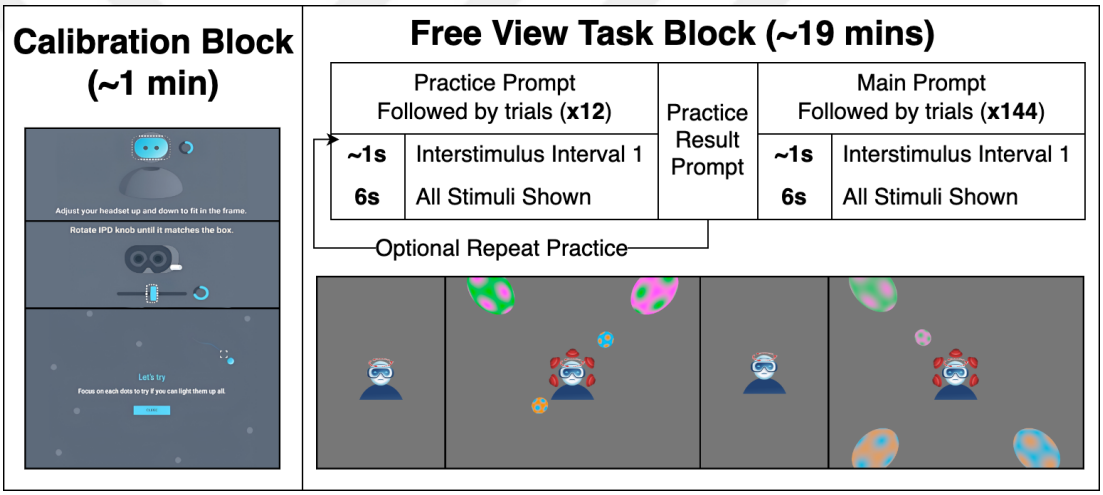


Figure 22. Trial structure of the free view task block

4.2.3.2 Search task block

The search task block starts with a prompt explaining the procedure. Each trial starts with an empty gray screen. The participant presses the trigger button to continue. The target visual stimulus for the current trial is shown in front of the participant for 1.0 seconds, followed by an empty gray screen for 1.0 seconds. Then, all four visual stimuli are shown in four of eight positions for 6.0 seconds, along with an accompanying sound (except for the mute conditions). The participant searches for

the target stimulus and, upon finding it, points their controller onto the target sphere like a laser pointer and presses the trigger button to select. Consequently, the stimuli disappear, and the trial is finished. Meanwhile, the eye-tracking data of the participant, response duration, and whether the response was correct are recorded. Figure 23 shows the trial design for the search task block.

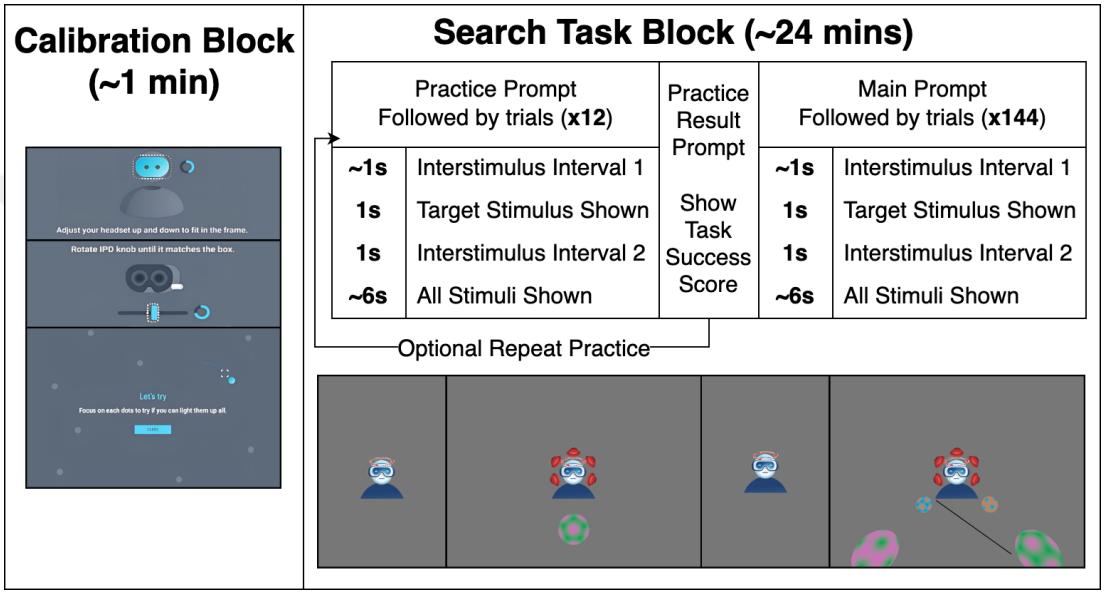


Figure 23. Trial structure of the search task block

4.2.3.3 Memory recall task block

The memory recall task block starts with a prompt explaining the procedure. Each trial starts with an empty gray screen. The participant presses the trigger button to continue. All four visual stimuli are shown in four of eight positions for 6.0 seconds, along with an accompanying sound (except for the mute conditions). The participant tries to memorize the stimuli. Then, an empty gray screen is shown for 1.0 seconds. Finally, the target visual stimulus is shown in front of the participant along with easily visible (twice the JND luminance level) gray spheres in all eight positions. The participant tries to remember in which position they have seen the target stimulus,

points their controller onto the gray sphere in that position like a laser pointer, and presses the trigger button to select. Consequently, all stimuli disappear, and the trial is finished. Meanwhile, the eye-tracking data of the participant, response duration, and whether the response was correct are recorded. Figure 24 shows the trial design for the memory recall task block.

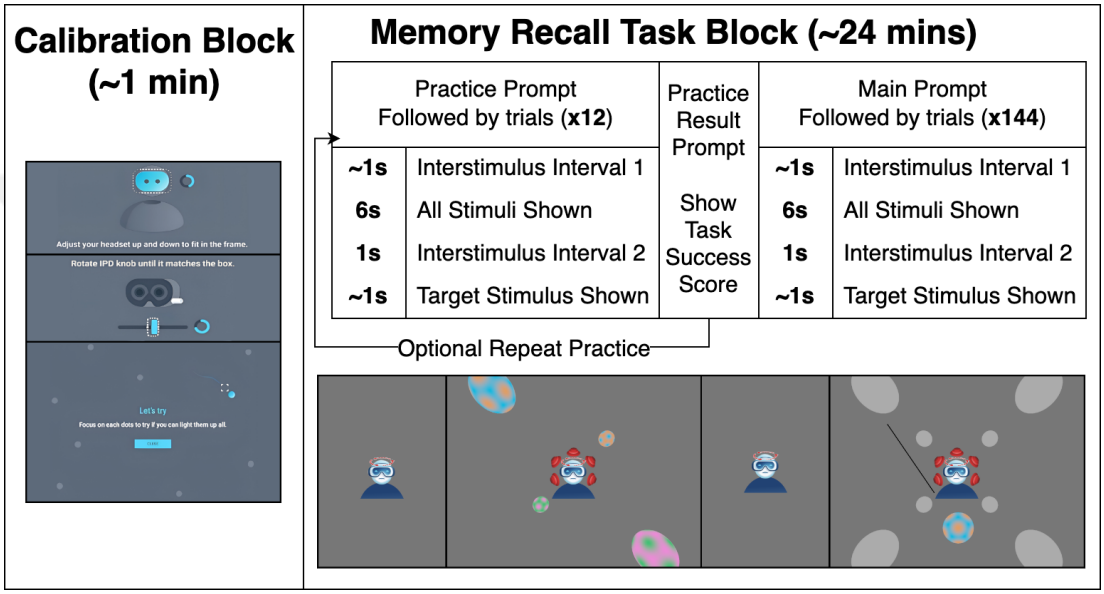


Figure 24. Trial structure of the memory recall task block

4.3 Data collection

4.3.1 Participants

A total of six voluntary researchers from Boğaziçi University (students and alumni) participated in the experiment. All participants had normal or corrected-to-normal vision either by wearing contact lenses or glasses. Neither of them had any auditory problems or color blindness. The study complied with the Boğaziçi University Ethics Coordinating Committee's requirements of research ethics and was approved by them, as shown in Appendix N. All data were collected via randomly assigned IDs to the participants to assure data anonymity and confidentiality. All participants were

either native or proficient speakers of both Turkish and English. The consent form, the questionnaire form, and the experimental instructions were given in either Turkish or English to the participants' preference.

4.3.2 Stimuli and apparatus

In the experiment, participants wore an HTC Vive Pro Eye VR headset to view and hear the stimuli while their eyes were tracked. The VR Headset was connected to a Dell Precision 5820 Tower Workstation Computer (Intel Xeon(R) W-2235 CPU, 16GB DDR4 RAM, NVIDIA Quadro RTX 5000 16GB graphics card) via a DisplayPort and a USB 3.0 cable. During the experiment, participants used a single controller with their dominant hand. The cable connectors of the headset were attached to an extensible and adjustable mechanism fixed to the ceiling with a hook, allowing the wearers to walk and rotate freely without the entanglement or restraint of the cable.

The experiment environment was programmed in MATLAB (The MathWorks Inc., 2023) using Psychtoolbox-3 (Brainard & Vision, 1997; Kleiner et al., 2007; Pelli, 1997), OpenAL (Hiebert et al., 2005), OpenGL (Khronos Group, 2017), IoSR Matlab Toolbox (IoSR-Surrey, 2023) to generate and display the stimulus, and to collect the data. A simple sound calibration routine was run to approximately calculate the maximum sound pressure level of headphone outputs via playing a pure-tone sound at different levels, recording the sound, and measuring the output level via a mobile app, Decibel X (Dinh, 2023), which resulted in 94.1 dB SPL for maximum loudness on 1-kHz pure-tones. This value is only used to find a baseline of initial sound output as the exact JND loudness values are calculated per participant via the JND session.

4.4 Results

4.4.1 Definitions

The target fixation rate in each trial is defined as the sum of fixation durations on the target position while all four visual stimuli are shown, divided by the duration these four visual stimuli were shown. Although the participants did not see the target stimuli any different than others in the free view task, for the sake of analyses, one of the four visual stimuli was chosen as the target stimulus for each trial with the same randomization routine for search and memory recall tasks. Likewise, the average fixation duration in each trial is defined as the average of all fixation durations while all four visual stimuli are shown. Finally, the average fixation duration is defined as the mean fixation duration of all fixations during the time four visual stimuli were shown per trial.

4.4.2 Data processing

First, projections of the raw eye-tracking data on an imaginary sphere with a radius equal to the distance of presented stimuli to the origin were calculated, centering the ray centers on the origin similar to the approach Llanes-Jurado, Marín-Morales, Guixeres, & Alcañiz (2020). Next up, the previously implemented I-DT algorithm to extract fixations from 360° eye-tracking data explained by Agtzidis & Dorr (2019) was used. Using that algorithm, consecutive sequences of raw gaze data with a maximum dispersion of 1.5 visual angles and with a minimum duration of 0.1 seconds were marked as fixations, saving the start, end, duration, and mean direction of each fixation. Using the output, fixations within the timeframe of the four visual stimuli shown in each trial were extracted. By averaging out the durations of each fixation per trial, average fixation durations were obtained. Then, the distance of

each fixation to the closest stimulus was calculated, and the fixations with centers closer than the radius of foveal vision of 2.5° to a spherical stimulus were marked as onto that stimulus. Finally, the fixation rates on the target stimulus position and the audio stimulus position were calculated by taking a sum of fixation durations and dividing by the time the stimuli were shown per trial. All these results were saved in a comma-separated text file for further analysis. All the further analyses are done via R version 4.3.1 (R Core Team, 2023), using the rstatix package (Kassambara, 2023). All plots are created via the ggplot2 package (Wickham, 2009).

4.4.3 Effects of task

A repeated measures ANOVA was conducted to examine the effect of task type on average fixation duration. Figure 25 shows the corresponding boxplot. The results showed that there was no significant main effect of task type on average fixation duration, $F(2, 10) = 1.546$, $p = 0.26$, partial $\eta^2 = 0.236$. This suggests that the average fixation duration did not significantly differ across the three task types.

A repeated measures ANOVA was conducted to investigate the effect of task type on target fixation rate. The results revealed a significant main effect of task type on target fixation rate, $F(2, 10) = 34.702$, $p < 0.001$, partial $\eta^2 = 0.874$. This indicates that the target fixation rate significantly differed across the three task types, with task type explaining 87.4% of the variance in target fixation rate. Follow-up post-hoc paired two-tail t-test comparisons with Bonferroni adjustment revealed that the target fixation rate in the search task was significantly higher than that of the free view task ($p = 0.011$) and the memory recall task ($p = 0.001$). However, there was no significant difference in target fixation rate between free view and memory recall tasks ($p > 0.05$). Figure 26 shows the corresponding boxplot.

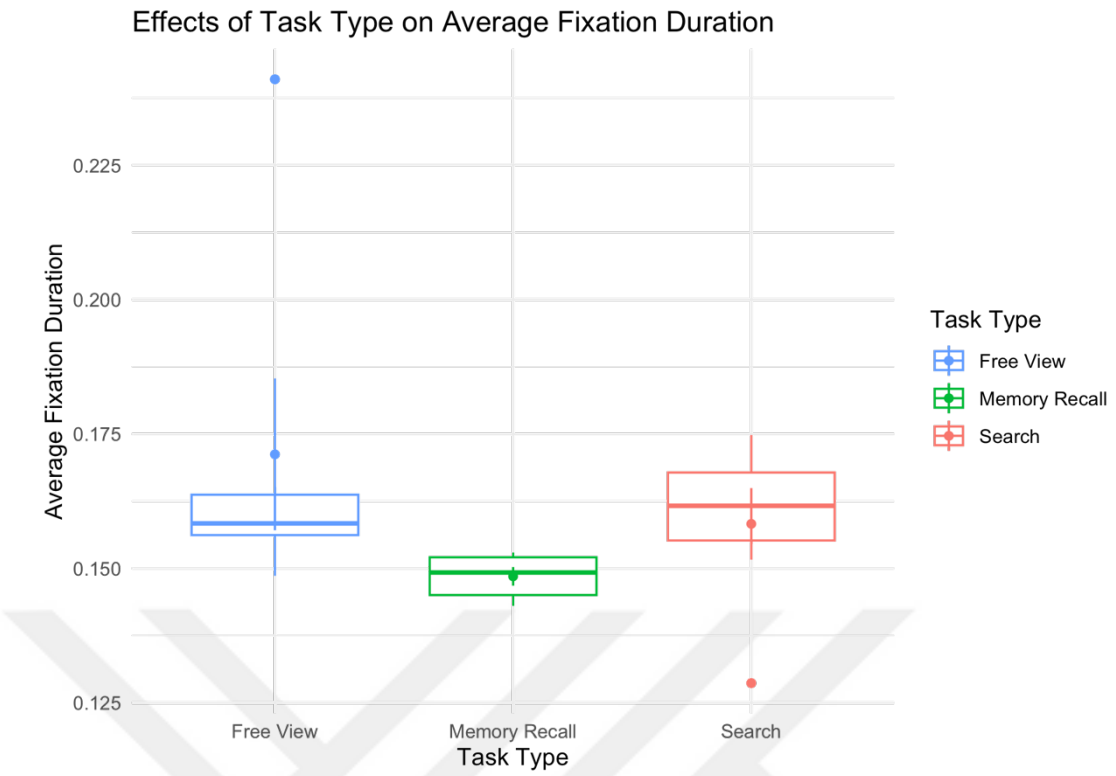


Figure 25. A boxplot showing average fixation durations across tasks

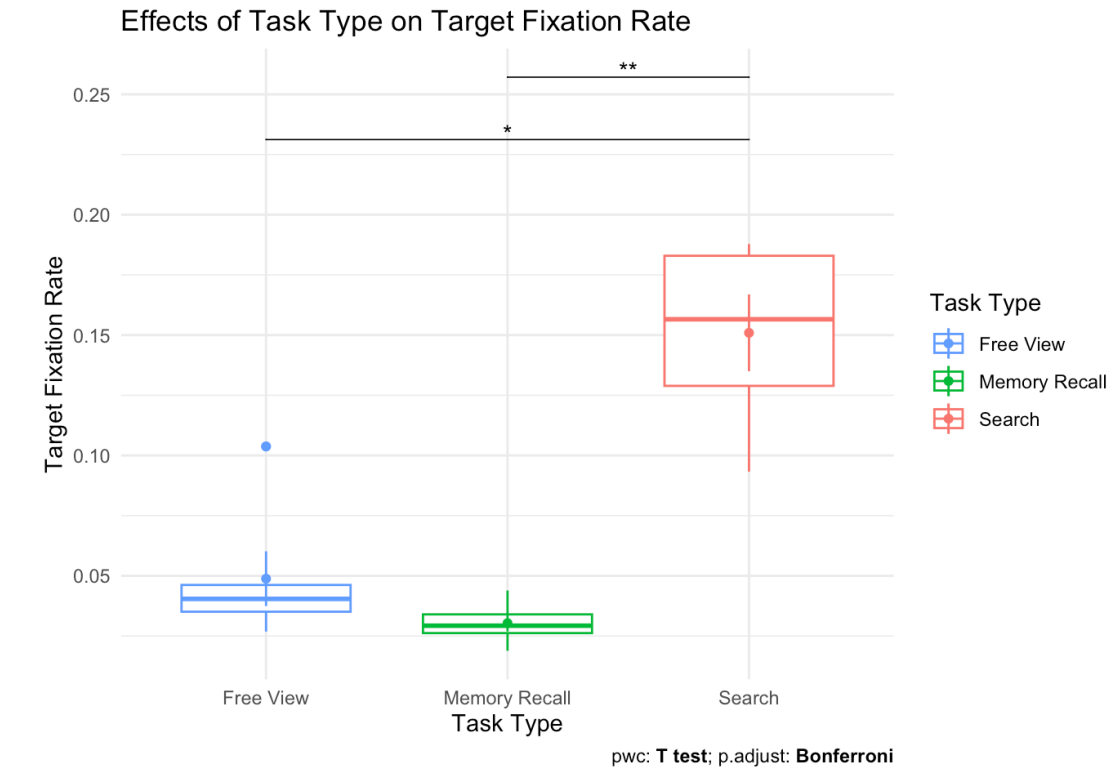


Figure 26. A boxplot showing target fixation rates across tasks

A repeated measures ANOVA was conducted to assess the effect of task type on response duration. The results showed a significant main effect of task type on response duration, $F(1, 5) = 10.214$, $p = 0.024$, partial $\eta^2 = 0.671$. Since there were only two types of tasks, no follow-up test was needed. This suggests that the response duration was significantly higher in search task than memory recall tasks, and is expected by the nature of the tasks. Figure 27 shows the corresponding boxplot.

A repeated measures ANOVA was conducted to examine the effect of task type on response correct rate. The results revealed a significant main effect of task type on response correct rate, $F(1, 5) = 21.523$, $p = 0.006$, partial $\eta^2 = 0.811$. Since there were only two types of tasks, no follow-up test was needed. The result indicates that the response correct rate was significantly higher in search tasks than in memory recall tasks. This result confirms the experiment design, such that the memory recall task is significantly harder than the search task as per design. Figure 28 shows the corresponding boxplot.

Any main effect of task type is skipped in the rest of the ANOVA results as they are always equivalent to section 4.4.3 results.

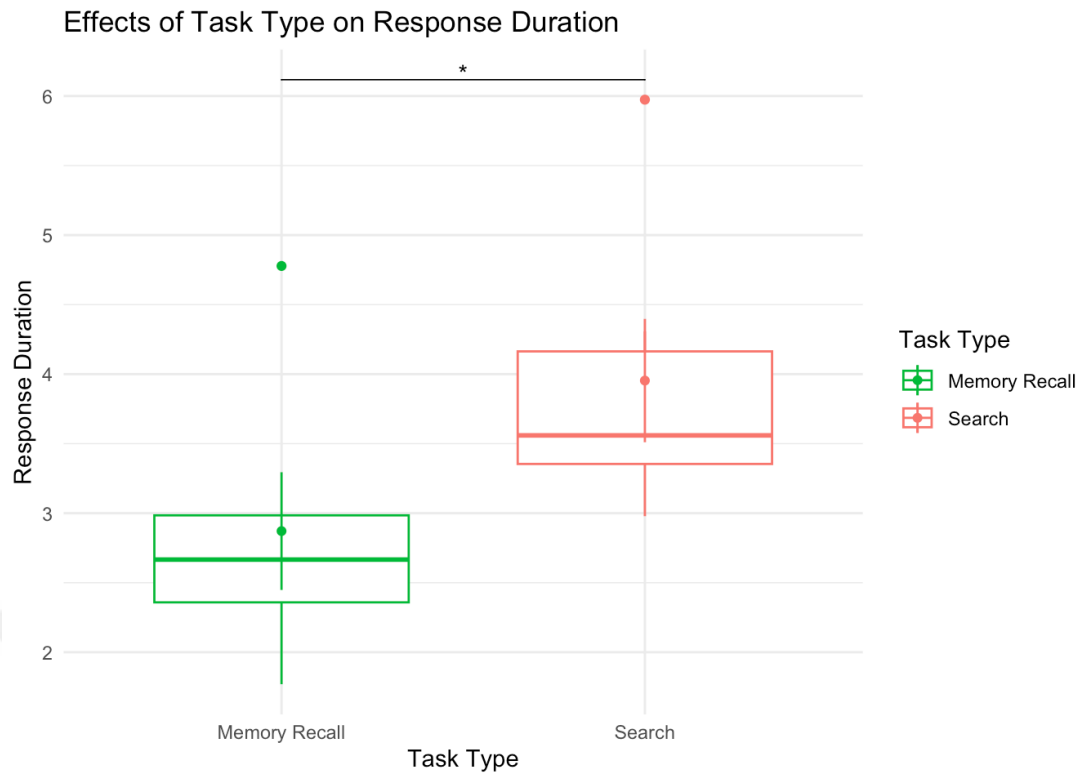


Figure 27. Boxplot showing response durations across tasks

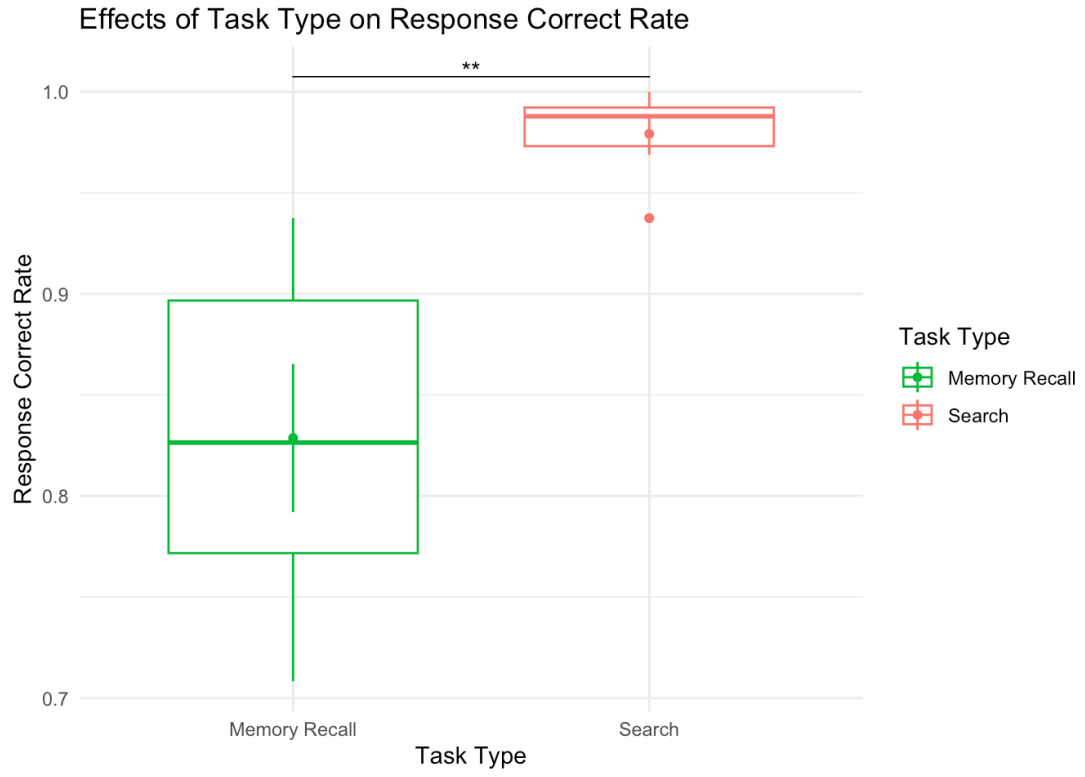


Figure 28. A boxplot showing correct response rate across tasks

4.4.4 Effects of position

A 2 x 2 x 2 x 3 repeated measures ANOVA was conducted to investigate the effects of target position (X, Y, and Z) and task type on target fixation rate. Figure 29 shows the corresponding boxplot. The assumption of sphericity was met for all factors and interactions, except for the interaction between Y, X, and task type, as assessed by Mauchly's test ($W = 0.064$, $p = 0.004$). For this interaction, a Greenhouse-Geisser correction was applied. The results showed significant main effects of Z position, $F(1, 5) = 10.891$, $p = 0.021$, partial $\eta^2 = 0.685$. This indicates that the target fixation rate significantly differed based on the target's Z position. There was also a significant interaction effect between Z position and task type, $F(2, 10) = 7.415$, $p = 0.011$, partial $\eta^2 = 0.597$. This suggests that the effect of the Z position on the target fixation rate varied depending on the task type. No other main effects of position or interactions were found to be significant (all $p > 0.05$). As there were two levels of Z position, the main effect corresponded to significantly higher target fixation rates if the target stimuli appeared in front rather than behind.

For the interaction effect, follow-up one-way repeated measures ANOVA analyses over task type with Bonferroni correction were conducted per Z position. Both analyses revealed a significant main effect of task type. ($F(2,10) = 43.5$, $p < 0.0001$, $\eta^2 = .897$ for the position "behind", and $F(2,10) = 21.9$, $p = 0.0004$, $\eta^2 = .814$ for the position "front") Therefore, follow-up post-hoc paired two-tailed t-tests were conducted for both positions with Bonferroni correction. The results revealed that for both behind and front target positions, the target fixation rate was significantly higher in search tasks than in free view tasks ($p = 0.003$ for behind, $p = 0.026$ for front) and in memory recall tasks ($p = 0.003$ for behind, $p = 0.002$ for front). Comparisons between memory recall and free view tasks did not reveal

any significant effect on target fixation rate in either front or behind positions (both $p > 0.05$). Figure 30 shows the corresponding boxplot.

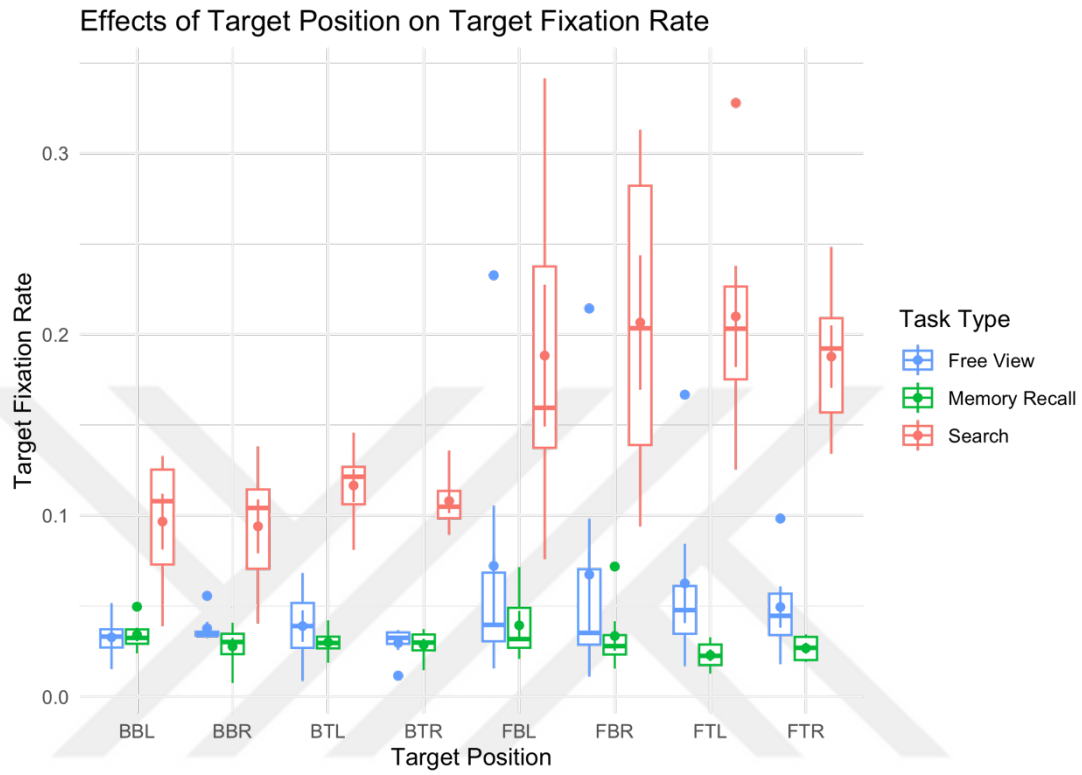


Figure 29. A boxplot showing target fixation rate across target positions per task

BBL corresponds to Behind Bottom Left, and FTR corresponds to Front Top Right. Other labels are similarly named. The positions were relative to the participants' positions just before the visual stimuli appeared.

A 2 x 2 x 2 x 2 repeated measures ANOVA was conducted to examine the effects of target position (X, Y, and Z) and task type on response duration. Results revealed significant main effects of the Z position, $F(1, 5) = 20.942$, $p = 0.006$, partial $\eta^2 = 0.807$. This indicates that the response duration significantly differed based on the target's Z position. Additionally, there was a significant interaction effect between Z position and task type, $F(1, 5) = 10.450$, $p = 0.023$, partial $\eta^2 = 0.676$. This suggests that the effect of Z position on response duration varied

depending on the task type. No other main effects of position or interactions were found to be significant (all $p > 0.05$). Figure 31 shows the corresponding boxplot.

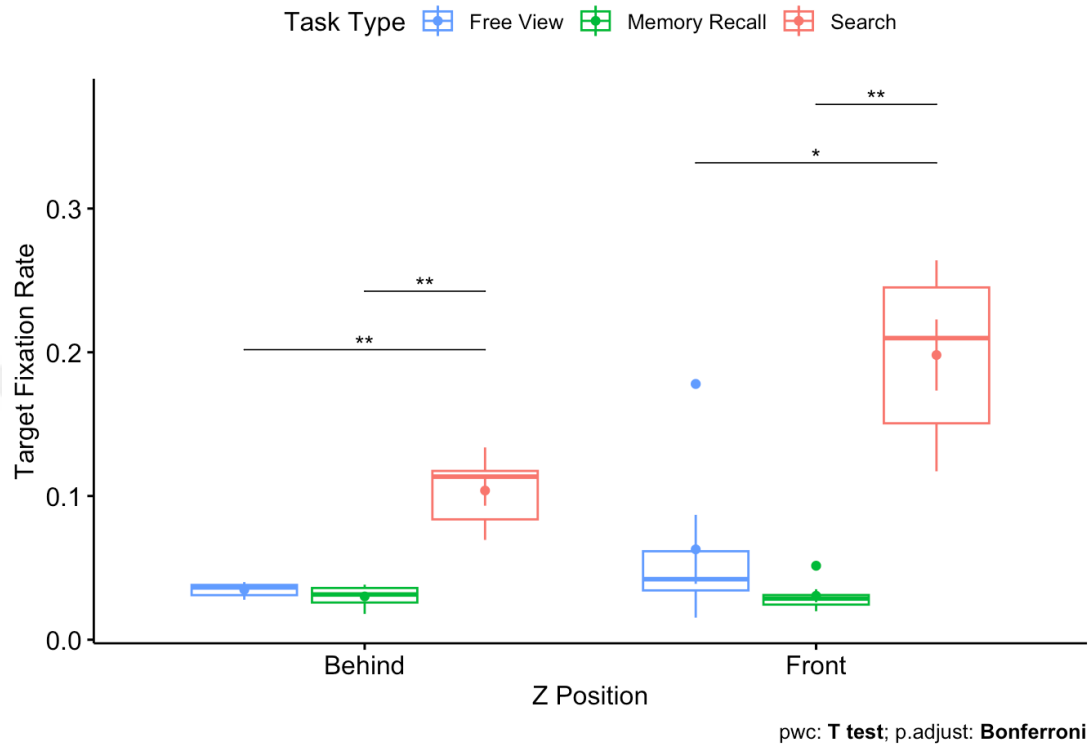


Figure 30. A boxplot showing target fixation rate task types per Z position

For the interaction effect, follow-up post-hoc paired two-tailed t-tests were conducted for both positions with Bonferroni correction. It revealed that participants responded significantly slower in search tasks than in memory recall tasks when the target element was behind ($p = 0.01$), but this effect was diminished when the target element was in front ($p > 0.05$). Figure 32 shows the corresponding boxplot.

A 2 x 2 x 2 x 2 repeated measures ANOVA was conducted to investigate the effects of target position (X, Y, and Z) and task type on response correct rate. Figure 33 shows the corresponding boxplot. The results showed no significant main effects of position or interactions (all $p > 0.05$).

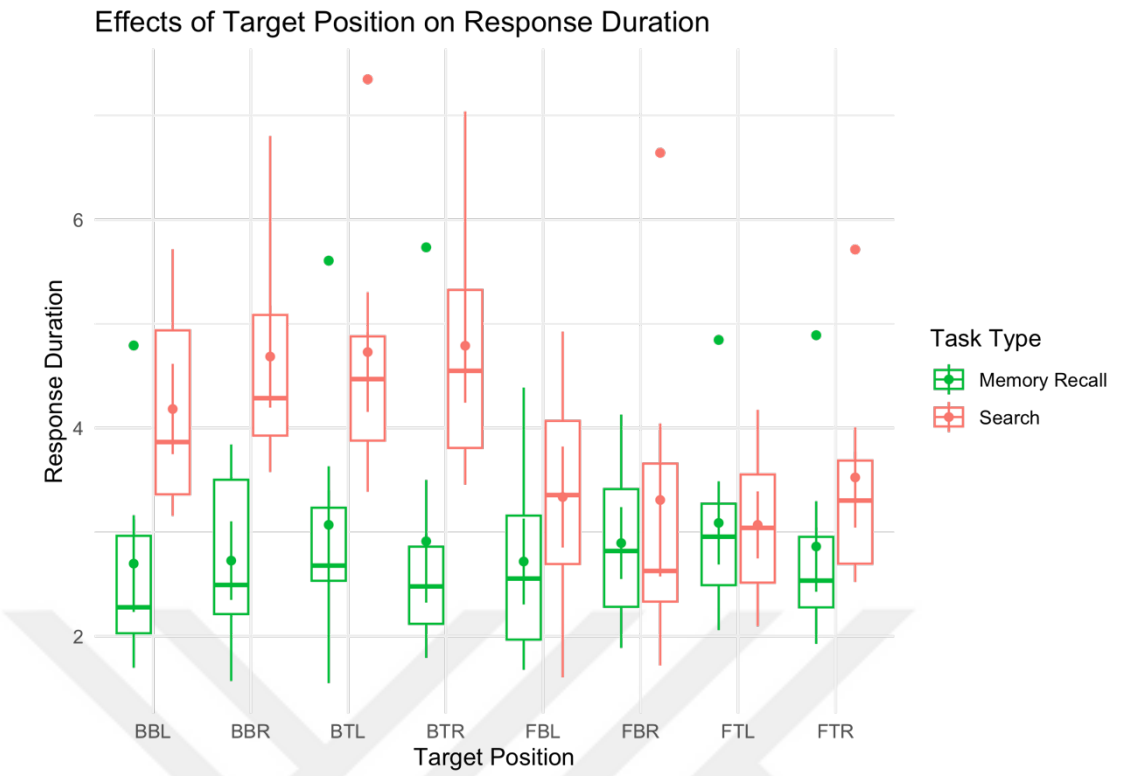


Figure 31. A boxplot showing response durations across target positions per task

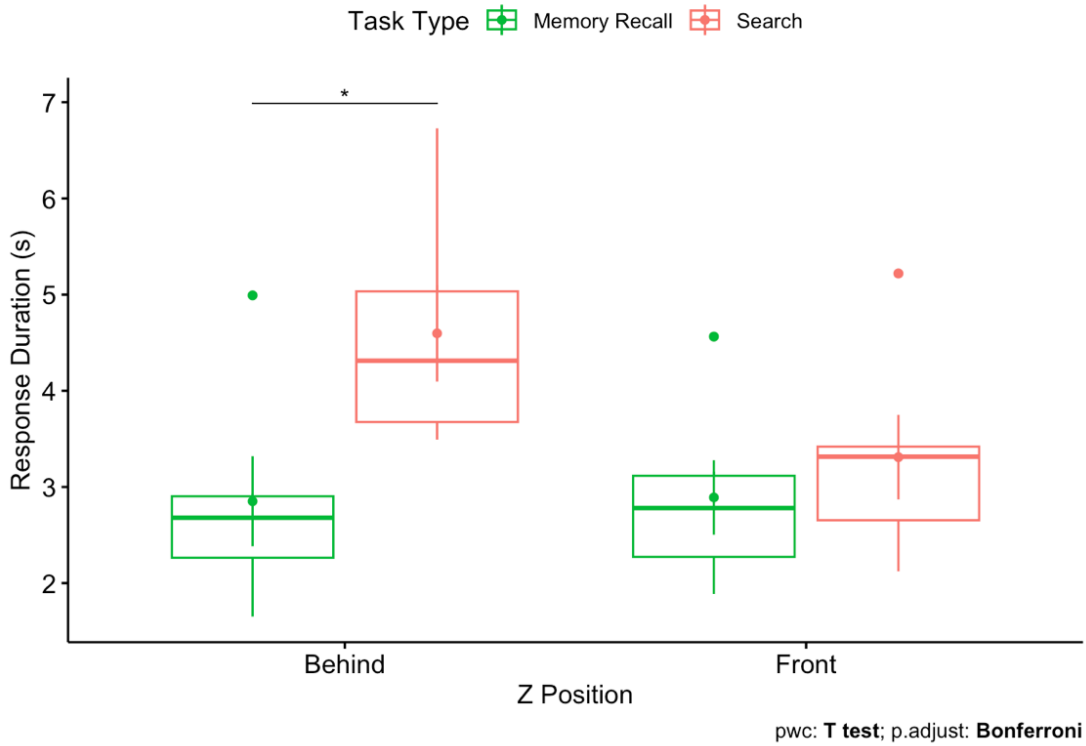


Figure 32. A boxplot showing durations across task types per Z position

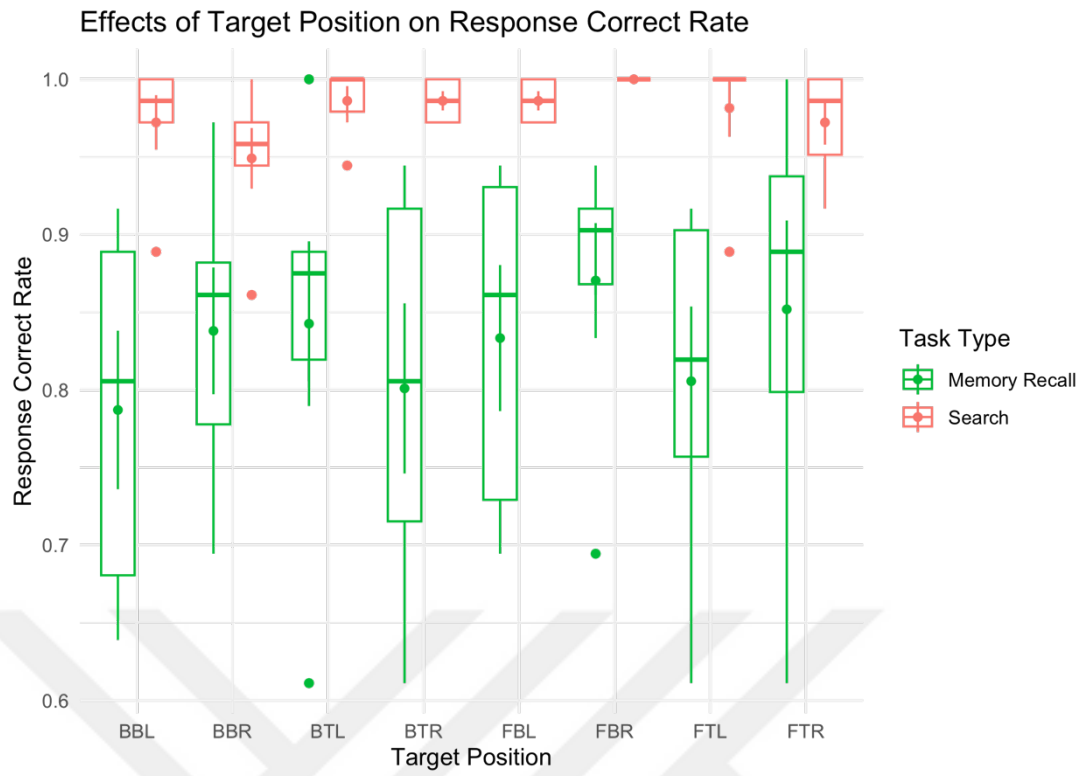


Figure 33. A boxplot showing correct response rate across target positions per task

4.4.5 Effects of color and luminance

A 2 x 2 x 3 repeated measures ANOVA was conducted to examine the effects of visual luminance, color contrast, and task type on average fixation duration. Figure 34 shows the corresponding boxplot. The assumption of sphericity was met for all factors and interactions, except for the interaction between color contrast and task type, as assessed by Mauchly's test ($W = 0.210$, $p = 0.044$). For this interaction, a Greenhouse-Geisser correction was applied. The results showed no significant main effects of visual luminance, $F(1, 5) = 1.235$, $p = 0.317$, partial $\eta^2 = 0.198$, or color contrast, $F(1, 5) = 1.260$, $p = 0.313$, partial $\eta^2 = 0.201$. This indicates that the average fixation duration did not significantly differ based on visual luminance or color contrast. Additionally, no interactions were found to be significant (all $p > 0.05$).

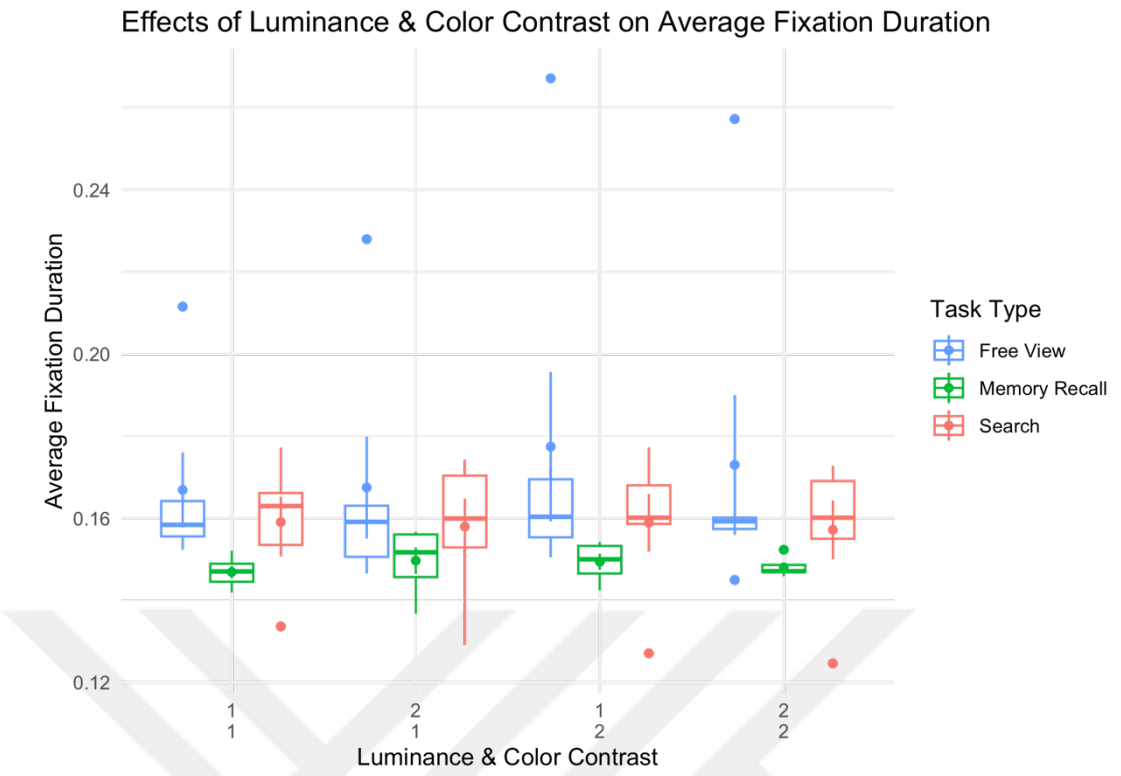


Figure 34. Average fixation durations across luminance and color contrast per task

On the horizontal axis, the top number corresponds to the luminance level; the bottom number corresponds to the color contrast level, such that "1" depicts the JND level (barely visible), and "2" depicts twice the JND level (easily visible).

A 2x2x3 repeated measures ANOVA was conducted to investigate the effects of luminance, color contrast, and task type on target fixation rate. It revealed a significant main effect of color contrast, $F(1, 5) = 7.328$, $p = 0.042$, partial $\eta^2 = 0.594$. Since there are only two levels of color contrast, no follow-up was needed. The result indicates that the target fixation rate is significantly higher in the easily visible color contrast level than in the barely visible contrast level. However, there was no significant main effect of visual luminance, $F(1, 5) = 0.005$, $p = 0.946$, partial $\eta^2 = 0.000997$, suggesting that luminance did not significantly influence the target fixation rate. Additionally, no interactions were found to be significant (all $p > 0.05$). Figure 35 and Figure 36 show the corresponding boxplots.

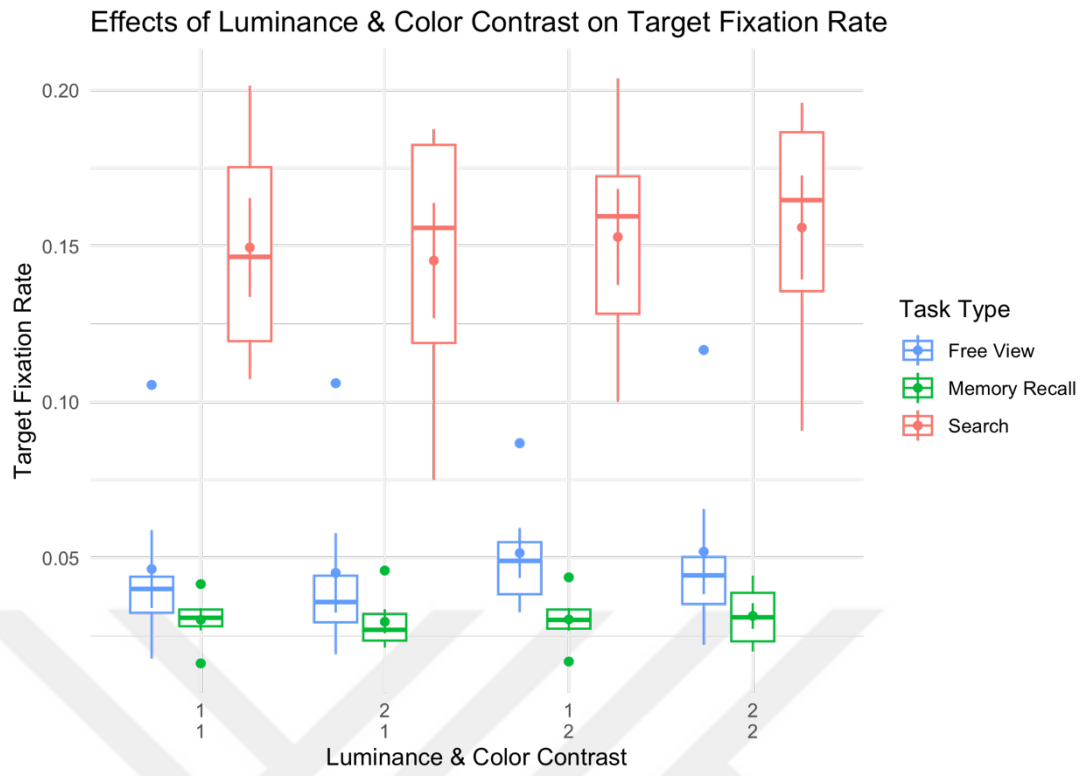


Figure 35. Target fixation rate across luminance and color contrast per task

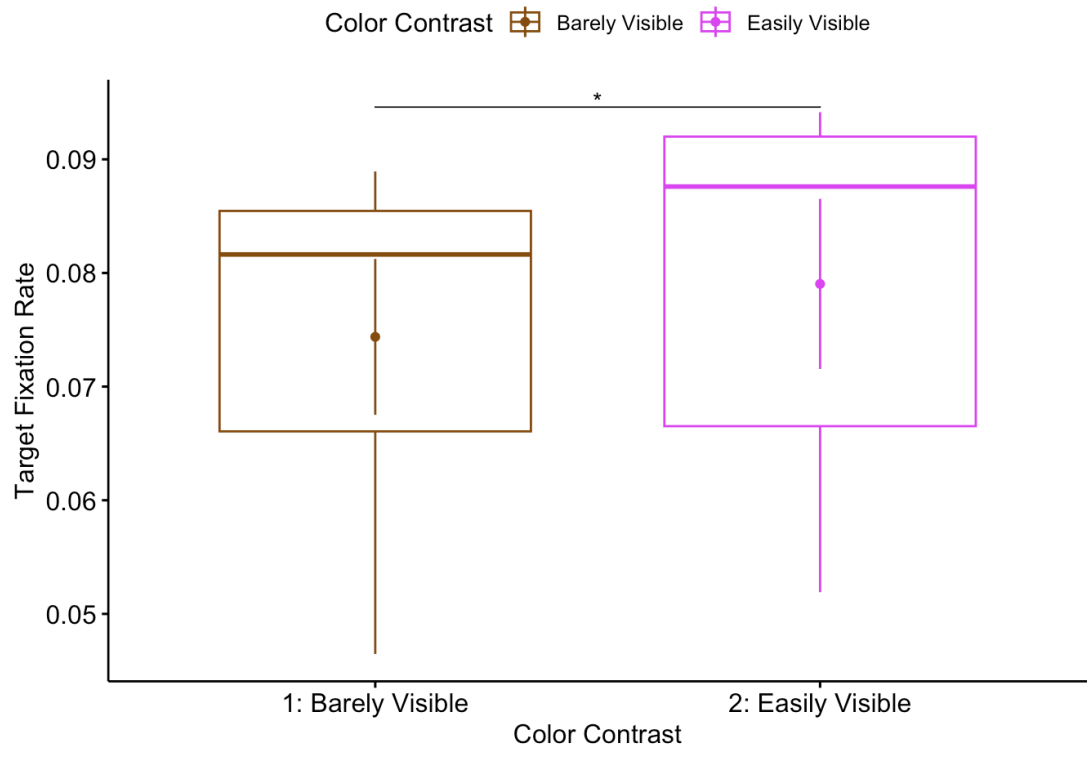


Figure 36. Target fixation rate across different levels of color contrast

A 2 x 2 x 2 repeated measures ANOVA was conducted to examine the effects of visual luminance, color contrast, and task type on response duration. Figure 37 shows the corresponding boxplot. The results showed significant main effects of color contrast, $F(1, 5) = 24.428$, $p = 0.004$, partial $\eta^2 = 0.830$. No follow-up test was needed as there were only two levels of color contrast. Hence, the response durations were significantly lower in the easily visible color contrast levels than in barely visible levels, indicating that higher color contrast levels helped participants finish and respond to tasks faster. However, there was no significant main effect of luminance, $F(1, 5) = 2.740$, $p = 0.159$, partial $\eta^2 = 0.354$, suggesting that visual luminance did not significantly influence the response duration. Interestingly, there was a significant interaction effect between color contrast and task type, $F(1, 5) = 16.939$, $p = 0.009$, partial $\eta^2 = 0.772$. This suggests that the effect of color contrast on response duration varied depending on the task type. No other interaction effects were found to be significant (all $p > 0.05$).

Following up on the interaction effect, post-hoc two-tailed paired t-tests at each color contrast level were conducted with Bonferroni correction. The results revealed that response durations were significantly longer in the search task than in the memory recall task in both barely visible ($p = 0.016$) and easily visible ($p = 0.041$) color contrast conditions. Figure 38 shows the corresponding boxplot.

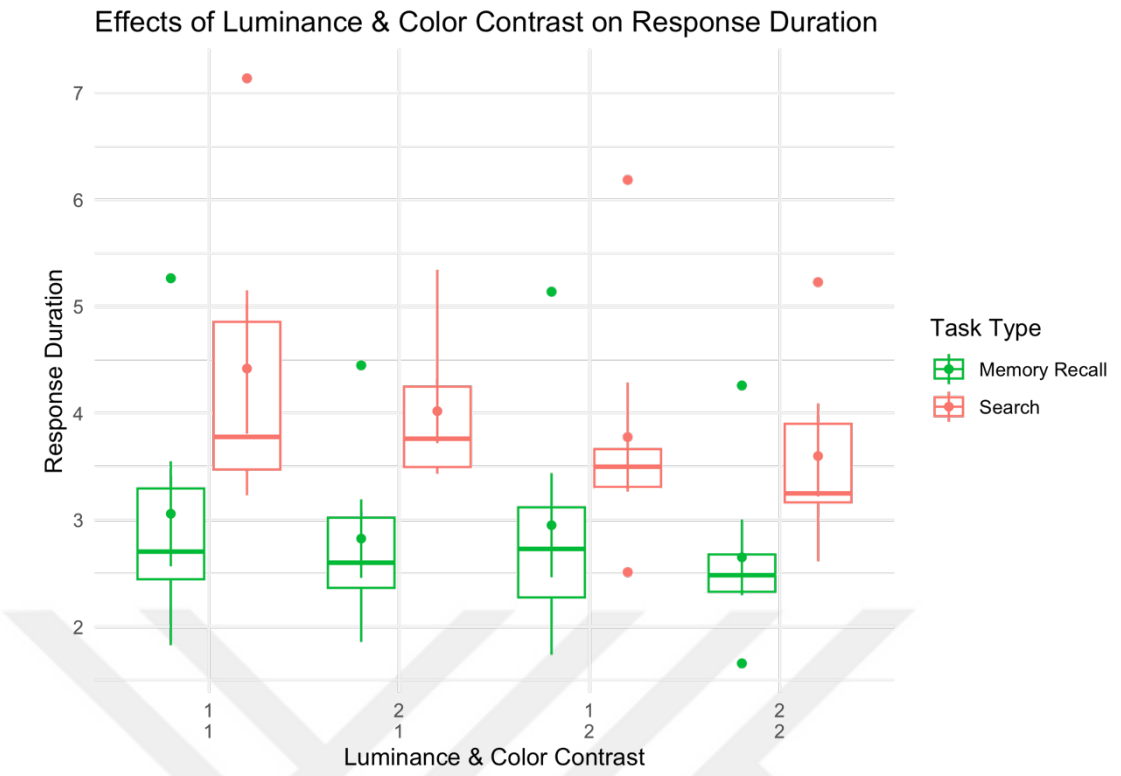


Figure 37. Response durations across luminance and color contrast per task

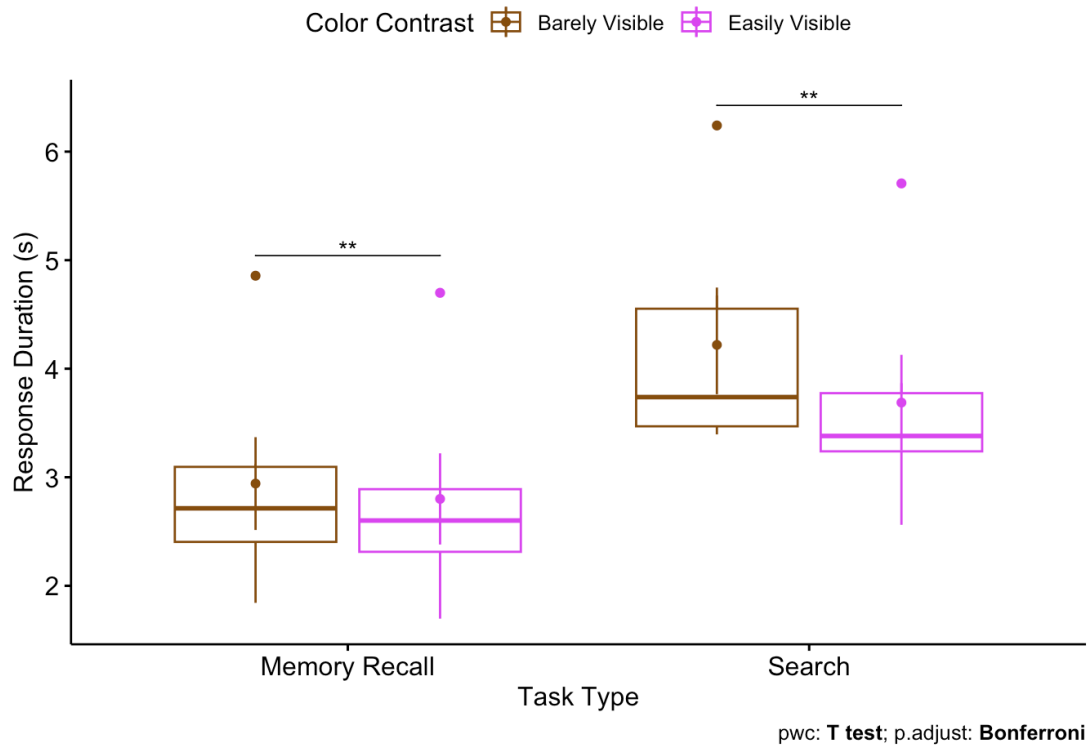


Figure 38. Response durations across different levels of color contrast per task

A 2 x 2 x 2 repeated measures ANOVA was conducted to investigate the effects of visual luminance, color contrast, and task type on response correct rate. Figure 39 shows the corresponding boxplot. The results showed no significant main effects of luminance, $F(1,5) = 1.111$, $p = 0.340$, partial $\eta^2 = 0.182$, or color contrast, $F(1,5) = 5.342$, $p = 0.069$, partial $\eta^2 = 0.517$, suggesting that visual luminance and color contrast did not significantly influence the response correct rate. Additionally, no interactions were found to be significant (all $p > 0.05$).

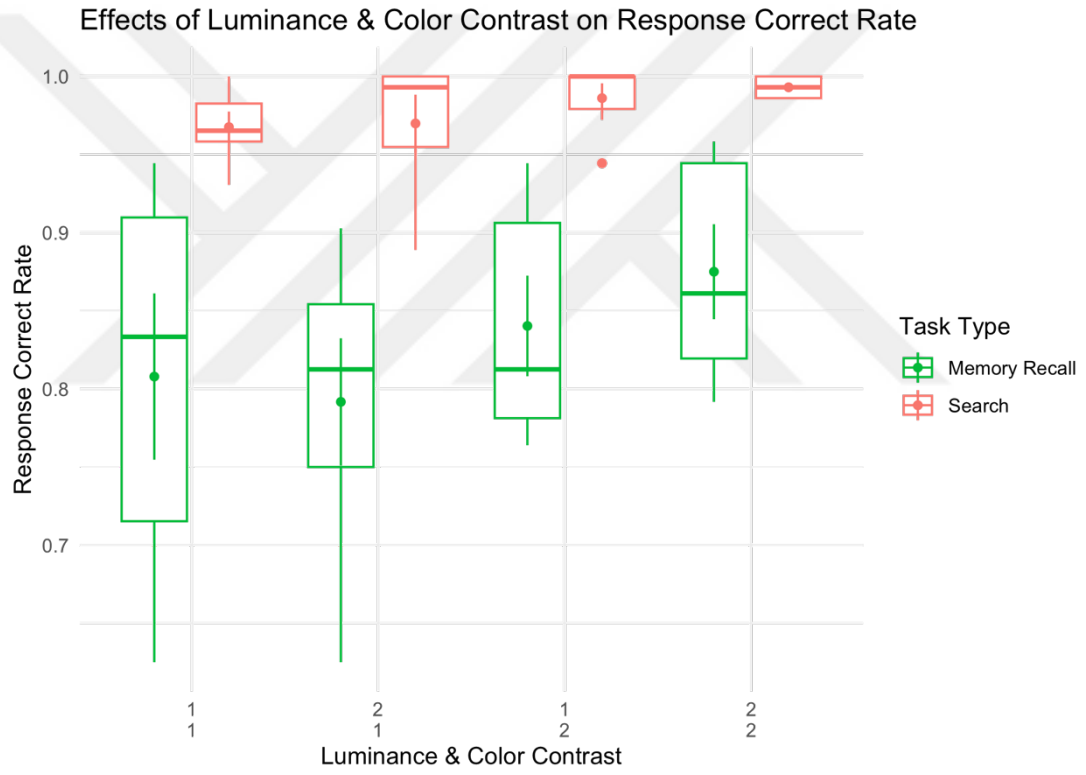


Figure 39. Correct response rate across luminance and color contrast per task

4.4.6 Effects of audio source

A 5 x 3 repeated measures ANOVA was conducted to examine the effects of audio source and task type on average fixation duration. Figure 40 shows the corresponding boxplot. The results showed no significant main effect of the audio source,

$F(4, 20) = 0.689$, $p = 0.608$, partial $\eta^2 = 0.121$. This indicates that the average fixation duration did not significantly differ across the five audio sources. Additionally, no interactions were found to be significant (all $p > 0.05$).

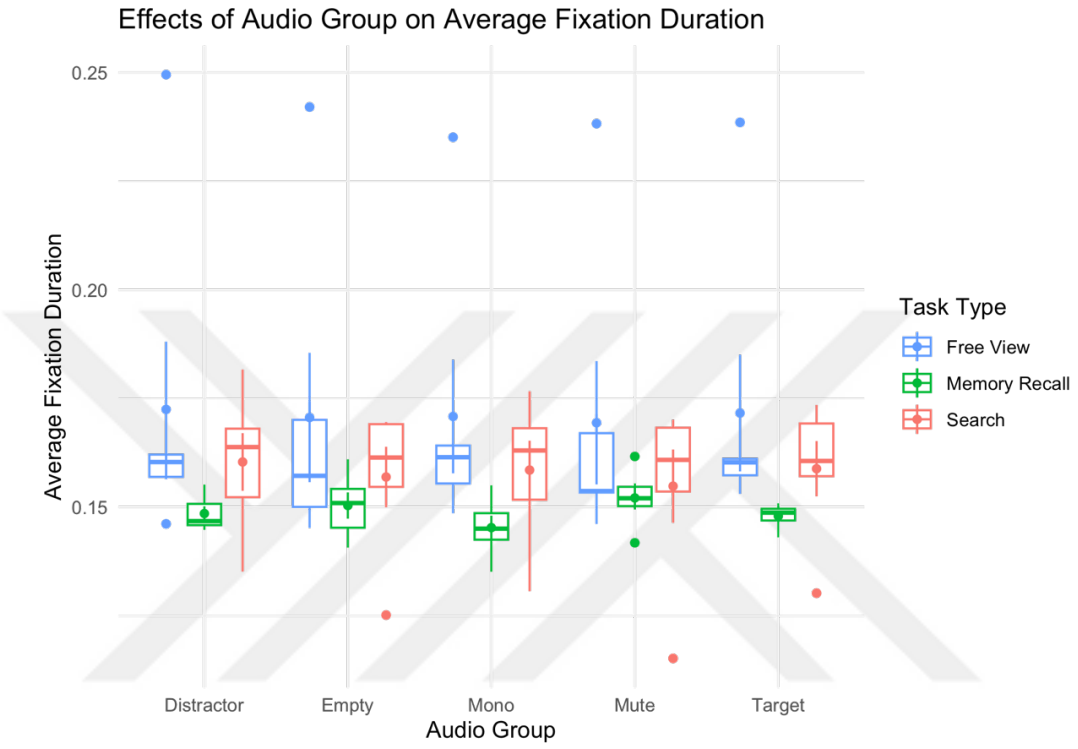


Figure 40. Average fixation durations across audio source groups per task

A 5 x 3 repeated measures ANOVA was conducted to investigate the effects of audio source and task type on target fixation rate. Figure 41 shows the corresponding boxplot. The results showed no significant main effect of the audio source, $F(4, 20) = 0.345$, $p = 0.844$, partial $\eta^2 = 0.065$, suggesting that the target fixation rate did not significantly differ across the five audio groups. Additionally, no interactions were found to be significant (all $p > 0.05$).

A 5 x 2 repeated measures ANOVA was conducted to examine the effects of the audio source and task type on response duration. Figure 42 shows the corresponding boxplot. The assumption of sphericity was violated for the audio

source factor, as assessed by Mauchly's test ($W = 0.000877$, $p = 0.009$). Therefore, a Greenhouse-Geisser correction was applied for the audio source factor. The results showed no significant main effect of the audio source, $F(1.36, 6.79) = 1.444$, $p = 0.285$, partial $\eta^2 = 0.224$ (Greenhouse-Geisser corrected), suggesting that the response duration did not significantly differ across the five audio sources. Additionally, no interactions were found to be significant (all $p > 0.05$).

A 5 x 2 repeated measures ANOVA was conducted to investigate the effects of audio group and task type on response correct rate. Figure 43 shows the corresponding boxplot. The results showed no significant main effect of the audio source, $F(4, 20) = 0.832$, $p = 0.521$, partial $\eta^2 = 0.143$, suggesting that the response correct rate did not significantly differ across the five audio sources. Additionally, no interactions were found to be significant (all $p > 0.05$).

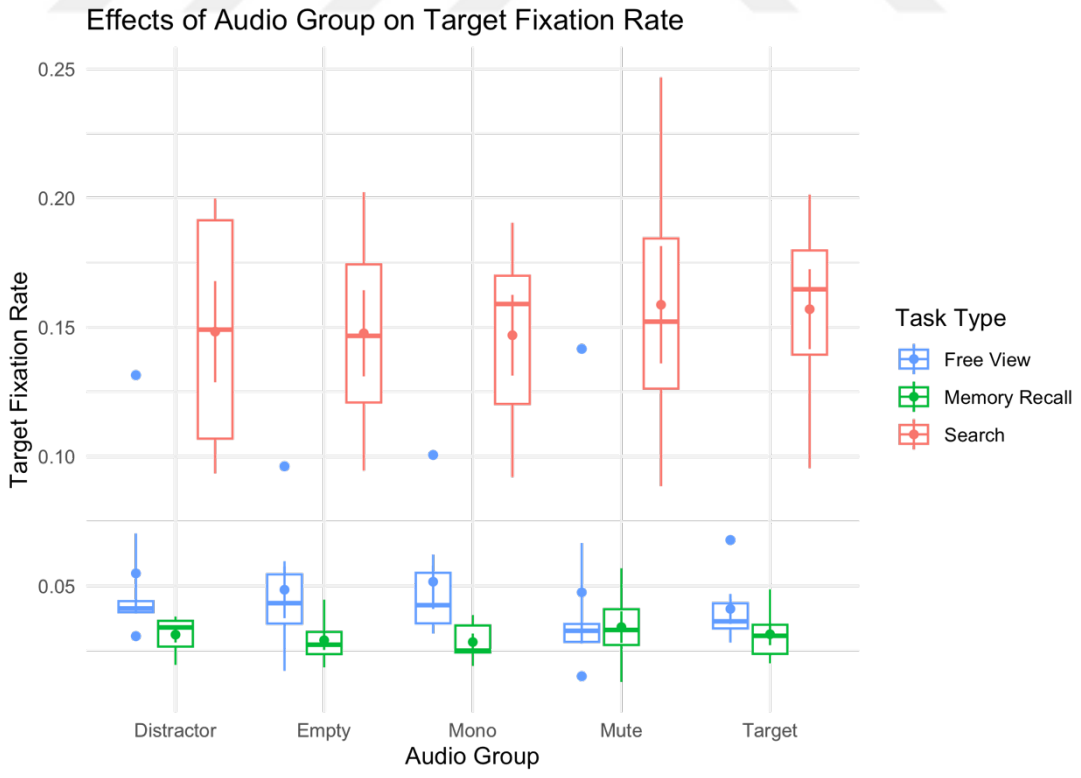


Figure 41. Target fixation rate across audio source groups per task

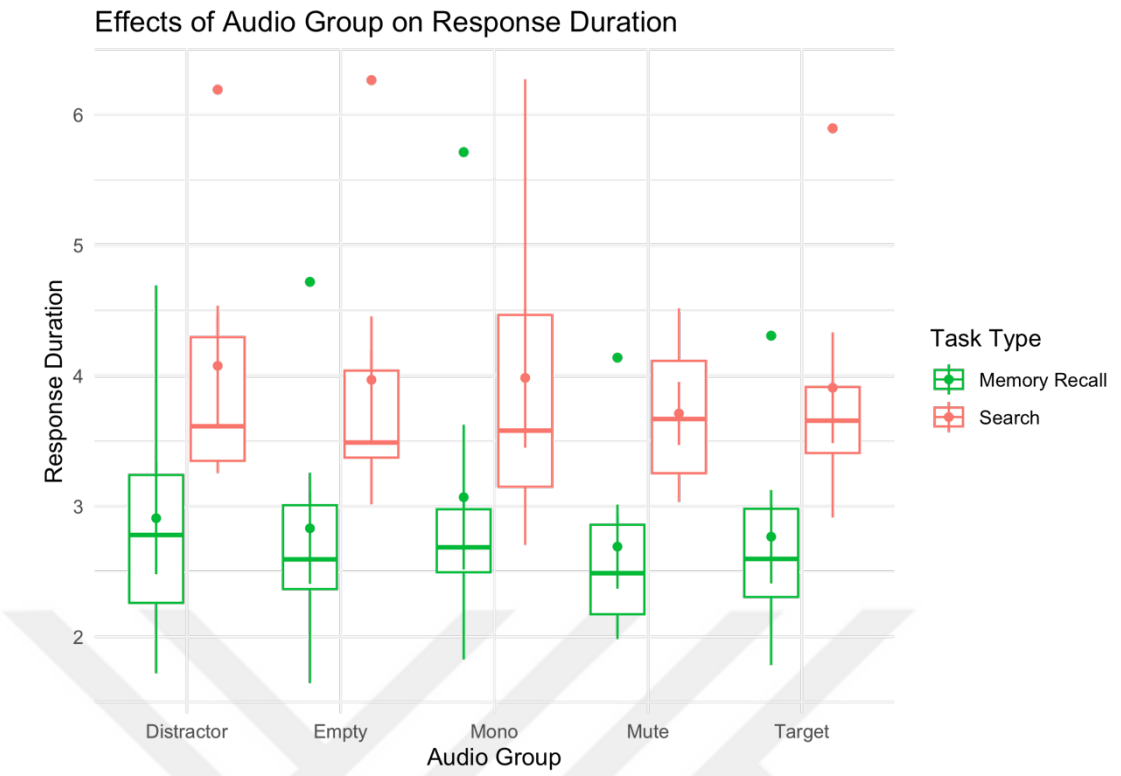


Figure 42. Response durations across audio source groups per task

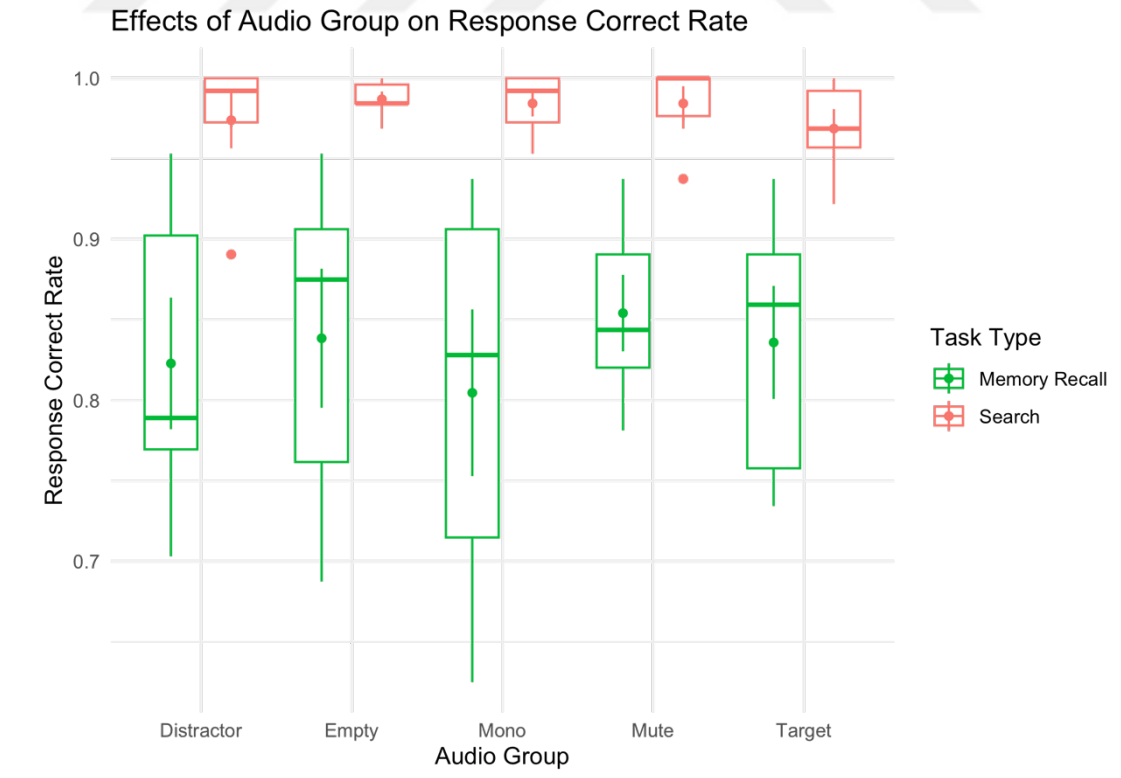


Figure 43. Correct response rate across audio source groups per task

4.4.7 Effects of loudness

A 3 x 3 repeated measures ANOVA was conducted to examine the effects of loudness and task type on average fixation duration. Figure 44 shows the corresponding boxplot. The results showed no significant main effect of loudness, $F(2, 10) = 0.423$, $p = 0.667$, partial $\eta^2 = 0.078$. This indicates that the average fixation duration did not significantly differ across the three levels of loudness. Additionally, no interactions were found to be significant (all $p > 0.05$).

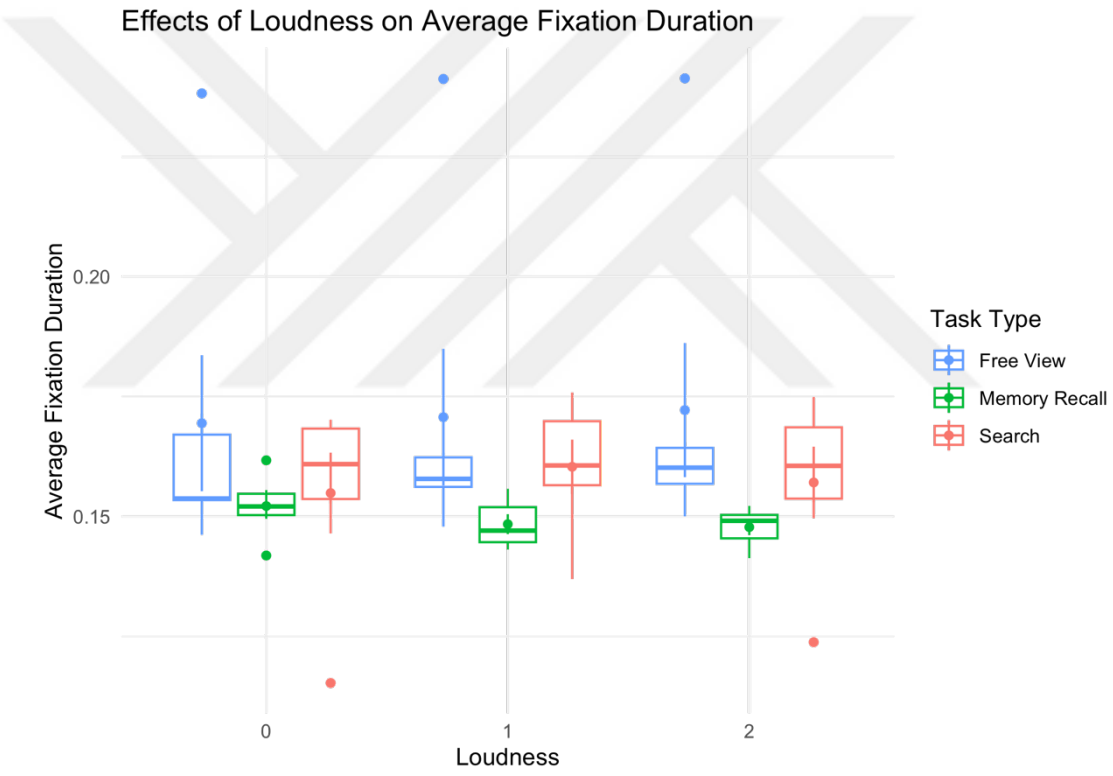


Figure 44. Average fixation duration across different loudness levels per task

In the figure, zero depicts mute, one depicts one times the JND level loudness (barely audible), and two depicts two times the JND level loudness (easily audible).

A 3 x 3 repeated measures ANOVA was conducted to investigate the effects of loudness and task type on target fixation rate. Figure 45 shows the corresponding boxplot. The results showed no significant main effect of loudness, $F(2, 10) =$

0.520, $p = 0.609$, partial $\eta^2 = 0.094$, suggesting that the target fixation rate did not significantly differ across the three levels of loudness. Additionally, no interactions were found to be significant (all $p > 0.05$).

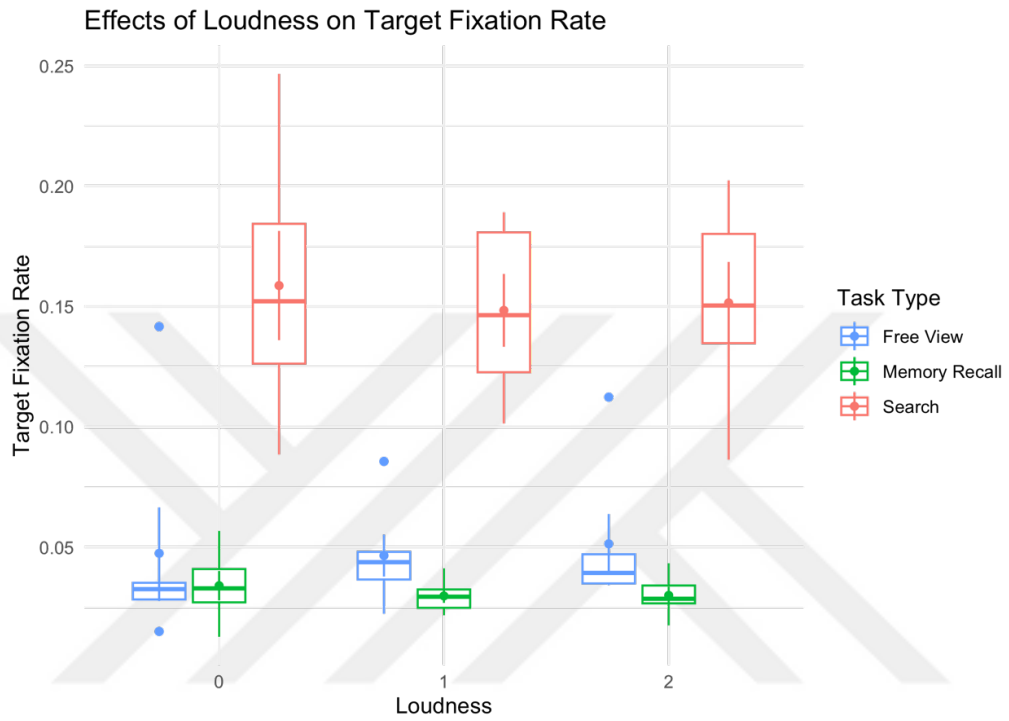


Figure 45. Target fixation rate across different loudness levels per task

A 3 x 2 repeated measures ANOVA was conducted to examine the effects of loudness and task type on response durations. Figure 46 shows the corresponding boxplot. The assumption of sphericity was violated for loudness ($W = 0.070$, $p = 0.005$), as assessed by Mauchly's test. Therefore, a Greenhouse-Geisser correction was applied for the loudness factor. The results showed no significant main effect of loudness, $F(1.04, 5.18) = 1.567$, $p = 0.266$, partial $\eta^2 = 0.239$ (Greenhouse-Geisser corrected), suggesting that the response duration did not significantly differ across the three levels of loudness. Additionally, no interactions were found to be significant (all $p > 0.05$).

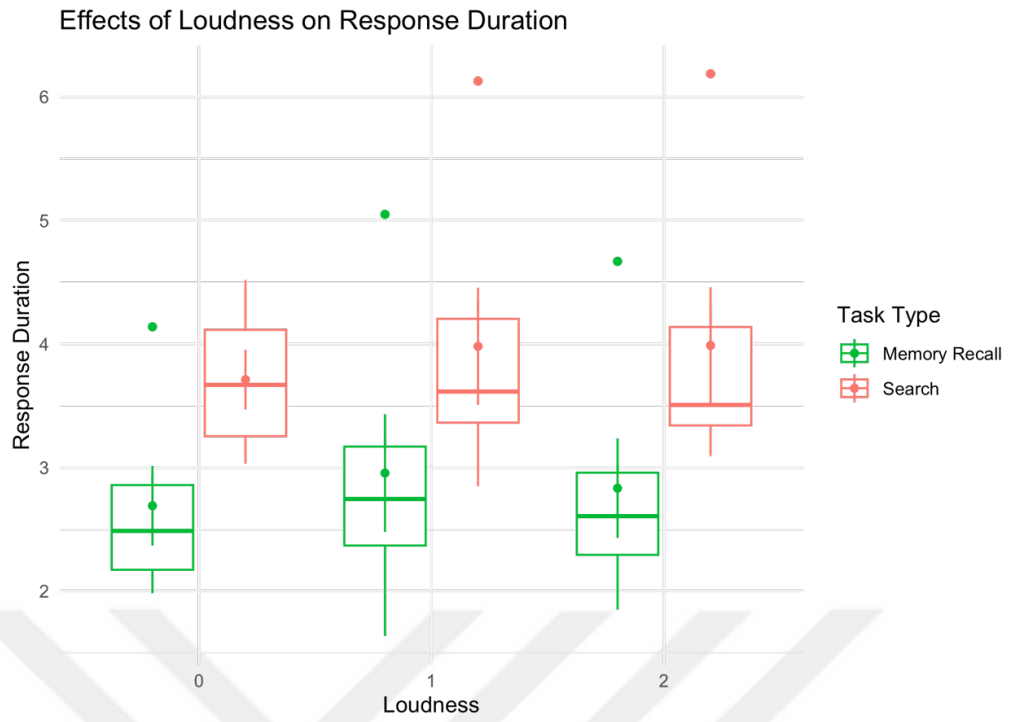


Figure 46. Response duration across different loudness levels per task

A 3 x 2 repeated measures ANOVA was conducted to investigate the effects of loudness and task type on response correct rate. Figure 47 shows the corresponding boxplot. The results showed no significant main effect of loudness, $F(2, 10) = 3.479$, $p = 0.071$, partial $\eta^2 = 0.410$, suggesting that the response correct rate did not significantly differ across the three levels of loudness. Additionally, no interactions were found to be significant (all $p > 0.05$).

A 2x3 repeated measures ANOVA was conducted to examine the effects of loudness and task type on audio element fixation rate. Figure 48 shows the corresponding boxplot. The results showed no significant main effect of loudness, $F(1, 5) = 0.540$, $p = 0.496$, partial $\eta^2 = 0.097$, suggesting that the audio element fixation rate did not significantly differ between the two levels of loudness. Additionally, no interactions were found to be significant (all $p > 0.05$).

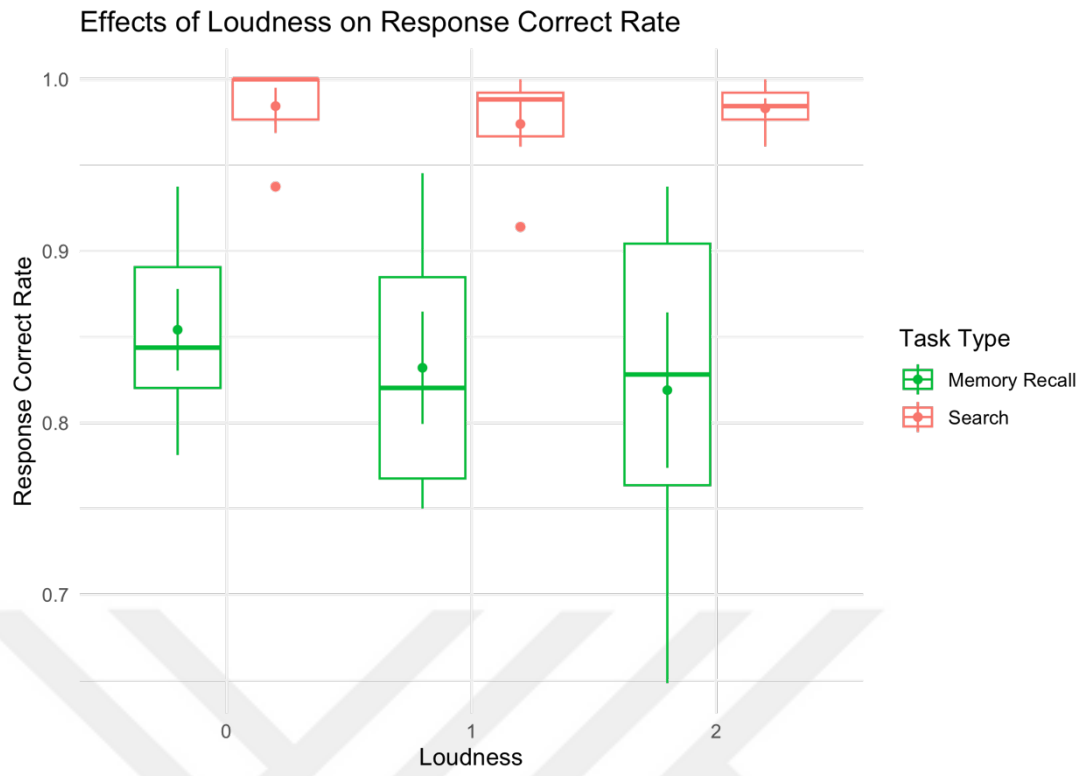


Figure 47. Correct response rate across different loudness levels per task type

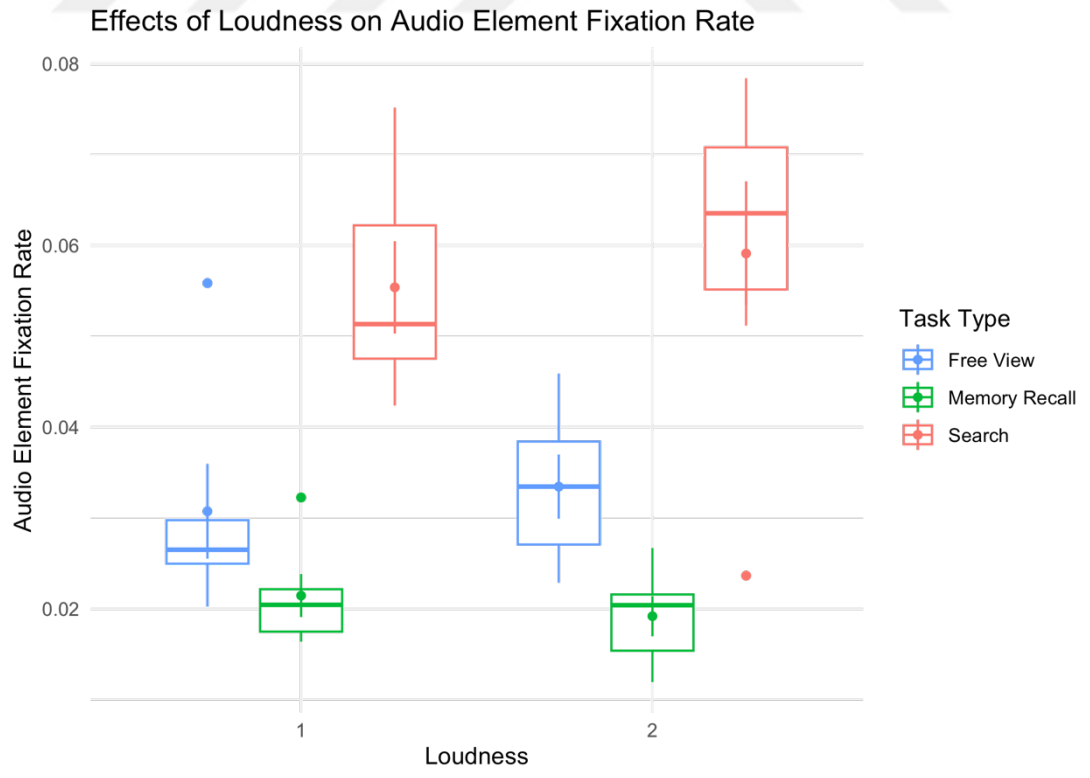


Figure 48. Audio element fixation rate across different loudness levels per task type

4.4.8 Effects of loudness and audio source

A 2 x 4 x 3 repeated measures ANOVA was conducted to investigate the effects of loudness, audio source, and task type on average fixation duration. Figure 49 shows the corresponding boxplot. The results showed no significant main effects of loudness, $F(1, 5) = 1.082$, $p = 0.346$, partial $\eta^2 = 0.178$, audio source, $F(3, 15) = 0.883$, $p = 0.472$, partial $\eta^2 = 0.150$. This indicates that the average fixation duration did not significantly differ across the loudness levels, or audio sources. Additionally, no interactions were found to be significant (all $p > 0.05$).

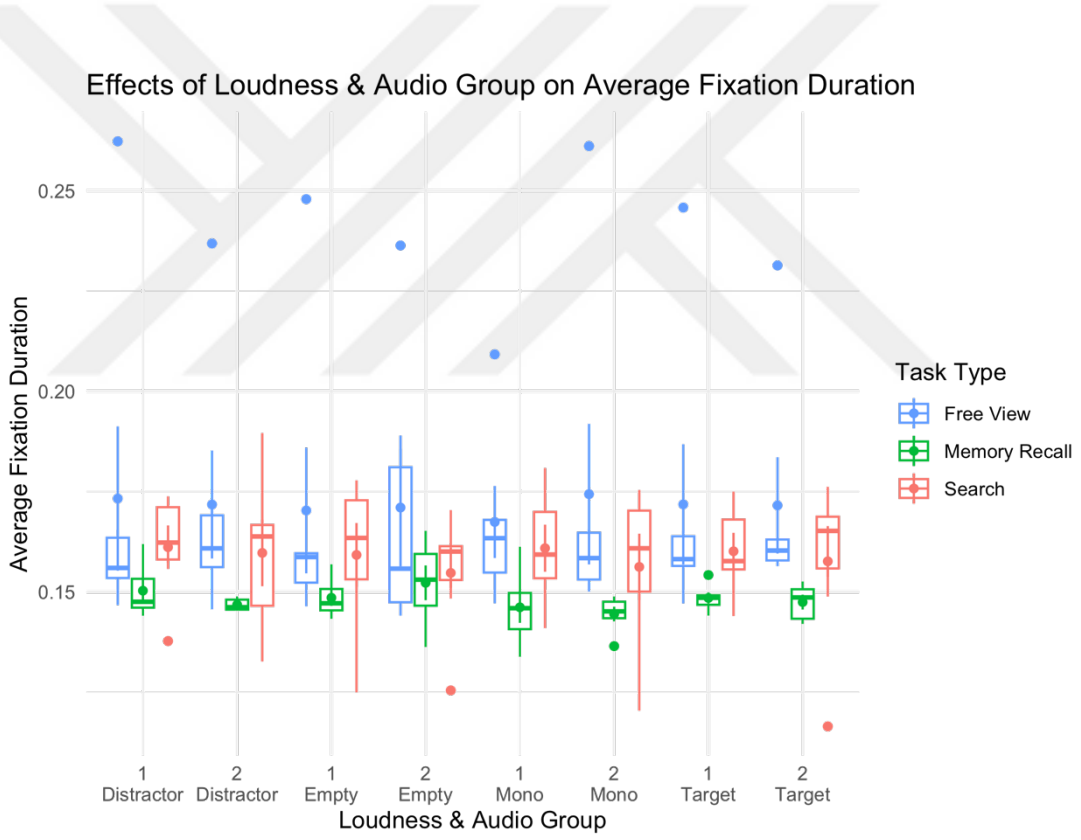


Figure 49. Average fixation duration across loudness and audio source per task

A 2 x 4 x 3 repeated measures ANOVA was conducted to examine the effects of loudness, audio source, and task type on target fixation rate. Figure 50 shows the corresponding boxplot. The results showed no significant main effects of loudness,

$F(1, 5) = 0.810$, $p = 0.409$, partial $\eta^2 = 0.139$, or audio source, $F(3, 15) = 0.225$, $p = 0.878$, partial $\eta^2 = 0.043$, suggesting that the target fixation rate did not significantly differ across the levels of audio amplitude or audio sources. Additionally, no interactions were found to be significant (all $p > 0.05$).

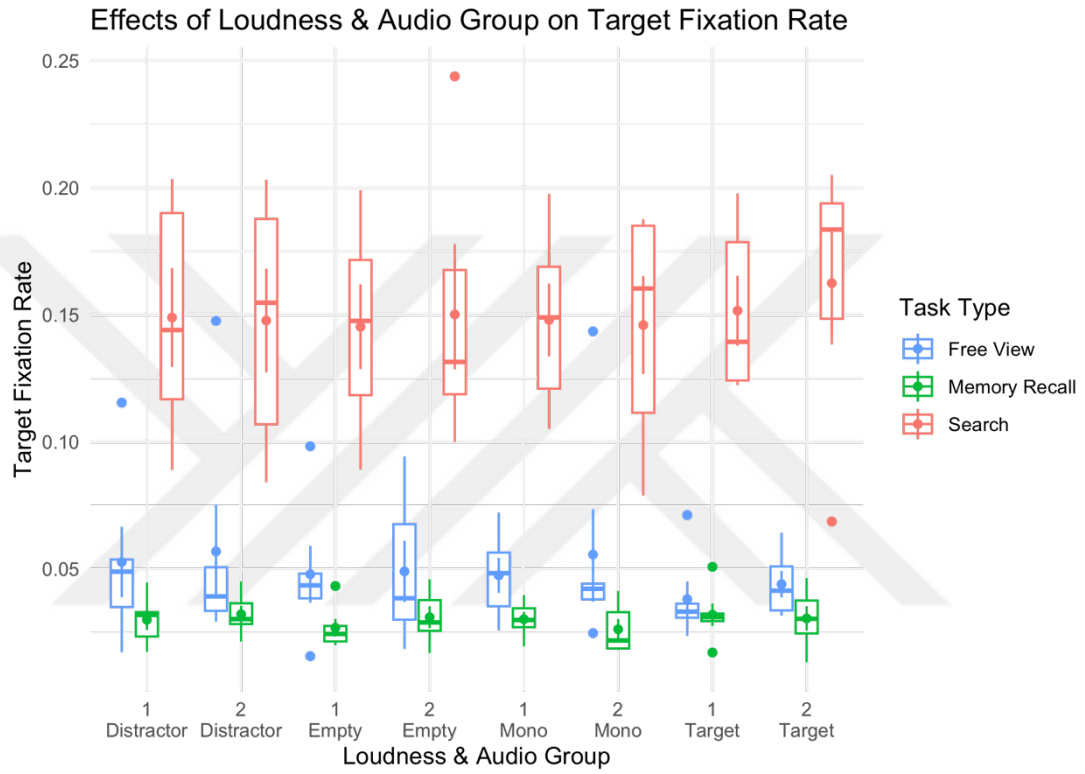


Figure 50. Target fixation rate across loudness and audio source per task

A 2 x 4 x 2 repeated measures ANOVA was conducted to investigate the effects of loudness, audio source, and task type on response duration. Figure 51 shows the corresponding boxplot. The assumption of sphericity was violated for the interaction between loudness and audio source, as assessed by Mauchly's test ($W = 0.020$, $p = 0.016$). Therefore, a Greenhouse-Geisser correction was applied for this interaction. The results showed no significant main effects of loudness, $F(1, 5) = 2.553$, $p = 0.171$, partial $\eta^2 = 0.338$, or audio source, $F(3, 15) = 1.283$, $p =$

0.316, partial $\eta^2 = 0.204$, suggesting that the response duration did not significantly differ across the levels of audio amplitude or audio sources. Additionally, no interactions were found to be significant (all $p > 0.05$).

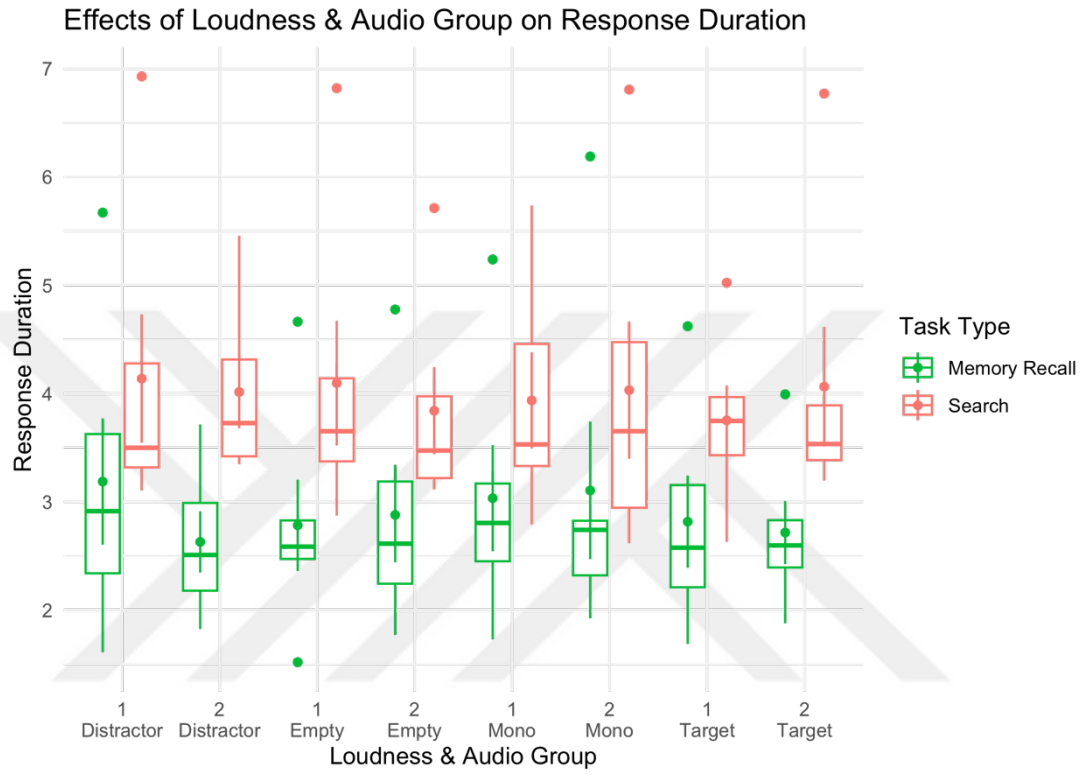


Figure 51. Response duration across loudness and audio source per task

A 2 x 4 x 2 repeated measures ANOVA was conducted to examine the effects of loudness, audio source, and task type on response correct rate. Figure 52 shows the corresponding boxplot. The results showed no significant main effects of loudness, $F(1, 5) = 0.126$, $p = 0.737$, partial $\eta^2 = 0.025$, or audio source, $F(3, 15) = 0.403$, $p = 0.753$, partial $\eta^2 = 0.075$, suggesting that the response correct rate did not significantly differ across the levels of audio amplitude or audio sources. Additionally, no interactions were found to be significant (all $p > 0.05$).

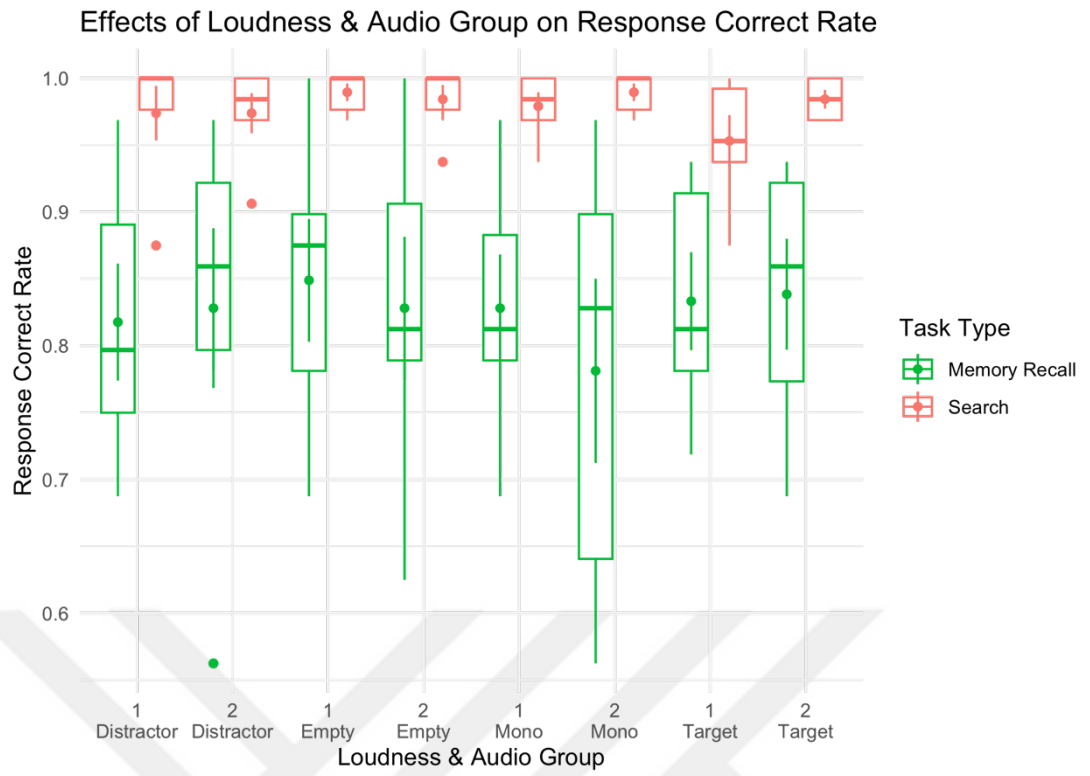


Figure 52. Correct response rate across loudness and audio source per task

4.5 Discussion

As expected, response durations and correct response rates varied across memory and search tasks. In memory recall, the participants needed to remember all four positions of visual stimuli along with their colors compared to remembering a single target color; the memory recall task was designed to be harder than the search task. Likewise, the correct response rate was significantly lower in memory recall. In the search task, the participant needed to look around and find the correct stimuli in order to respond, whereas, in the memory recall task, the participant could almost instantly respond if they remembered the target's location. Therefore, response durations for the search task were significantly higher than memory recall, as expected by the nature of these tasks. Thus, the tasks were sufficiently different from each other. The target fixation rate was significantly higher in the search than in

other tasks, again as expected, because the participant did not need to fixate on other stimuli once they found the target. In order to quickly scan all positions, we expected the average fixation duration to be lower in the memory recall task, but even though it was somewhat lower, it did not significantly differ across tasks. Likewise, the average fixation durations did not change through any other condition modifications, suggesting that neither the time pressure from the task nor the low-level factors of the stimuli can affect average fixation durations during exploration behavior.

The position of the visual stimuli did not significantly affect the correct response rate in any of the tasks, indicating that the participants did not employ any strategy to prioritize any positions but tried to maximize their overall scores by attending all positions. The response durations were significantly lower when the target stimuli were shown front rather than behind. This effect was more strong in the search task. Moreover, the results showed that the fixation rate on targets presented from the front was significantly higher than those presented from behind. Also, this effect was diminished in the memory recall task and increased in the search task. It suggests that in the memory recall task, participants explored all positions equally with the time pressure induced by the task. The increase in search, however, can be explained by the decreased response duration as the participant did not need to look at other stimuli, overall suggesting that the participants searched the target more commonly from starting in the front. Vertical or horizontal target position did not have any significant effect.

Luminance levels of the visual stimuli did not have any significant effect, indicating the participants were more focused on the colors as the tasks required discriminating colors but nothing special about the luminance. Participants had a significantly higher target fixation rate in easily noticeable color contrast levels

compared to barely noticeable color contrast levels. This could be explained by the participants struggling less to find the exact positions of the stimuli and, hence, attending to empty space less when the color contrast level is higher. Additionally, participants responded significantly more quickly with easily noticeable color contrast than barely noticeable color contrast. Similarly, the correct response rates were higher in easily noticeable color contrast but not statistically significant. These results are expected as the color contrast levels both change the difficulty of the task and the saliency of the visual stimuli.

Interestingly, neither the position of the sound (whether it came from the target, distractor, empty space, or mute) nor its loudness nor their interaction had any significant effect. It might be because the participants had a task related to color, reducing any attention to other features. Alternatively, different analyses via other derived measurements from eye tracking data such as saccade behavior, fixation orderings, or scan paths might still reveal effects of luminance, audio conditions, loudness, and audio directionality.

Overall, our findings suggest that the commonly referred low-level factors on saliency are only effective as long as they are task-dependent. Moreover, although the stimuli in the front are easier to fixate on than behind, humans can attend all directions equally if their task requires so.

CHAPTER 5

CONCLUDING REMARKS

Overall, the human attention system, or even saliency as a subset, is a very complex phenomenon. Therefore, it requires many case-specific analyses to develop a thorough understanding. All research conducted within the scope of this thesis covers several portions of those since covering it as a whole and simultaneously in-depth is infeasible by a single researcher. In this manner, the dataset created in Chapter 3 was designed to accumulate as much relevant information as possible to provide a systematic, structured, and general-purpose resource with diverse content to support various future research by other researchers. The saliency modeling literature lacks but requires similar curated datasets, especially for 360°-degree spatial audio videos with eye-tracking data. Therefore, we aspire to help ameliorate the literature by providing the prepared resource.

Virtual reality is a remarkable tool for conducting experiments due to its ecological validity and the ability to bring immersive experiences to participants. Furthermore, embedded eye trackers enable a tremendous amount of new research opportunities. Nonetheless, their use in psychophysics literature has only recently grown; therefore, most of the psychological research tools still need to mature. Therefore, the experiment in Chapter 4 not only helped investigate our research questions but also provided methodological advancements by adapting two-dimensional stimuli used by psychophysics research to three dimensions and establishing guidelines on how to conduct similar research.

Regarding the actual research questions investigated throughout this thesis, there are several findings and several new research opportunities. To begin with,

active tasks play a major role in determining how the exploration behavior of humans is characterized, therefore completely replacing saliency mechanisms, possibly by re-classifying low-level features according to relevance to both active and future tasks. This idea comes from the results of Chapter 4, in which the participants were informed of all the tasks at the beginning, and although the color feature of the stimuli was only relevant to memory recall and search tasks, the analyses only revealed the effects of color feature for the free viewing (no active task) task as well, indicating that auditory stimuli and the controlled luminance levels played surprisingly no role in the participants' eye fixation behavior. Similar research with uninformed or misinformed designs could be conducted to investigate the effects of active and future tasks on saliency and exploration behavior. Likewise, the research can be repeated with tasks focusing on different features of stimuli (such as using luminance or shapes instead of colors) can be conducted to further support our prediction. Finally, the research task can be adapted to other modalities, such as replacing the tasks to search and remember the locations of simultaneous auditory stimuli with varying characteristic features (such as pitch and loudness).

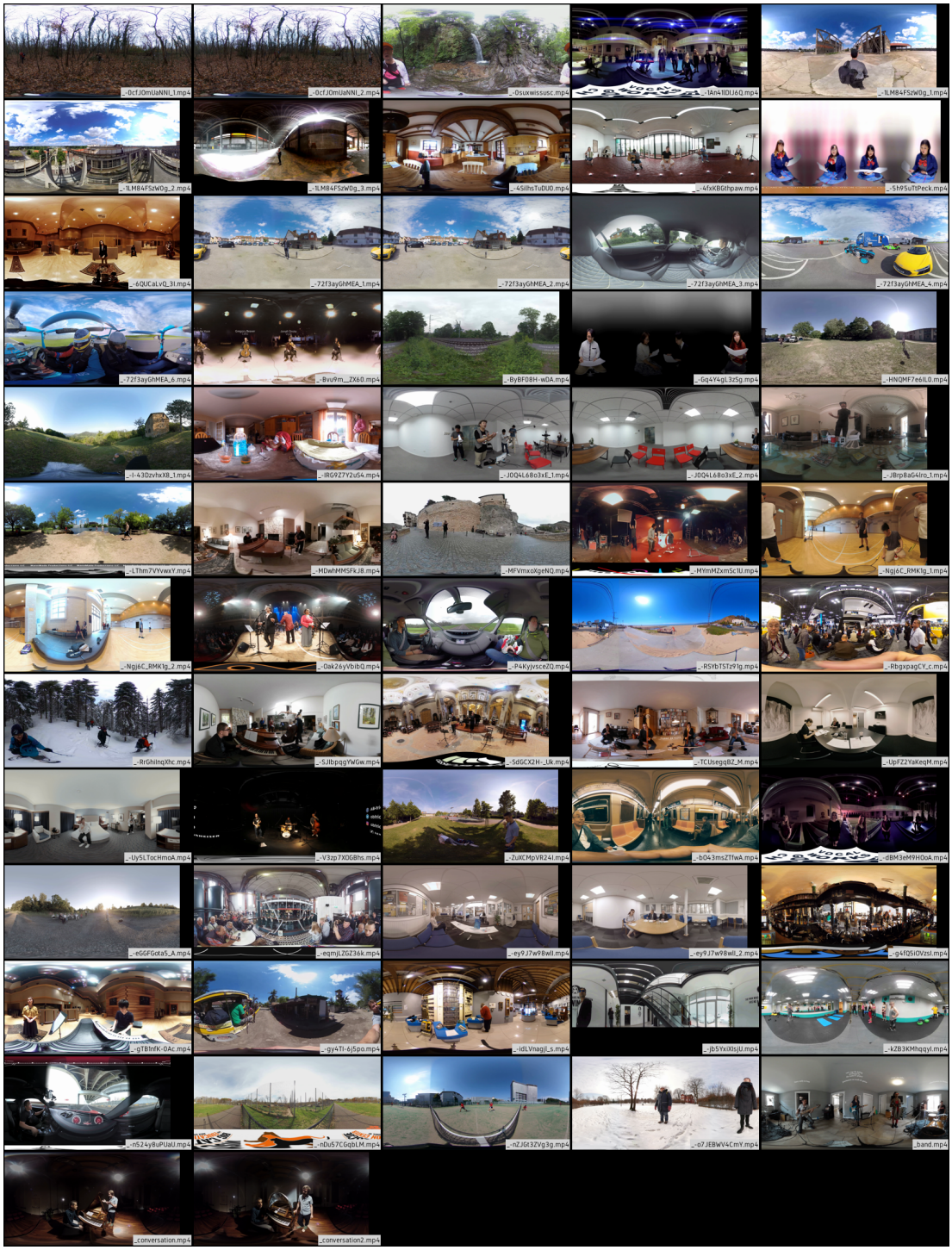
Regarding the computer vision literature, saliency modeling research is mainly conducted within free-view task conditions. Therefore, the results of the experiment in Chapter 4 hint that introducing various tasks could completely change the actual human exploration behavior and fixated regions, rendering current saliency models useless but also offering an opportunity to develop even more sophisticated saliency models. Even though contemporary research started incorporating several semantic estimations as factors in proposed saliency models instead of direct signal processing methods, the findings in Chapter 3 show that saliency is a complex phenomenon that requires separate investigations and possibly models based on

different scenes, audio categories, and semantic content. More specifically, observer exploration behavior, as captured by inter-subject-consistency scores, completely changed across different audio and place categories. Therefore, more sophisticated complex saliency models can be created by first detecting the contents of the scene to match a previously defined (either through fixed categories or unsupervised clustering methods) resemblance group of multi-modal scenes and only then estimating saliency via the saliency model designed per resemblance group. Finally, the collected qualitative feedback data pointed out that participants had very mixed preferences over video content, which, in turn, affected their exploration behavior and fixated regions when they viewed the videos. Therefore, it is noteworthy to say that although saliency models can assume this effect is randomly distributed over participants and, therefore, be modeled as random error, scan path modeling might also benefit from estimating user preferences from prior scan behaviors.

As an attempt to explain the analyses' results from all chapters, similar to the approach in updated guided search model by Wolfe (2021), we theorize that saliency is a complex phenomenon based on evaluating bottom-up signals selected by both active and planned tasks, personal preferences, multi-modal scene characteristics, and semantic content. Unless these factors determine it as irrelevant, a multi-modal stimulus with meaningful content continues to drive the exploration behavior of observers even when its signal features are reduced while its modalities are kept intact until the signal is completely diminished.

APPENDIX A

VIDEO CLIPS IN ASOD60K DATASET

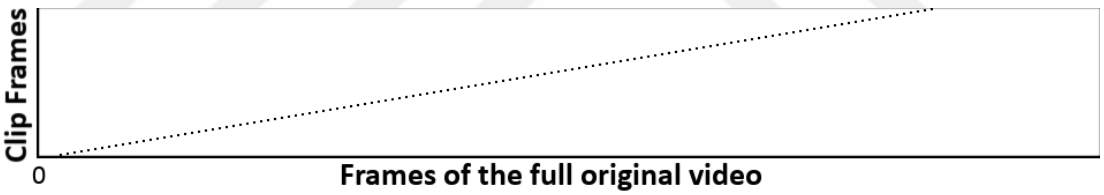


Source: [ASOD60K, 2021] Thumbnails of all 62 video clips in ASOD60K (Yi Zhang, 2021) dataset used in the experiment is presented as a collage. The frames used as thumbnail images are taken from the middle of each video clip.

APPENDIX B

VIDEO FRAME MATCHING METHOD

The video frame matching method works as follows. After downloading each video, in order to match a video clip with its full version, the most significant bits in three color channels for each frame of the video clip and its corresponding full version are extracted. Next up, using the binary XOR function, a most similar frame from the full video is estimated for each frame of the video clip by comparing those extracted bits of the frame to that of the full video's frame. Finally, for each video clip, temporal best-fit lines are calculated using the RANSAC algorithm for these estimated frame matchings to match the video clip as a whole in the full video, finding the most similar clip with the same duration.



An example video clip's frame-matching results over the original uncropped video is shown. Columns (x-axis) correspond to original video frames, rows (y-axis) correspond to frames of the clip. A black pixel indicates matched frames. Shown figure is a typical example of the results, and an almost perfect match for every frame, therefore all points in the figure seem very regular.

APPENDIX C

SIMILARITY METRIC CALCULATIONS

To calculate the similarity metrics, the human eye fixation density maps need to be calculated from the fixations, first. By applying a Gaussian filter to a sum of discrete fixation points in a map, one can obtain a density map; however, the distortions of the equirectangular format also need to be accounted for (Agtzidis & Dorr, 2019). John et al. (2018) reviewed different ways to overcome this problem and proposed a modified Gaussian kernel that comes up with a good trade-off of accuracy and speed. This method was implemented and used for converting fixation data to density maps. Generalizing these comparisons over images to videos is another challenge. To do so, the methods and implementation specified by Salient360! Benchmark was followed such that density maps corresponding to 200ms chunks of the video were compared to fixations corresponding to that period (Gutierrez et al., 2018). Yet, unlike the original implementation, instead of collecting 200ms chunks that are adjacent to one another by the boundary, 200ms chunks that overlap by 100ms with adjacent ones were collected for better interpolation. Moreover, as the original metrics are designed for two-dimensional environments, the implementations in Salient360 which include modified metric calculations for omnidirectional environments and equirectangular images were used. Doing so, AUC, NSS, CC, SIM, and KLD scores for chunks of each video clip comparing fixations when the observers watched it with sound and without sound were calculated. Finally, by averaging scores over chunks, corresponding scores for each video clip were found.

APPENDIX D

CUSTOM VIDEO ANNOTATION WEBSITE TOOL SCREENSHOTS

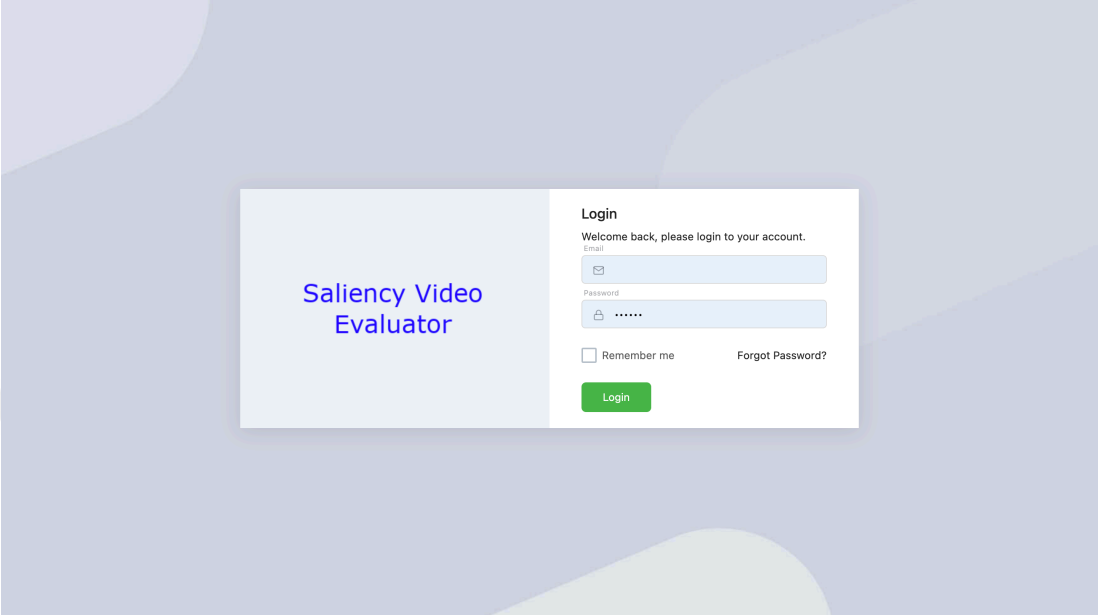


Figure 53. Login page screenshot of the custom video annotation website

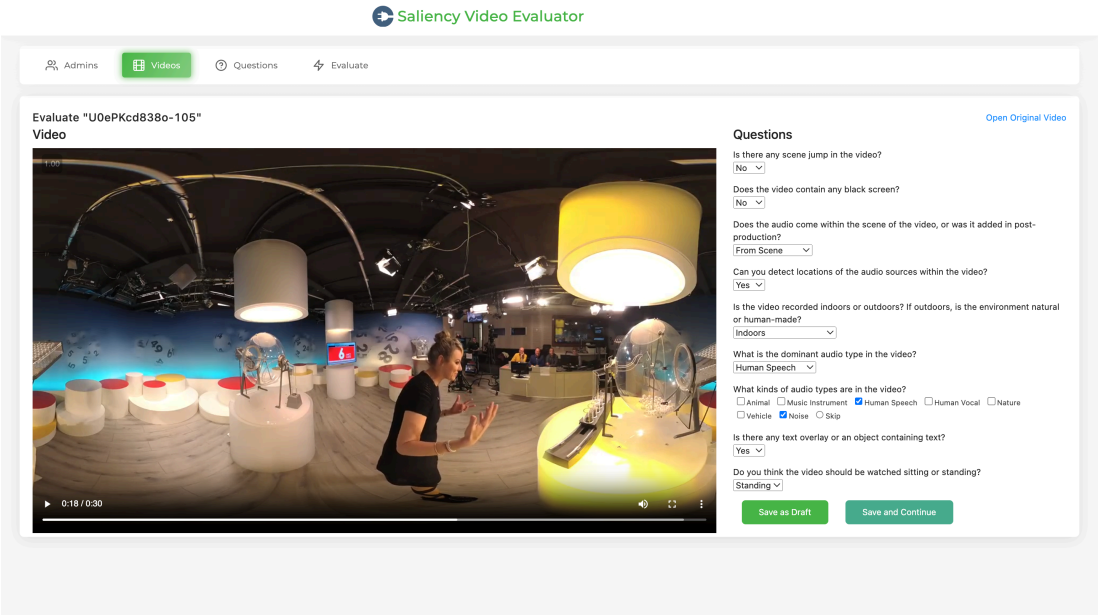


Figure 54. Video annotation page screenshot of the custom website

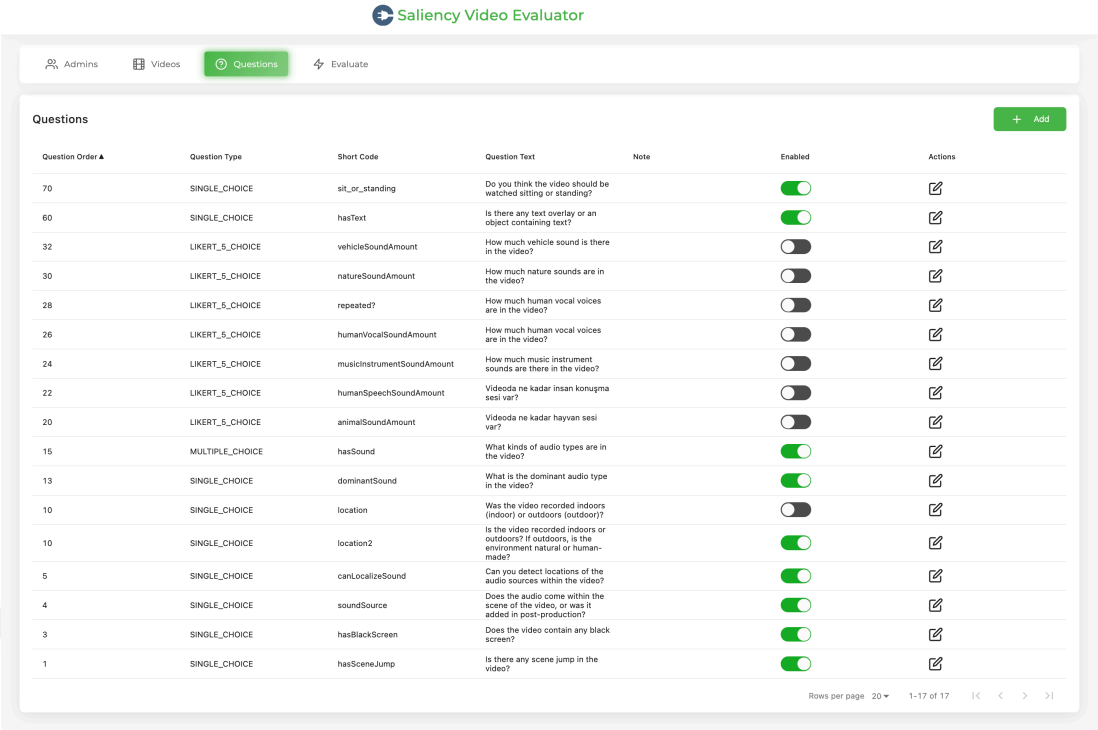


Figure 55. Question list page screenshot of the custom website.

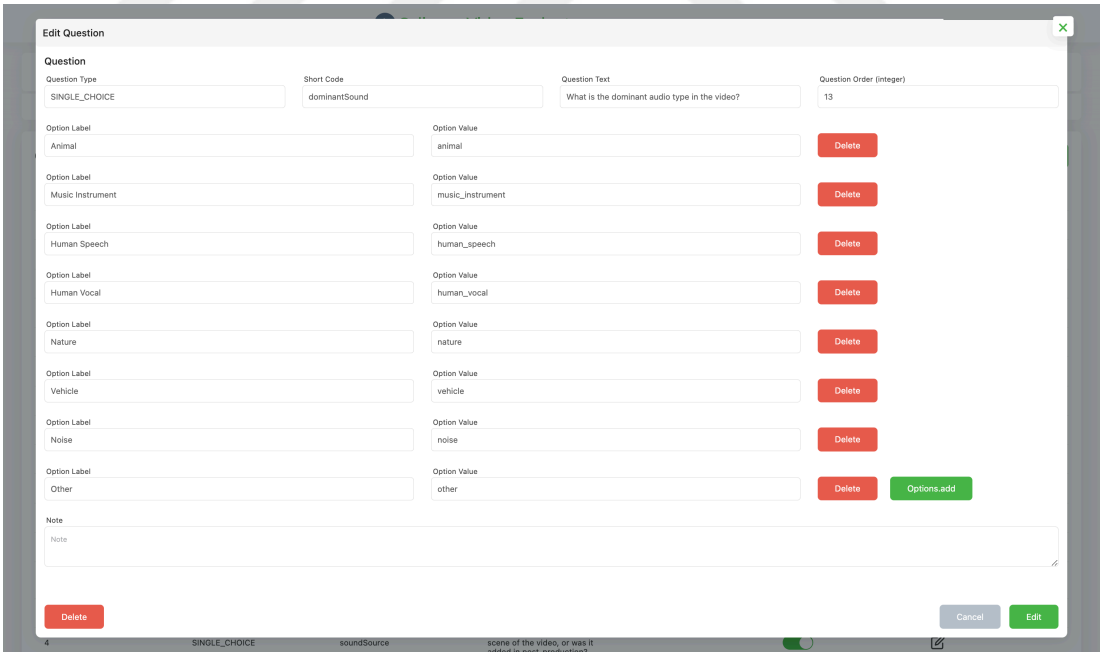


Figure 56. Edit question details modal screenshot of the custom website

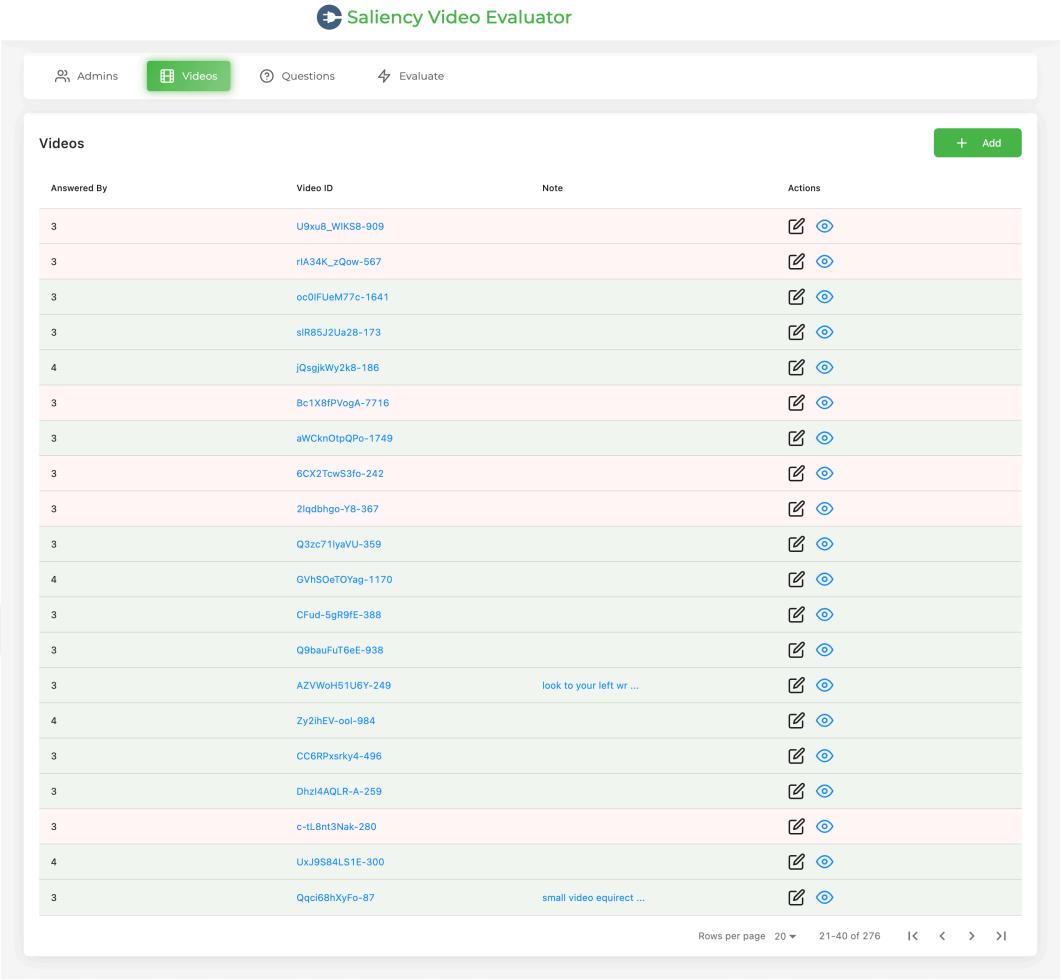


Figure 57. Video list page screenshot of the custom website

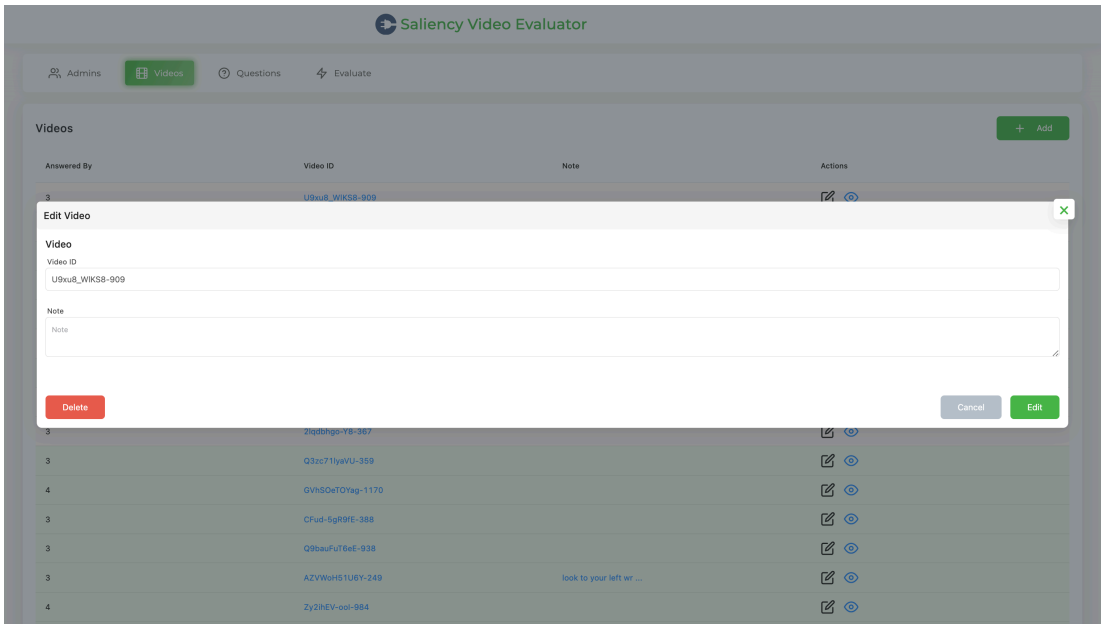


Figure 58. Edit video details modal screenshot of the custom website

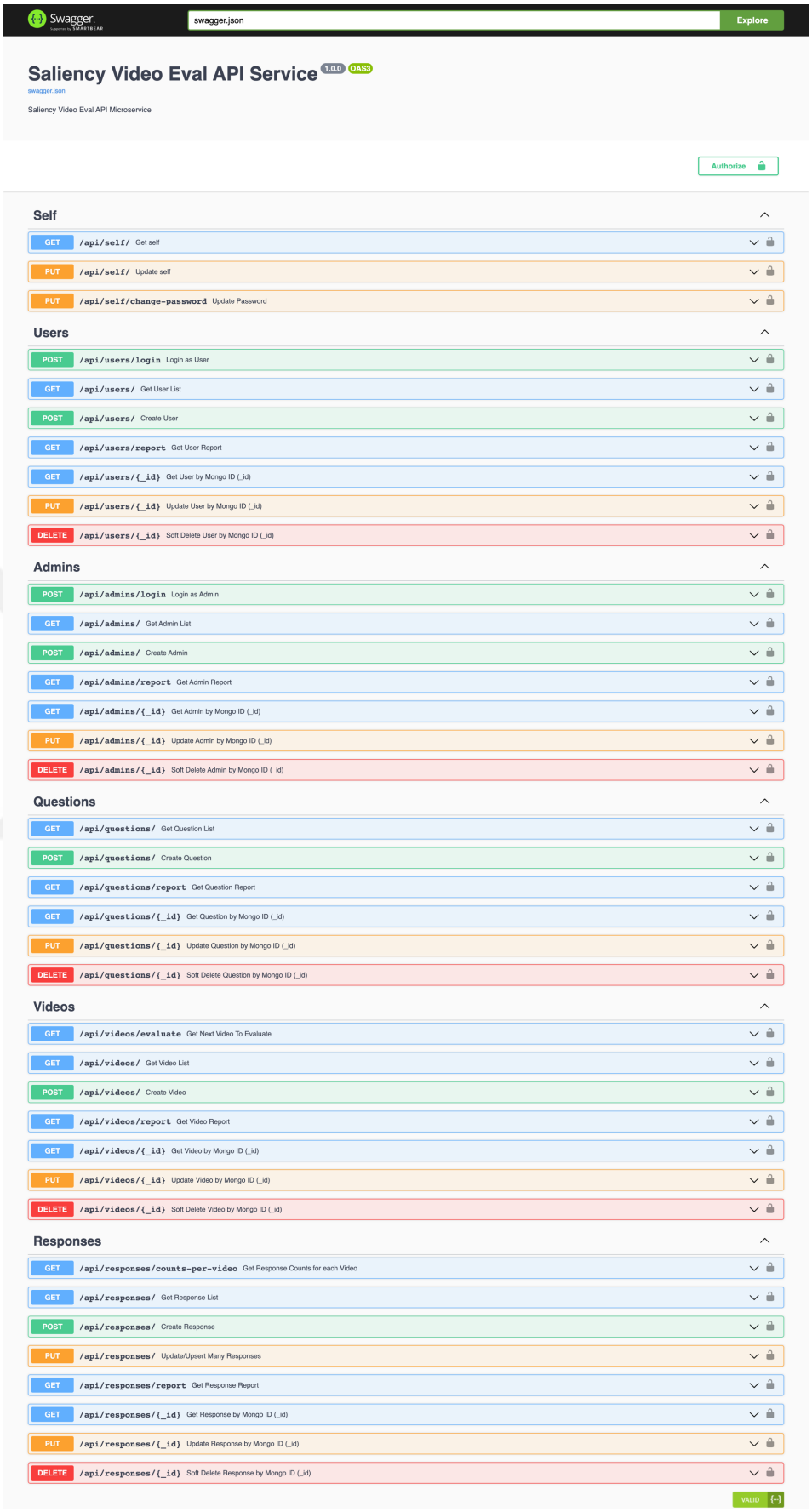
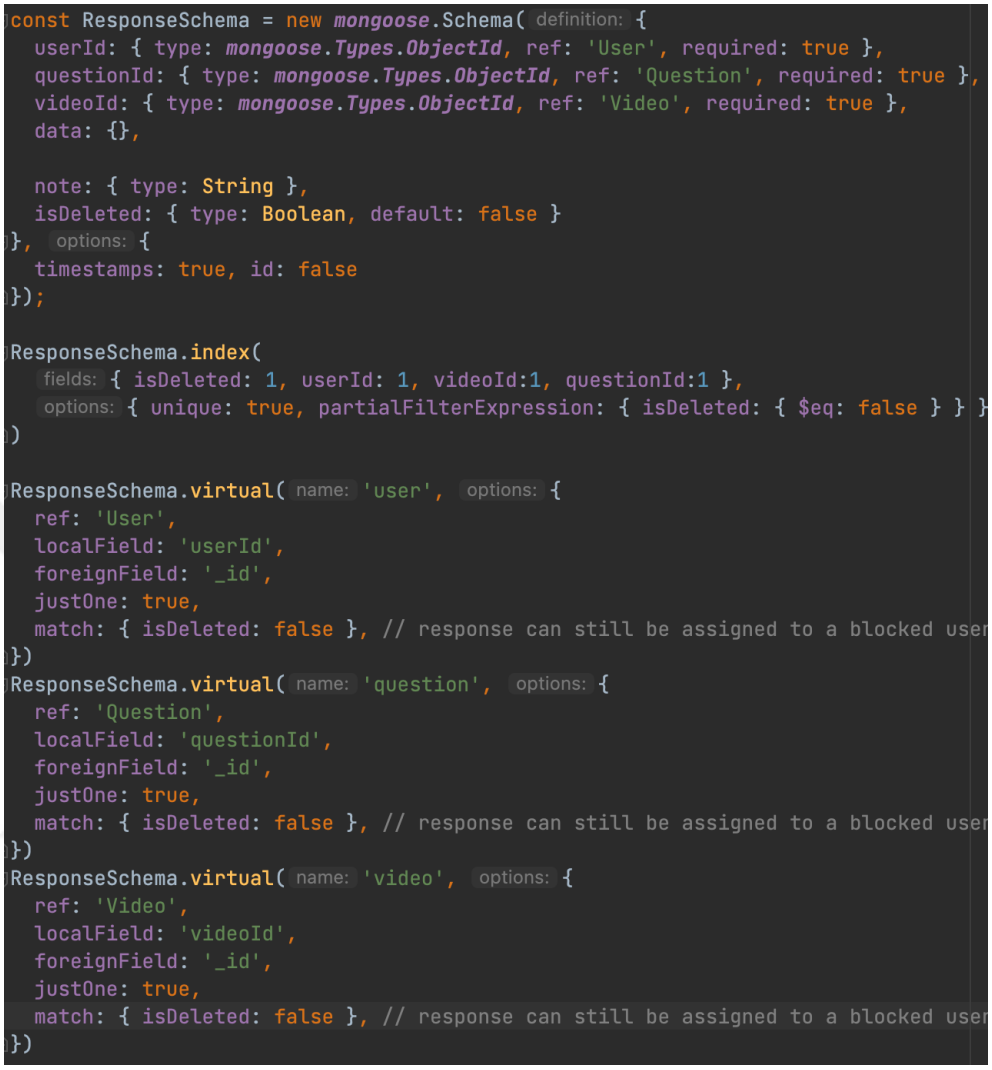


Figure 59. Application programming interface documentation for the website

MongoDB (MongoDB Inc., 2022) was used as the database.



```
const ResponseSchema = new mongoose.Schema( definition: {
  userId: { type: mongoose.Types.ObjectId, ref: 'User', required: true },
  questionId: { type: mongoose.Types.ObjectId, ref: 'Question', required: true },
  videoId: { type: mongoose.Types.ObjectId, ref: 'Video', required: true },
  data: {},

  note: { type: String },
  isDeleted: { type: Boolean, default: false }
}, options: {
  timestamps: true, id: false
});

ResponseSchema.index(
   fields: { isDeleted: 1, userId: 1, videoId:1, questionId:1 },
   options: { unique: true, partialFilterExpression: { isDeleted: { $eq: false } } }
)

ResponseSchema.virtual( name: 'user', options: {
  ref: 'User',
  localField: 'userId',
  foreignField: '_id',
  justOne: true,
  match: { isDeleted: false }, // response can still be assigned to a blocked user
})
ResponseSchema.virtual( name: 'question', options: {
  ref: 'Question',
  localField: 'questionId',
  foreignField: '_id',
  justOne: true,
  match: { isDeleted: false }, // response can still be assigned to a blocked user
})
ResponseSchema.virtual( name: 'video', options: {
  ref: 'Video',
  localField: 'videoId',
  foreignField: '_id',
  justOne: true,
  match: { isDeleted: false }, // response can still be assigned to a blocked user
})
```

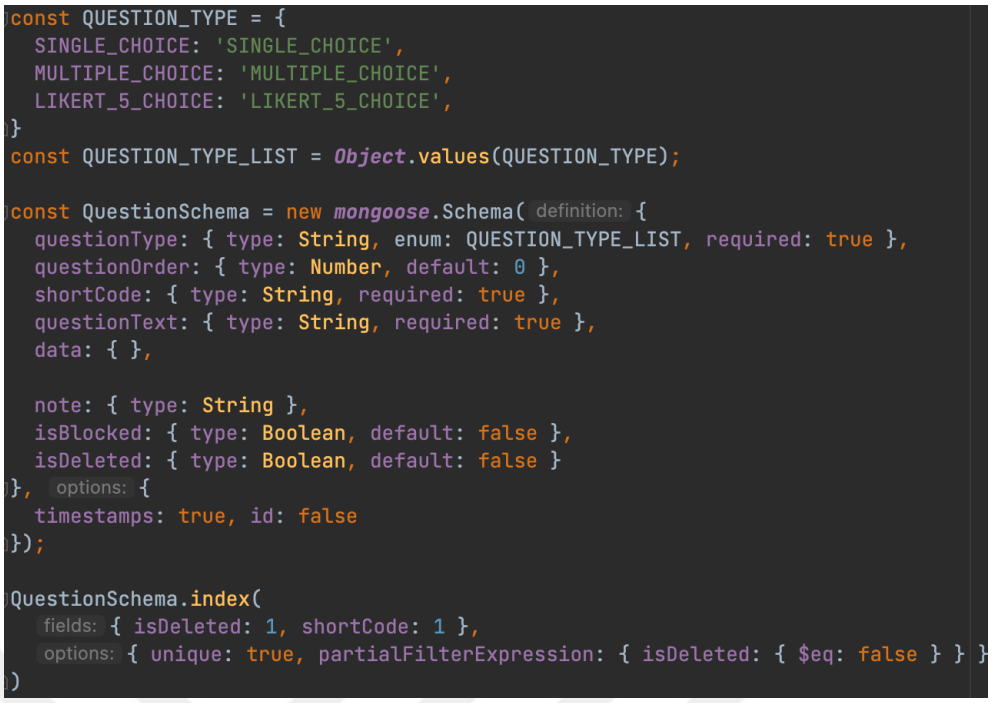
Figure 60. Mongo model definition for responses in the custom website



```
const VideoSchema = new mongoose.Schema( definition: {
  videoId: { type: String, required: true },
  note: { type: String },
  isBlocked: { type: Boolean, default: false },
  isDeleted: { type: Boolean, default: false }
}, options: {
  timestamps: true, id: false
});

VideoSchema.index(
   fields: { isDeleted: 1, videoId: 1 },
   options: { unique: true, partialFilterExpression: { isDeleted: { $eq: false } } }
)
```

Figure 61. Mongo model definition for videos in the custom website



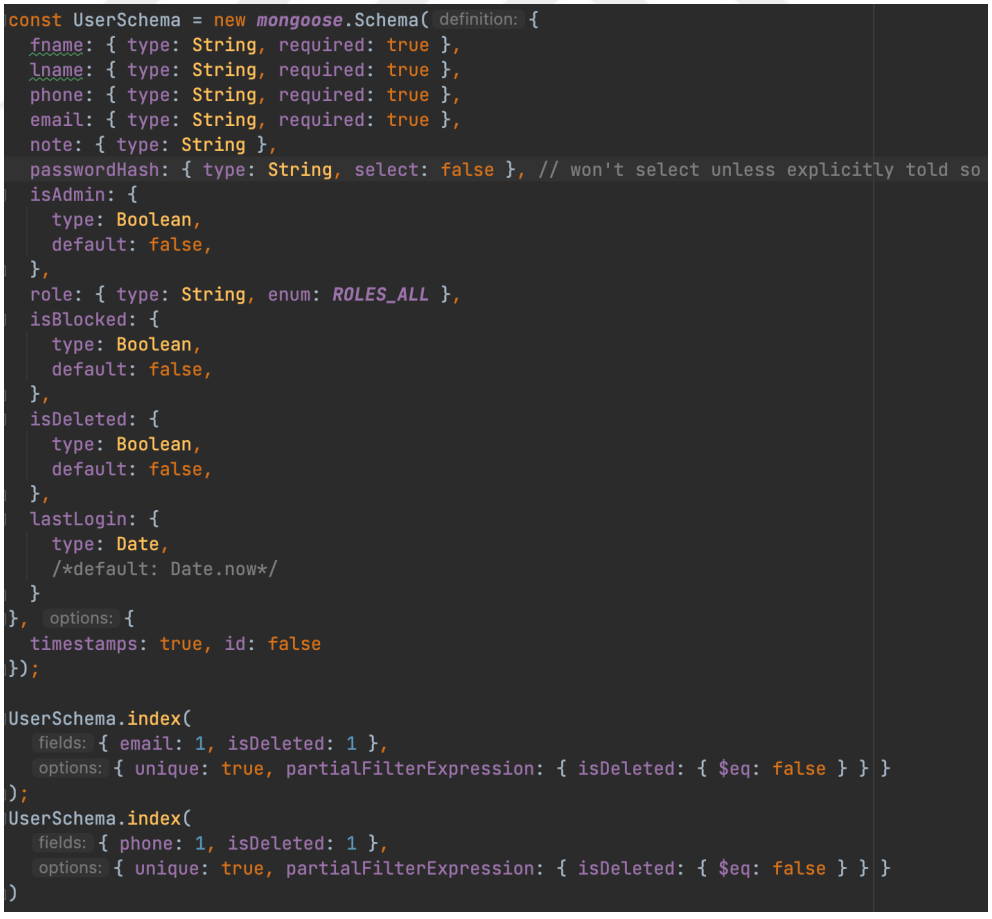
```
const QUESTION_TYPE = {
  SINGLE_CHOICE: 'SINGLE_CHOICE',
  MULTIPLE_CHOICE: 'MULTIPLE_CHOICE',
  LIKERT_5_CHOICE: 'LIKERT_5_CHOICE',
}
const QUESTION_TYPE_LIST = Object.values(QUESTION_TYPE);

const QuestionSchema = new mongoose.Schema( definition: {
  questionType: { type: String, enum: QUESTION_TYPE_LIST, required: true },
  questionOrder: { type: Number, default: 0 },
  shortCode: { type: String, required: true },
  questionText: { type: String, required: true },
  data: { },

  note: { type: String },
  isBlocked: { type: Boolean, default: false },
  isDeleted: { type: Boolean, default: false }
}, options: {
  timestamps: true, id: false
});

QuestionSchema.index(
   fields: { isDeleted: 1, shortCode: 1 },
   options: { unique: true, partialFilterExpression: { isDeleted: { $eq: false } } }
)
```

Figure 62. Mongo model definition for questions in the custom website



```
const UserSchema = new mongoose.Schema( definition: {
  fname: { type: String, required: true },
  lname: { type: String, required: true },
  phone: { type: String, required: true },
  email: { type: String, required: true },
  note: { type: String },
  passwordHash: { type: String, select: false }, // won't select unless explicitly told so
  isAdmin: {
    type: Boolean,
    default: false,
  },
  role: { type: String, enum: ROLES_ALL },
  isBlocked: {
    type: Boolean,
    default: false,
  },
  isDeleted: {
    type: Boolean,
    default: false,
  },
  lastLogin: {
    type: Date,
    /*default: Date.now*/
  }
}, options: {
  timestamps: true, id: false
});

UserSchema.index(
   fields: { email: 1, isDeleted: 1 },
   options: { unique: true, partialFilterExpression: { isDeleted: { $eq: false } } }
);
UserSchema.index(
   fields: { phone: 1, isDeleted: 1 },
   options: { unique: true, partialFilterExpression: { isDeleted: { $eq: false } } }
)
```

Figure 63. Mongo model definition for users in the custom website

APPENDIX E

CURATED VIDEO CLIPS FOR NEW YT360-EYE-TRACKING DATASET

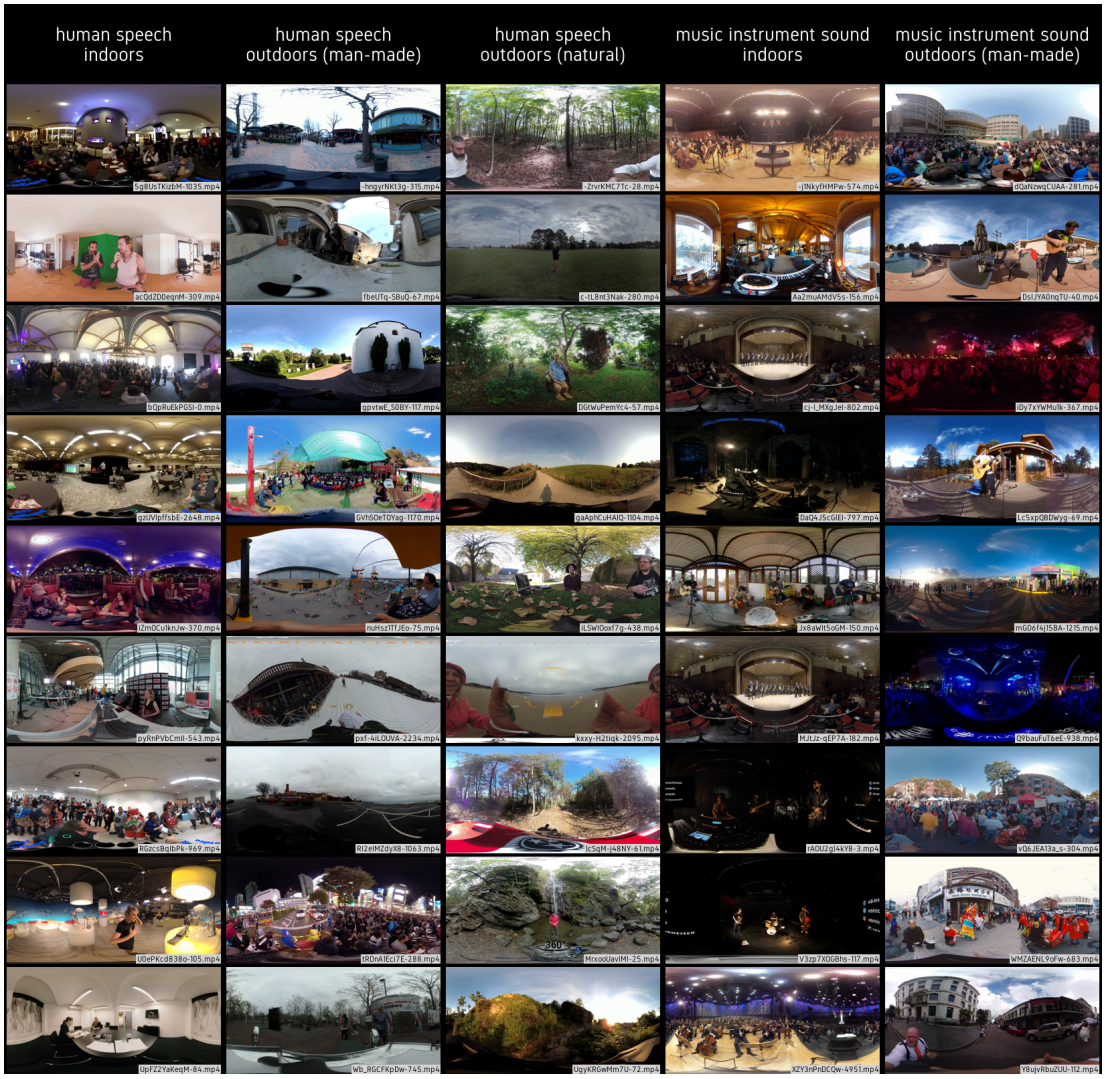


Figure 64. The first half of the video clips in the created dataset

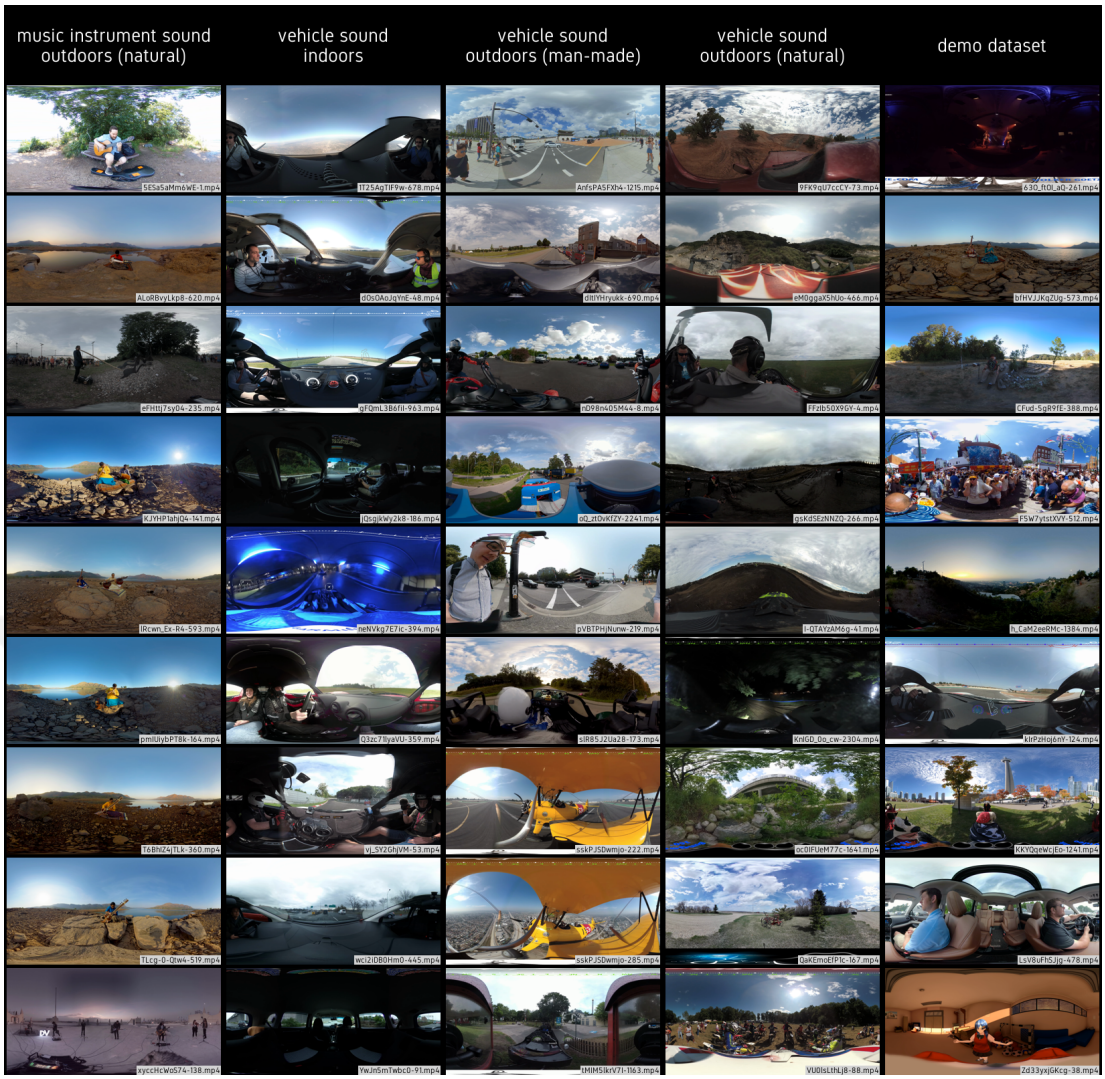


Figure 65. The second half of the video clips in the created dataset

APPENDIX F

QUESTIONNAIRE FORM (ENGLISH)

Demographic Information Form

Your age:

Your gender:

Your height:

Your weight:

Virtual Reality Experience

Have you ever had a virtual reality experience before this experiment? (Yes/No)

If you have had a virtual reality experience before:

- When was the first and last time you used this technology:
- How long do you estimate your total experience time to be:
- What has been the frequency of your use:

Video Game Experience

Have you ever played any video games (computer/play station, etc.) before? (Yes/No)

If you have played video games before:

- When was the first and last time you played a video game:
- What is your estimated total duration of playing video games:
- How frequently do you usually play video games:

APPENDIX G

QUESTIONNAIRE FORM (TURKISH)

Demografik Bilgi Formu

Yaşınız:

Cinsiyetiniz:

Boyunuz:

Kilonuz:

Sanal Gerçeklik Deneyimi

Bu deneyden önce hiç sanal gerçeklik deneyiminiz olmuş muydu? (Evet/ Hayır)

Eğer daha önceden bir sanal gerçeklik deneyiminiz olduysa:

- Bu teknolojiyi ilk ve son olarak ne zaman kullanmıştınız:
- Tahmini toplam deneyim süreniz ne kadardır:
- Kullanım sıklığınız ne olmuştur:

Video Oyunu Deneyimi

Daha önceden herhangi bir video oyunu (bilgisayar/play station vs.) oynadınız mı? (Evet/ Hayır)

Eğer daha önceden video oyunu oynadıysanız:

- İlk ve son olarak ne zaman video oyunu oynamıştınız:
- Tahmini toplam video oyunu oynama süreniz ne kadardır:
- Video oyunu oynama sıklığınız ne olmuştur:

APPENDIX H

ETHICS COMMITTEE APPROVAL FOR 360° VIDEO EXPERIMENT

Evrak Tarih ve Sayısı: 10.04.2022-61405

T.C.
BOĞAZİÇİ ÜNİVERSİTESİ
SOSYAL VE BEŞERİ BİLİMLER YÜKSEK LİSANS VE DOKTORA TEZLERİ ETİK İNCELEME KOMİSYONU
TOPLANTI KARAR TUTANAĞI

Toplantı Sayısı	:	30
Toplantı Tarihi	:	07.04.2022
Toplantı Saati	:	10:00
Toplantı Yeri	:	Zoom Sanal Toplantı
Bulunanlar	:	Prof. Dr. Ebru Kaya, Dr. Öğr. Üyesi Yasemin Sohtorik İlkmen
Bulunmayanlar	:	

Halit Özsoy
Psikoloji

Sayın Araştırmacı,
"Sanal Gerçeklikte 360 Derece Panoramik Videolarda Görsel-İşitsel Belirginlik Haritaları" başlıklı projeniz ile ilgili olarak yaptığınız SBB-EAK 2022/28 sayılı başvuru komisyonumuz tarafından 7 Nisan 2022 tarihli toplantıda incelenmiş ve uygun bulunmuştur.

Bu karar tüm üyelerin toplantıya çevrimiçi olarak katılımı ve oybirliği ile alınmıştır. COVID-19 önlemleri kapsamında kurul üyelerinden ıslak imza alınamadığı için bu onay mektubu üye ve raportör olarak Yasemin Sohtorik İlkmen tarafından bütün üyeler adına e-imzalanmıştır.

Saygılarımızla, bilgilerinizi rica ederiz.

Dr. Öğr. Üyesi Yasemin
SOHTORİK İLKMEN
ÜYE

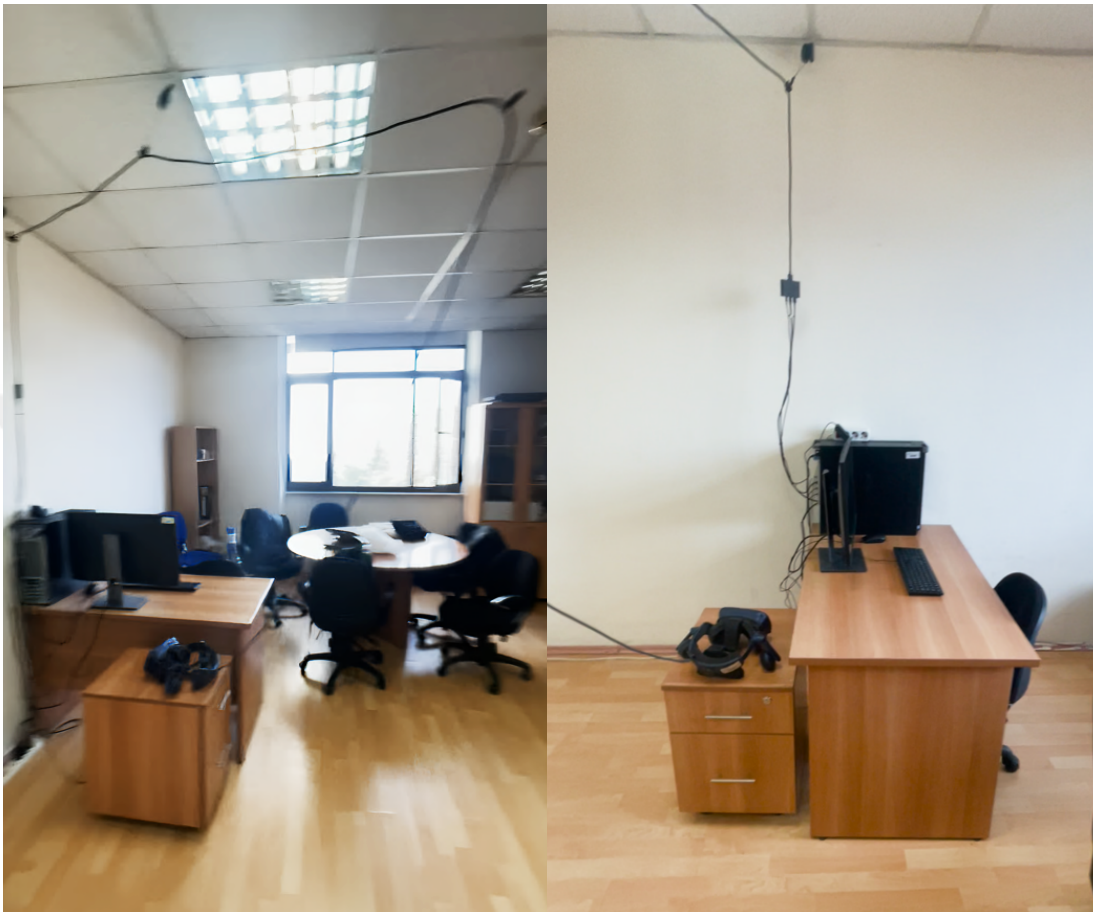
e-imzalıdır
Dr. Öğr. ÜyesiYasemin Sohtorik
İlkmen
Öğretim Üyesi
Raportör

SOBETİK 30 07.04.2022

Bu belge, güvenli elektronik imza ile imzalanmıştır.

APPENDIX I

EXPERIMENT SETUP AND ENVIRONMENT



Experiment environment and setup: Cable extension mechanism is shown

APPENDIX J

EXPERIMENTER PROMPT FOR 360° VIDEO EXPERIMENT (ENGLISH)

Hello, thank you for taking part in our research study. You will watch a series of 360-degree videos in a standing position. Your eye movements will be tracked as you watch the videos. A 360-degree video is a type of video that allows you to experience a scene from all angles. So; right, left, right, front, back, up, down, every direction will be video. Watching 360-degree videos can sometimes cause motion sickness. It can have effects such as mild dizziness, loss of balance, nausea. Especially if there is movement in the video or if the camera is not stabilized very well, it happens more, be aware. If you feel uncomfortable, tell us and we will stop the experiment. There are no jump-scare content in the videos. During the experiment you will watch 81 videos, each one lasting 30 seconds. There will be a 10-second break between each two videos, during which you will see a black screen. The videos will be divided into three sessions. Each will last 18 minutes. There will be a 3-5 minute break between the two sessions. We will also do a calibration process before each session. We will check whether we have worn the headgear properly and see if your eye data is coming. Now, I will explain the video content. Some of the videos will be with sound, some without sound, some in color, some in grayscale. No special instructions will be given while watching the videos. You can turn freely while watching the videos, but try not to walk too far. If you walk too far there will be blue boundary lines, if you see them, walk back a few steps and you will be fine. Walking does not have any special effect, but it is usually easier to turn via steps when turning around. If you have any questions, please ask. If you do not have any more questions, we can start whenever you want.

APPENDIX K

EXPERIMENTER PROMPT FOR 360° VIDEO EXPERIMENT (TURKISH)

Merhaba, araştırma çalışmamıza katıldığınız için teşekkür ederiz. Ayakta durarak bir dizi 360 derece video izleyeceksiniz. Videoları izlerken göz hareketleriniz takip edilecek. 360 derece video, bir sahneyi tüm açılardan deneyimlemenizi sağlayan bir video türüdür. Yani; sağ, sol, ön, arka, yukarı, aşağı, her yönünüz video olacak. 360 derece videoları izlerken bazen taşıt tutması olabiliyor. Hafif baş dönmesi, denge kaybı, mide bulantısı gibi etkileri olabilir. Özellikle video içerisinde hareket varsa veya kamera çok iyi sabitlenmediyse daha çok oluyor, haberiniz olsun. Eğer rahatsız hissederseniz, söylersiniz, deneyi durdururuz. Bir anda beliren veya korku unsuru içeren video yok. Deney sırasında 81 video izleyeceksiniz, her biri 30 saniye sürecek. Her iki video arasında 10 saniyelik bir ara olacak, o sırada siyah ekran görüyor olacaksınız. Videolar üç seansa ayrılacak. Her biri 18 dakika sürecek. İki seans arasında 3-5 dakikalık bir ara verilecektir. Her seans öncesinde bir de kalibrasyon işlemi yapacağız. Hem başlığı düzgün takmış mıyız diye kontrol edeceğiz, hem de göz verileriniz geliyor mu diye bakacağız. Biraz da video içeriğinden bahsedeyim. Videoların bazıları sesli, bazıları sessiz, bazıları renkli, bazıları siyah-beyaz olacak. Videoları izlerken herhangi özel talimat verilmeyecektir. Videoları izlerken serbestçe dönebilirsiniz, ancak çok uzağa yürümemeye çalışın. Eğer çok uzağa yürürseniz mavi sınır çizgileri çıkıyor, onları görünce birkaç adım geri gelirseniz sorun olmaz. Yürümenin özel bir etkisi yok, fakat arkaya dönerken falan yürüyerek dönmek genelde daha kolay oluyor. Sorularınız varsa alabilirim. Başka sorunuz yoksa istediğinizde başlayabiliriz.

APPENDIX L

VIDEO THUMBNAILS WITH HEATMAP OVERLAY OF FIXATION MAPS



Figure 66. Heatmap overlays from human speech, indoors category

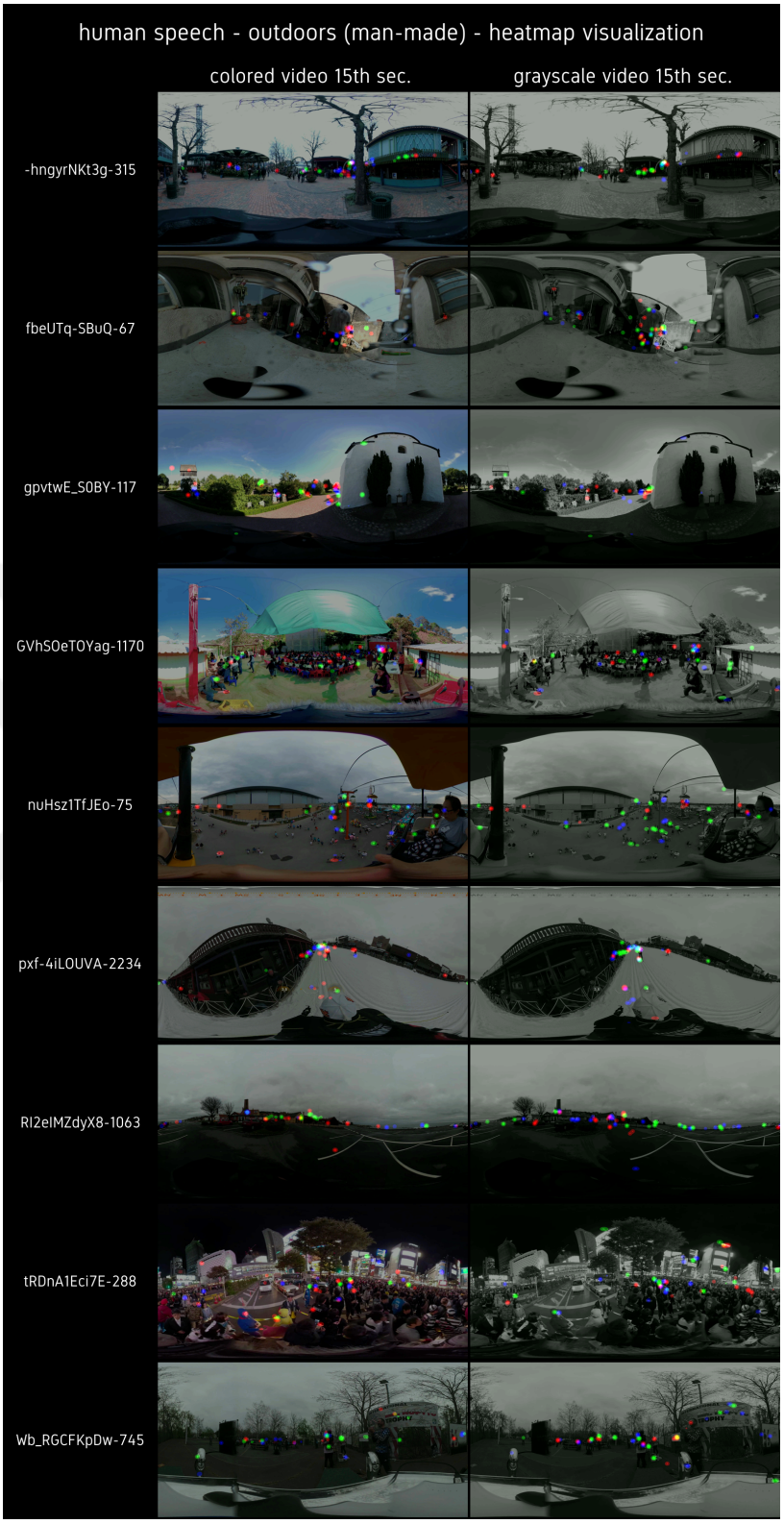


Figure 67. Heatmap overlays from human speech, outdoors (human-made) category

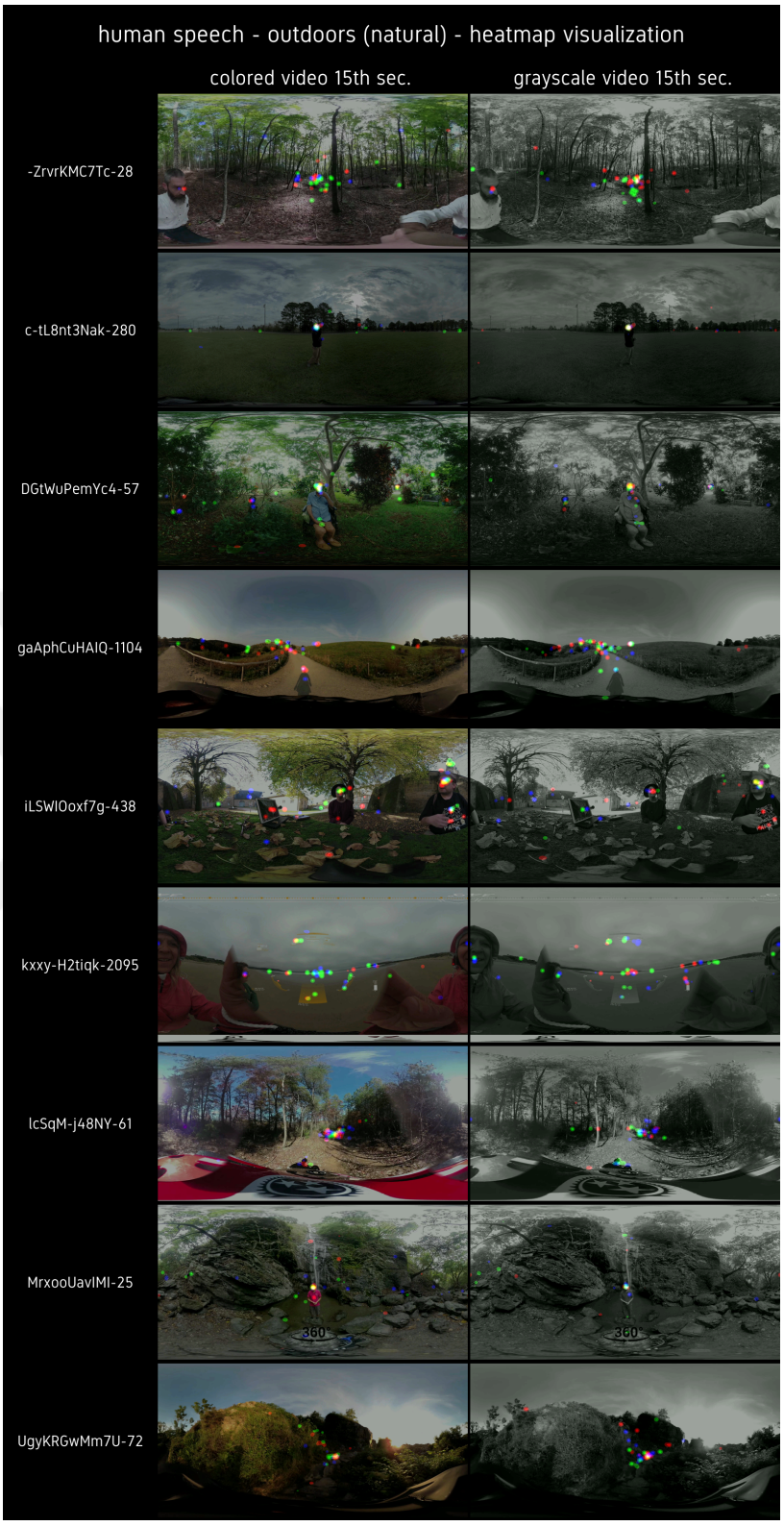


Figure 68. Heatmap overlays from human speech, outdoors (natural) category

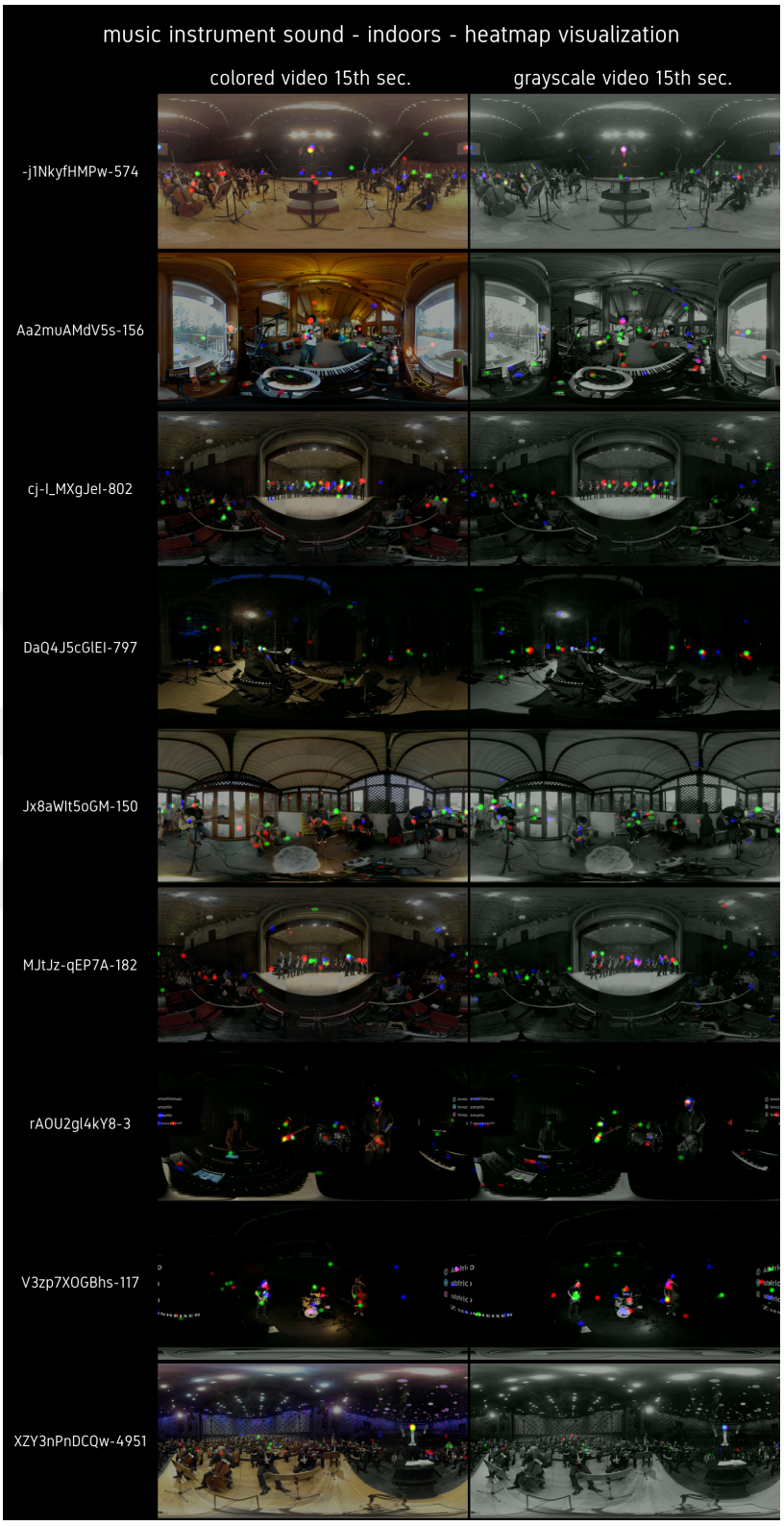


Figure 69. Heatmap overlays from music instrument, indoors category

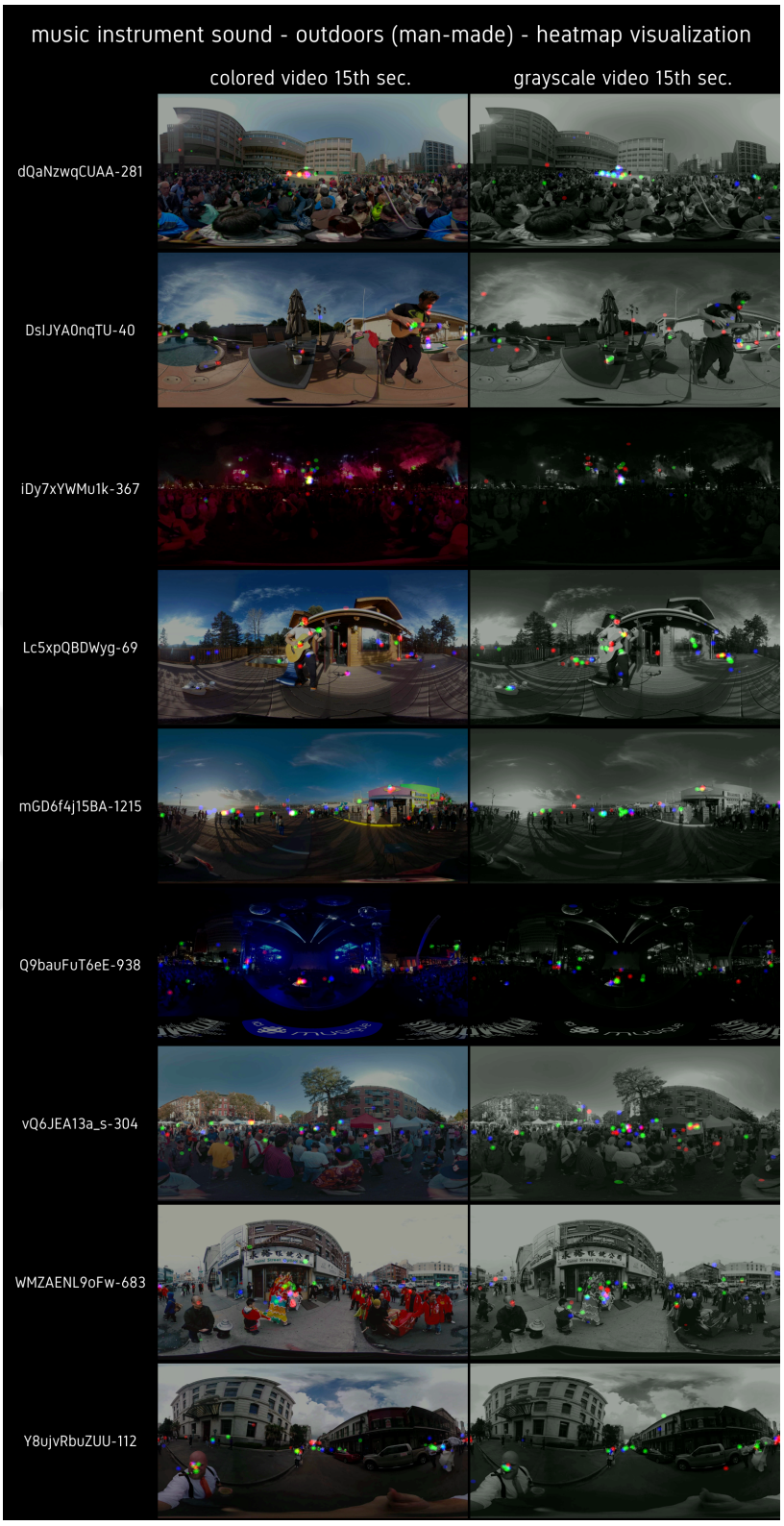


Figure 70. Heatmap overlays from music instrument, outdoors (human-made)



Figure 71. Heatmap overlays from music instrument, outdoors (natural)



Figure 72. Heatmap overlays from vehicle sound, indoors category



Figure 73. Heatmap overlays from vehicle sound, outdoors (human-made)

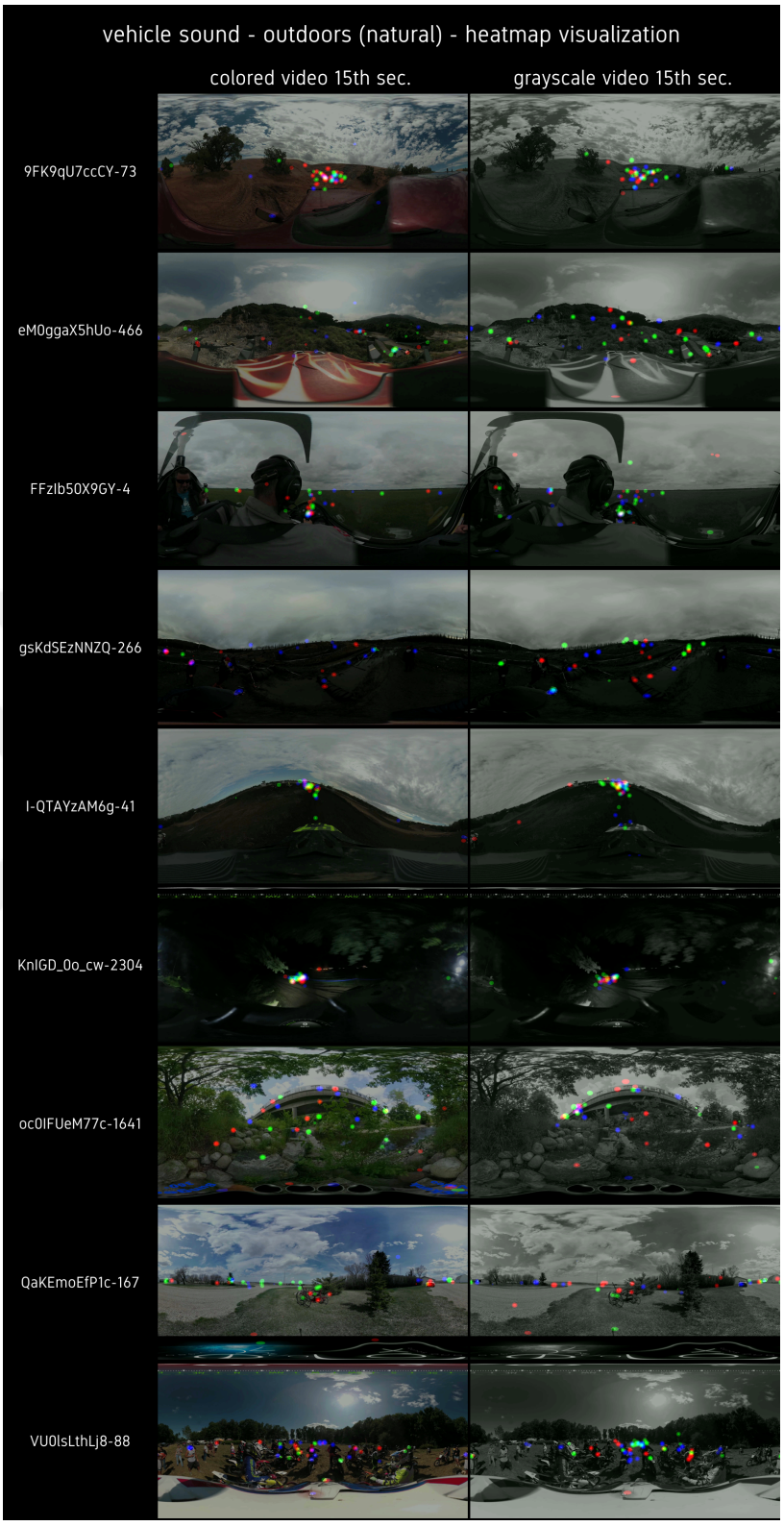


Figure 74. Heatmap overlays from vehicle sound, outdoors (natural) category

APPENDIX M

ADDITIONAL PLOTS OF CONSISTENCY SCORES

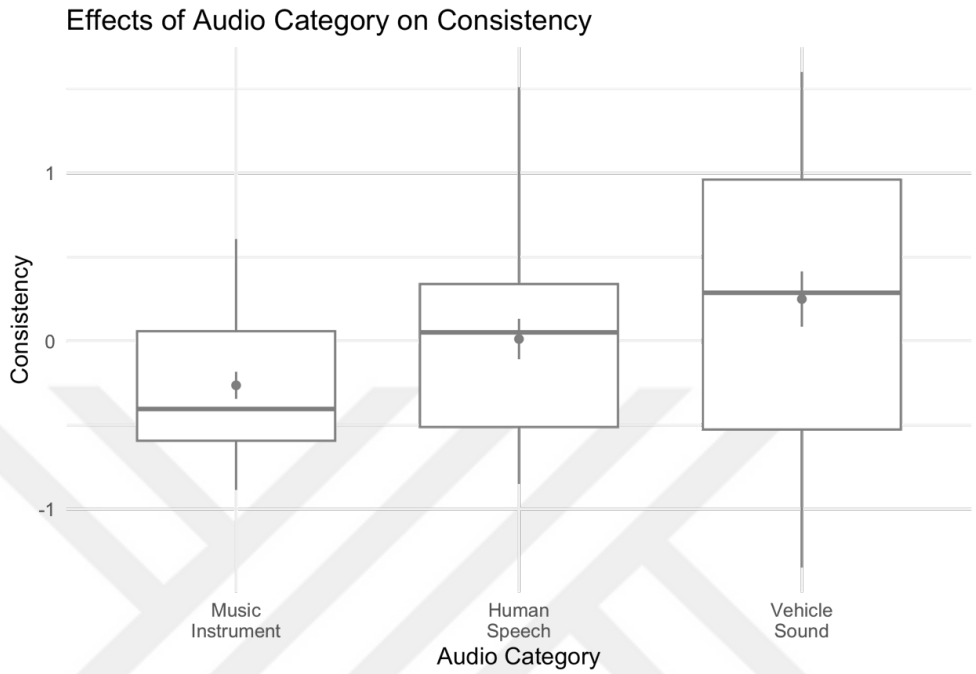


Figure 75. Boxplot of consistency scores across audio categories

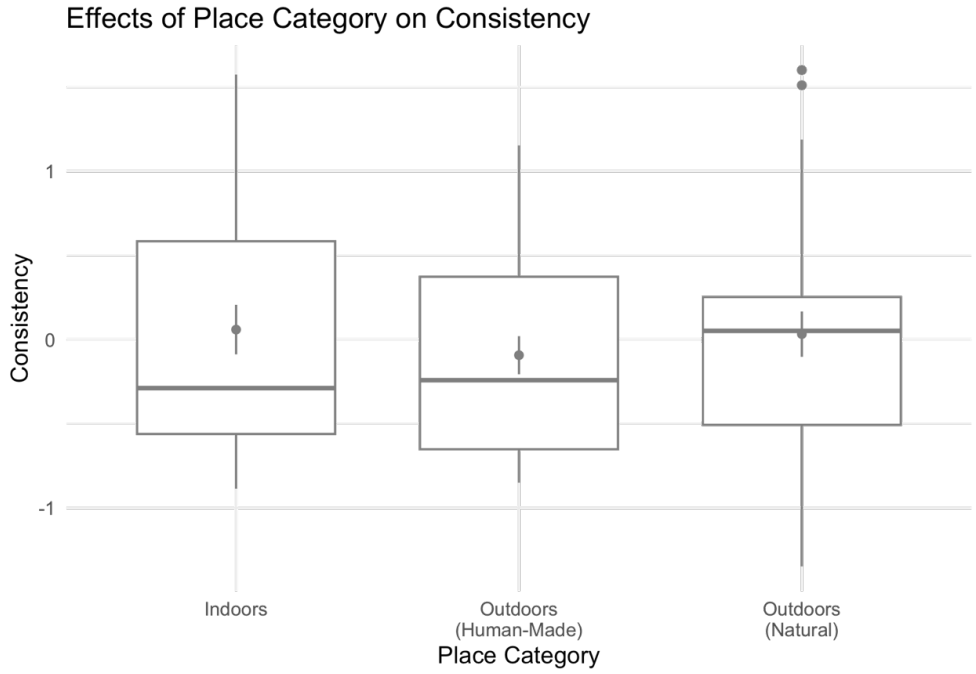


Figure 76. Boxplot of consistency scores across place categories

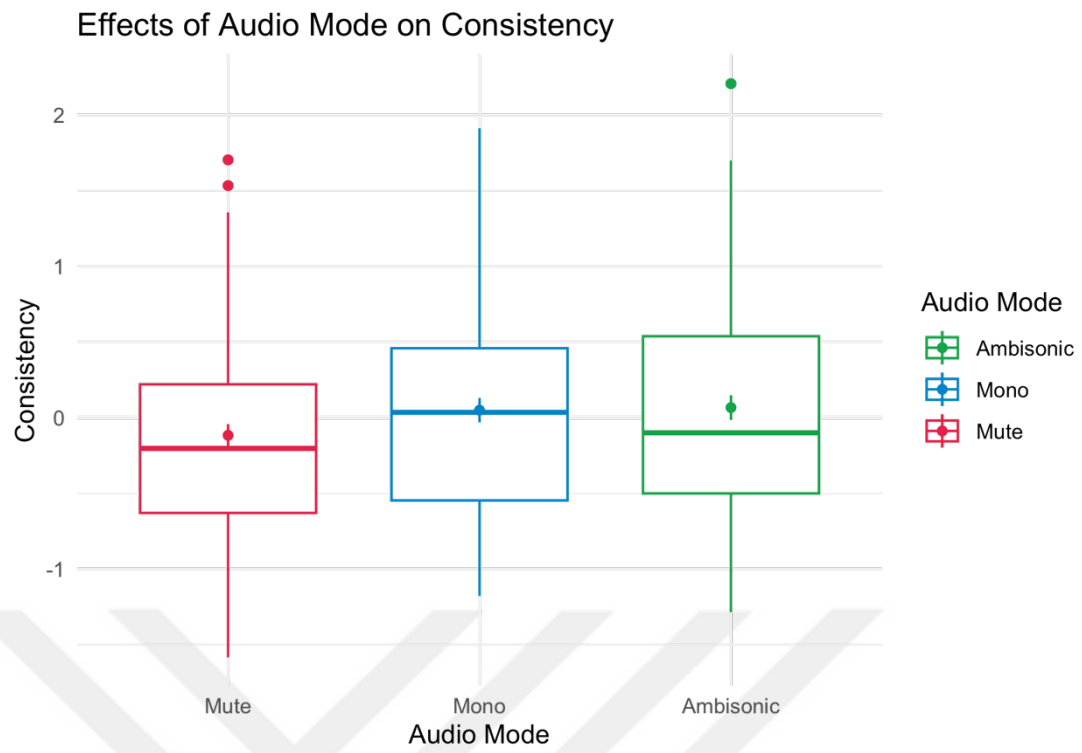


Figure 77. Boxplot of consistency scores across audio modes

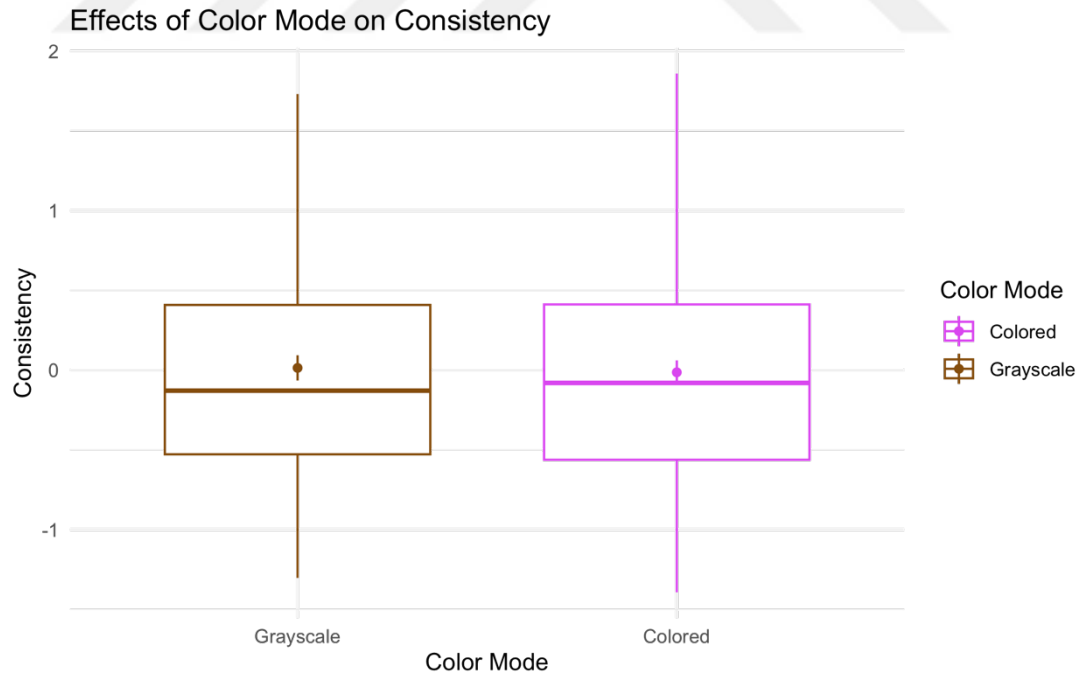


Figure 78. Boxplot of consistency scores across color modes

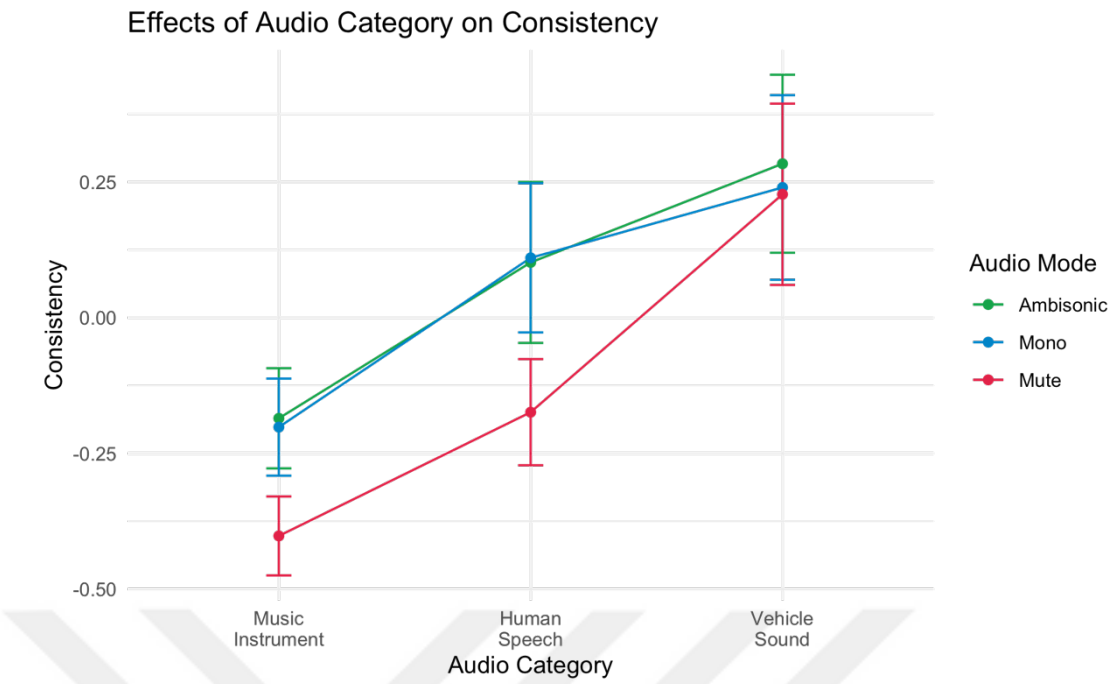


Figure 79. Consistency scores line plot of audio categories per audio mode

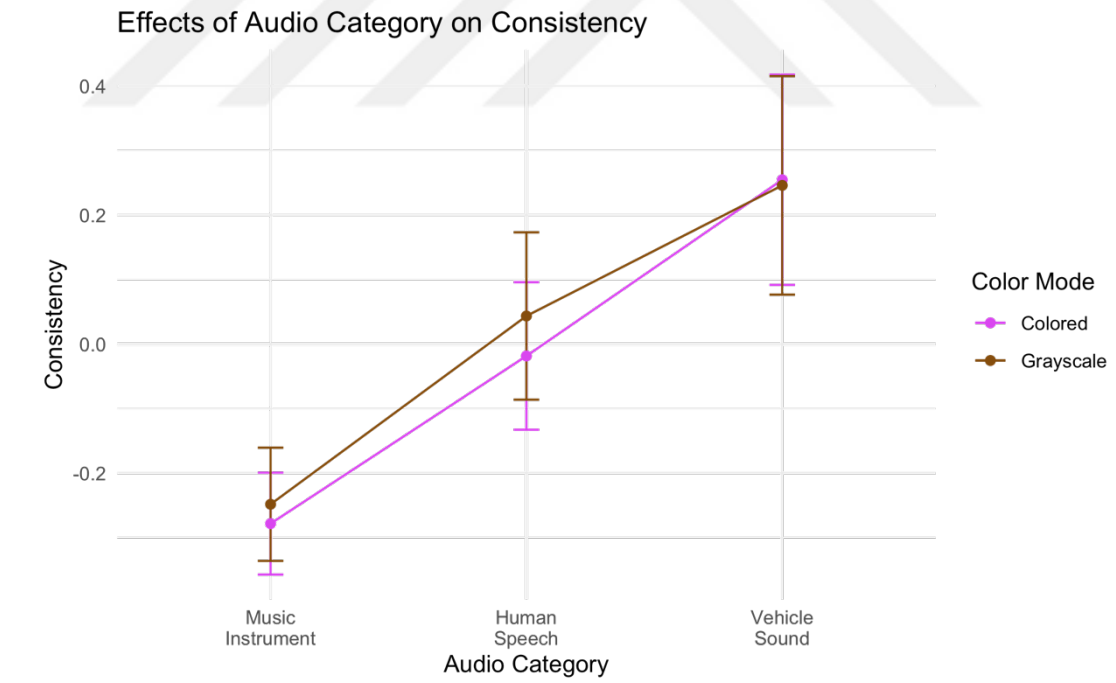


Figure 80. Consistency scores line plot of audio categories per color mode

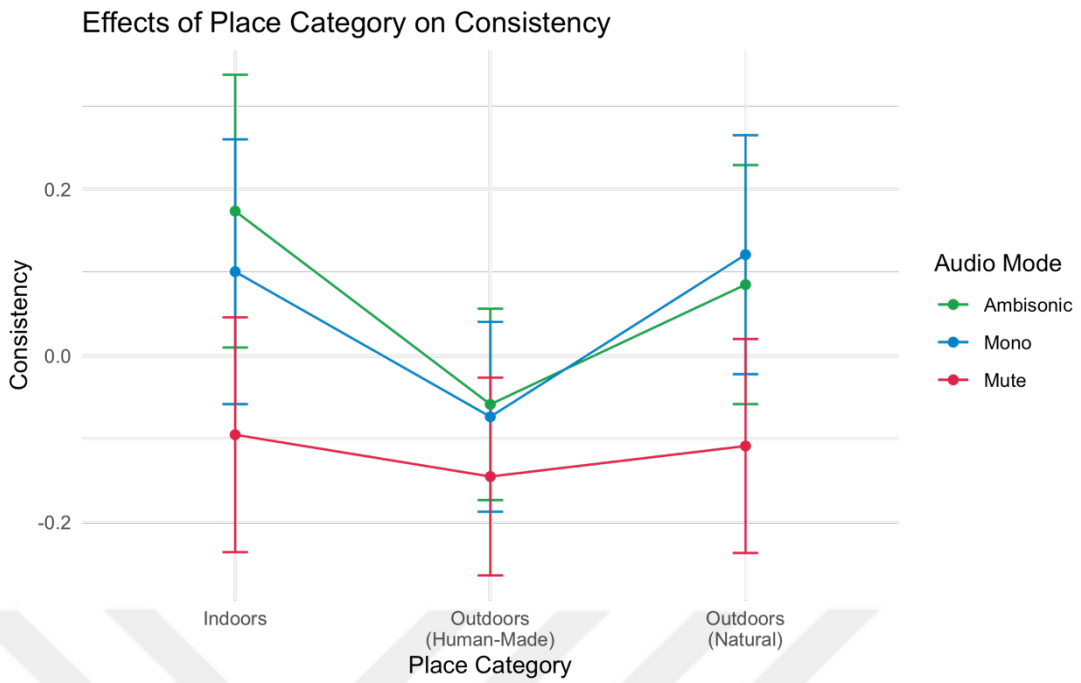


Figure 81. Consistency scores line plot of place categories per audio mode

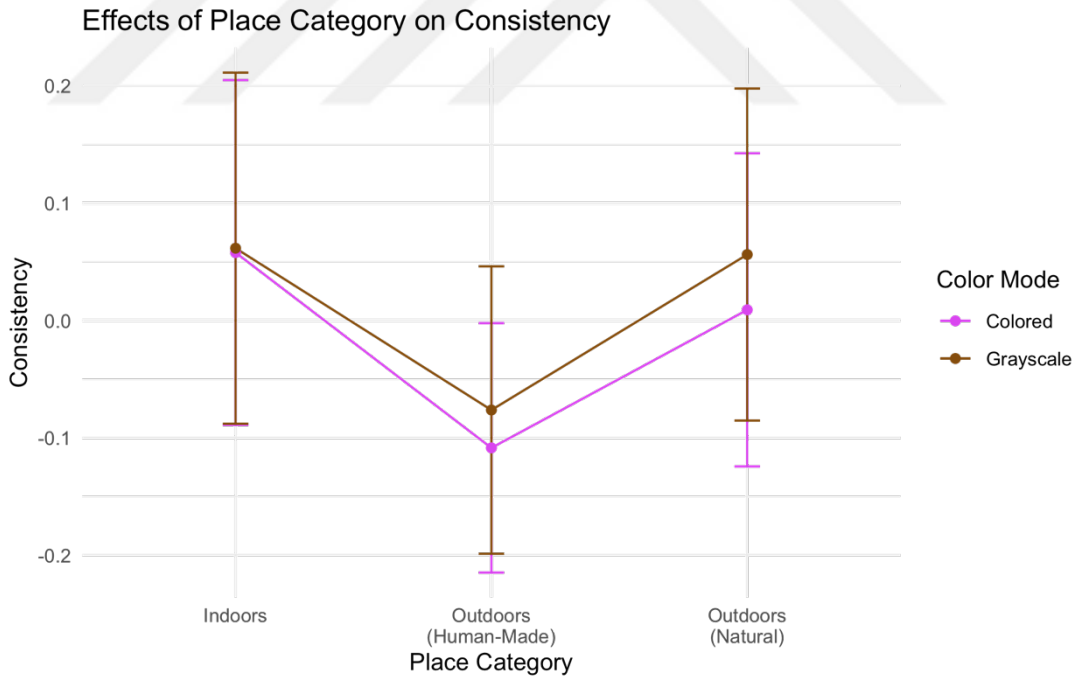


Figure 82. Consistency scores line plot of place categories per color mode

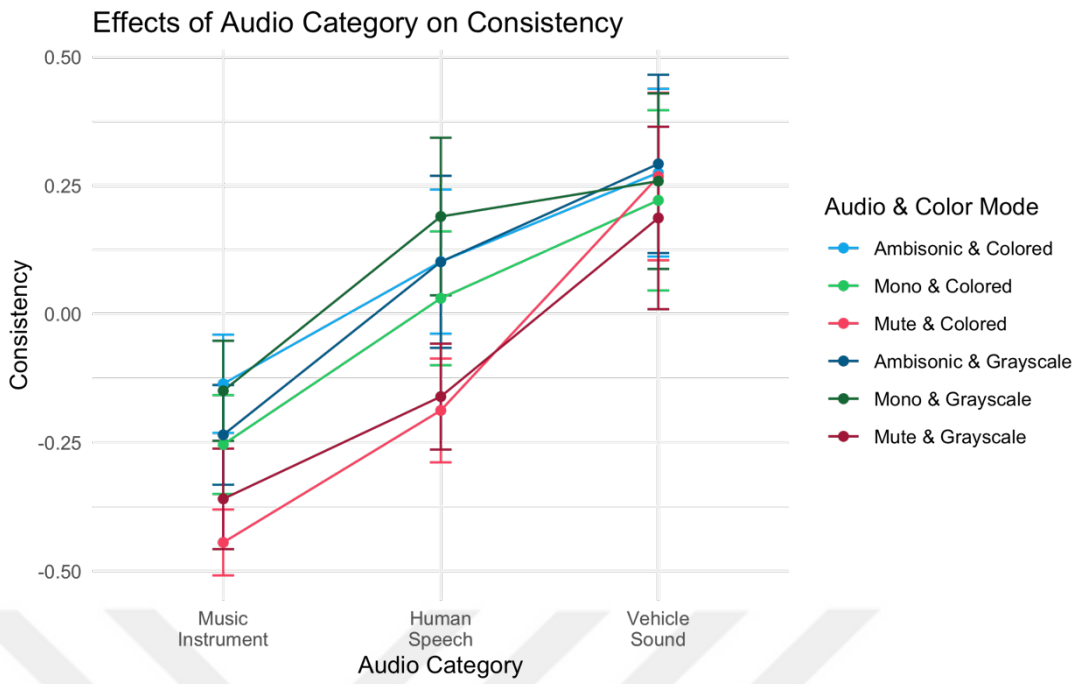


Figure 83. Consistency scores of audio categories per audio and color mode

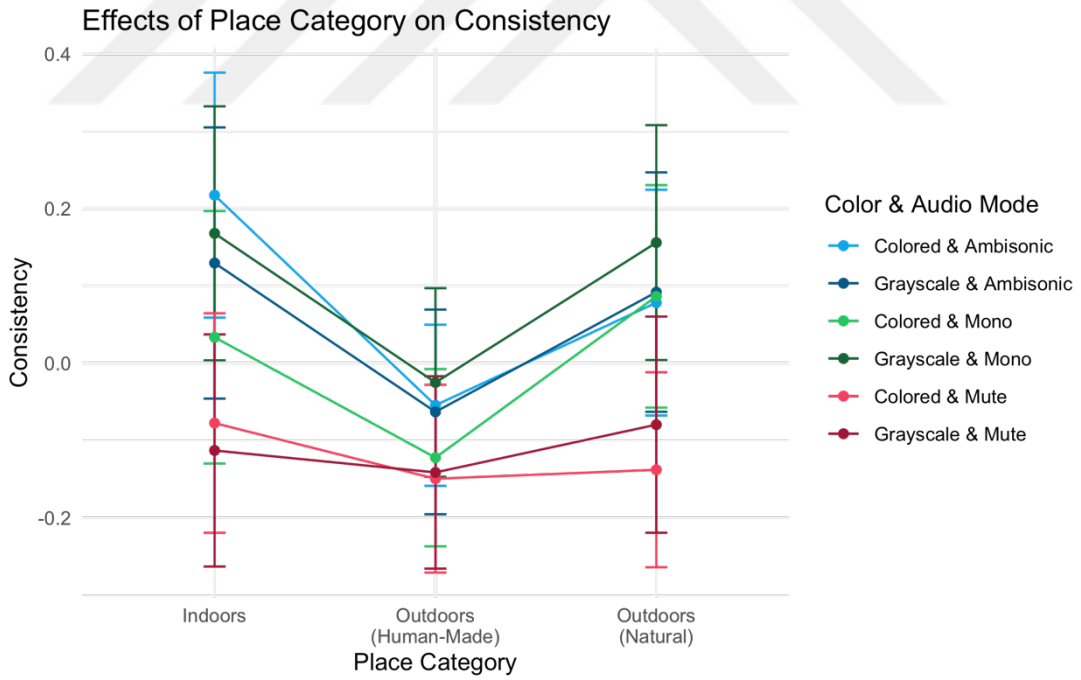


Figure 84. Consistency scores of place categories per color and audio mode

APPENDIX N

ETHICS COMMITTEE APPROVAL FOR PSYCHOPHYSICS EXPERIMENT

Evrak Tarih ve Sayısı: 03.07.2023-134080

T.C.
BOĞAZİÇİ ÜNİVERSİTESİ
SOSYAL VE BEŞERİ BİLİMLER YÜKSEK LİSANS VE DOKTORA TEZLERİ ETİK İNCELEME KOMİSYONU
TOPLANTI KARAR TUTANAĞI

Toplantı Sayısı	:	46
Toplantı Tarihi	:	14.06.2023
Toplantı Saati	:	14:00
Toplantı Yeri	:	Zoom Sanal Toplantı
Bulunanlar	:	Prof. Dr. Feyza Çorapçı, Dr. Öğr. Üyesi Ayşegül Metindoğan, Doç. Dr. Senem Yıldız
Bulunmayanlar	:	Doç.Dr. Arhan S. Ertan, Dr. Öğr. Üyesi Yasemin Sohtorik İlkmen

Halit Özsoy
Bilişsel Bilim

Sayın Araştırmacı,

"Çok Yönlü Panoramik Sahnelerde Görsel-İşitsel Belirginlik (Audio-Visual Saliency in Omnidirectional Panoramic Scenes)" başlıklı projeniz ile ilgili olarak yaptığınız SBB-EAK 2023/54 sayılı başvuru komisyonumuz tarafından 14 Haziran 2023 tarihli toplantıda incelenmiş ve uygun bulunmuştur.

Bu karar üyelerin toplantıya çevrimiçi olarak katılımı ve oy birliği ile alınmıştır. Onay mektubu üye ve raportör olarak Yasemin Sohtorik İlkmen tarafından toplantıya katılan bütün üyeler adına e-imzalanmıştır.

Saygılarımızla, bilgilerinizi rica ederiz.

Dr. Öğr. Üyesi Yasemin
SOHTORİK İLKMEN
Öğretim Üyesi

e-imzalıdır
Dr. Öğr. ÜyesiYasemin Sohtorik
İlkmen
Öğretim Üyesi
Raportör

SOBETİK 46 14.06.2023

Bu belge, güvenli elektronik imza ile imzalanmıştır.

REFERENCES

Agtzidis, I., & Dorr, M. (2019). Getting (more) real: Bringing eye movement classification to HMD experiments with equirectangular stimuli. *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*. doi:10.1145/3314111.3319829

Bertet, S., Daniel, J., & Moreau, S. (2006). 3D sound field recording with higher order ambisonics - objective measurements and validation of spherical microphone. *Journal of the Audio Engineering Society*, *6857*. Retrieved from https://aes2.org/publications/elibrary-page/?id13661

Blais, J. (2011). Discrete spherical harmonic transforms of nearly equidistributed global data. *Journal of Geodetic Science*, *1*(3), 251–258. doi:10.2478/v10156-011-0003-1

Blignaut, P. (2009). Fixation identification: The optimum threshold for a dispersion algorithm. *Attention, Perception, & Psychophysics*, *71*(4), 881–895. doi:10.3758/app.71.4.881

Borji, A., Cheng, M.-M., Hou, Q., Jiang, H., & Li, J. (2019). Salient object detection: A survey. *Computational Visual Media*, *5*(2), 117–150. doi:10.1007/s41095-019-0149-9

Bradski, G. (2000). The OpenCV library. *Dr. Dobb's Journal of Software Tools*.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*(4), 433–436. doi:10.1163/156856897x00357

Bruce, N., & Tsotsos, J. (2005). Saliency based on information maximization. *Advances in Neural Information Processing Systems*, *18*. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2005/file/0738069b244a1c43c83112b735140a16-Paper.pdf

Bundesen, C. (1990). A theory of visual attention. *Psychological Review*, *97*(4), 523–547. doi:10.1037/0033-295x.97.4.523

Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., & Durand, F. (2019). What do different evaluation metrics tell us about saliency models? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(3), 740–757. doi:10.1109/tpami.2018.2815601

Bylinskii, Z., Recasens, A., Borji, A., Oliva, A., Torralba, A., & Durand, F. (2016). Where should saliency models look next? *Computer Vision – ECCV 2016*, 809–824. doi:10.1007/978-3-319-46454-1_49

Carter, B. T., & Luke, S. G. (2020). Best practices in eye tracking research. *International Journal of Psychophysiology*, *155*, 49–62. doi:10.1016/j.ijpsycho.2020.05.010

Celikcan, U., Askin, M. B., Albayrak, D., & Capin, T. K. (2020). Deep into visual saliency for immersive VR environments rendered in real-time. *Computers & Graphics*, *88*, 70–82. doi:10.1016/j.cag.2020.03.006

Cerf, M., Frady, E. P., & Koch, C. (2009). Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of Vision*, *9*(12), 10–10. doi:10.1167/9.12.10

Chao, F.-Y., Ozcinar, C., Zhang, L., Hamidouche, W., Deforges, O., & Smolic, A. (2020). Towards audio-visual saliency prediction for omnidirectional video with spatial audio. *2020 IEEE International Conference on Visual Communications and Image Processing (VCIP)*. doi:10.1109/vcip49819.2020.9301766

Chen, H., Xie, W., Vedaldi, A., & Zisserman, A. (2020). Vggsound: A large-scale audio-visual dataset. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. doi:10.1109/icassp40776.2020.9053174

Chiariotti, F. (2021). A survey on 360-degree video: Coding, quality of experience and streaming. *Computer Communications*, *177*, 133–155. doi:10.1016/j.comcom.2021.06.029

Ciortan, I. M., Dinet, E., & Tremeau, A. (2014). Influence of color on visual saliency in short videos. *2014 IEEE International Conference on Image Processing (ICIP)*. doi:10.1109/icip.2014.7025231

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, *3*(3), 201–215. doi:10.1038/nrn755

Coutrot, A., & Guyader, N. (2014). How saliency, faces, and sound influence gaze in dynamic social scenes. *Journal of Vision*, *14*(8), 5–5. doi:10.1167/14.8.5

Coutrot, A., Guyader, N., Ionescu, G., & Caplier, A. (2012). Influence of soundtrack on eye movements during video exploration. *Journal of Eye Movement Research*, *5*(4). doi:10.16910/jemr.5.4.2

Dahou, Y., Tliba, M., McGuinness, K., & O'Connor, N. (2021). ATSal: An attention based architecture for saliency prediction in 360° videos. *Pattern Recognition. ICPR International Workshops and Challenges*, 305–320. doi:10.1007/978-3-030-68796-0_22

David, E. J., Gutiérrez, J., Coutrot, A., Da Silva, M. P., & Le Callet, P. (2018). A dataset of head and eye movements for 360° videos. *Proceedings of the 9th ACM Multimedia Systems Conference*. doi:10.1145/3204949.3208139

David, E., Gutiérrez, J., Vo, M. L.-H., Coutrot, A., Da Silva, M. P., & Le Callet, P. (2023). The salient360! toolbox: Processing, visualising and comparing gaze data in 3D. *2023 Symposium on Eye Tracking Research and Applications*. doi:10.1145/3588015.3588406

de Gelder, B., Kätsyri, J., & de Borst, A. W. (2018). Virtual reality and the new Psychophysics. *British Journal of Psychology*, *109*(3), 421–426. doi:10.1111/bjop.12308

de Santana Correia, A., & Colombini, E. L. (2022). Attention, please! A survey of neural attention models in deep learning. *Artificial Intelligence Review*, *55*(8), 6037–6124. doi:10.1007/s10462-022-10148-x

Ding, N., & Simon, J. Z. (2012). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences*, *109*(29), 11854–11859. doi:10.1073/pnas.1205381109

Dinh, T. (2023) Decibel x: dB sound level meter [Mobile application software] Retrieved from http://itunes.apple.com/

Djilali, Y. A. D., McGuinness, K., & O'Connor, N. E. (2021). Simple baselines can fool 360° saliency metrics. *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*. doi:10.1109/iccvw54120.2021.00418

Elwardy, M., Zepernick, H.-J., & Sundstedt, V. (2019). Annotated 360-degree image and video databases: A comprehensive survey. *2019 13th International Conference on Signal Processing and Communication Systems (ICSPCS)*. doi:10.1109/icspcs47537.2019.9008715

Fairchild, M. D. (2013). *Color appearance models* (3rd ed.). Wiley.

Fan, S., Shen, Z., Jiang, M., Koenig, B. L., Xu, J., Kankanhalli, M. S., & Zhao, Q. (2018). Emotional attention: A study of image sentiment and visual attention. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. doi:10.1109/cvpr.2018.00785

Farell, B., & Pelli, D. G. (1998). Psychophysical methods, or how to measure a threshold, and why. *Vision Research: A Practical Guide to Laboratory Methods*, 129–136. doi:10.1093/acprof:oso/9780198523192.003.0005

Fastl, H., & Zwicker, E. (2007). *Psychoacoustics* (3rd ed.). New York, NY: Springer.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. doi:10.3758/brm.41.4.1149

Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., … Ritter, M. (2017). Audio set: An ontology and human-labeled dataset for audio events. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. doi:10.1109/icassp.2017.7952261

Gerzon, M. A. (1973). Periphony: With-height sound reproduction. *Journal of the Audio Engineering Society*, *21*(1), 2–10. Retrieved from https://aes2.org/publications/elibrary-page/?id=2012

GoLuckyRyan. (2017, August 14). Spherical harmonics and platonic solids [Web log post]. Retrieved from https://nukephysik101.wordpress.com/2017/08/14/spherical-harmonics-and-platonic-solids/

Gutierrez, J., David, E. J., Coutrot, A., Da Silva, M. P., & Callet, P. L. (2018). Introducing UN Salient360! benchmark: A platform for evaluating visual attention models for 360° contents. *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*. doi:10.1109/qomex.2018.8463369

Haskins, A. J., Mentch, J., Botch, T. L., & Robertson, C. E. (2020). Active vision in immersive, 360° real-world environments. *Scientific Reports*, *10*(1). doi:10.1038/s41598-020-71125-4

Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., Moore, R. C., … Wilson, K. (2017). CNN architectures for large-scale audio classification. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. doi:10.1109/icassp.2017.7952132

Herwig, A., & Schneider, W. X. (2014). Predicting object features across saccades: Evidence from object recognition and visual search. *Journal of Experimental Psychology: General*, *143*(5), 1903–1922. doi:10.1037/a0036781

Hiebert, G., Gordon, R., Peacock, D., & Vogelsang, C. (2005). OpenAL 1.1 specification and reference. *Creative Labs*. Retrieved from https://openal.org/documentation/OpenAL_Programmers_Guide.pdf

Hu, Z., Bulling, A., Li, S., & Wang, G. (2021). FixationNet: Forecasting eye fixations in task-oriented virtual environments. *IEEE Transactions on Visualization and Computer Graphics*, *27*(5), 2681–2690. doi:10.1109/tvcg.2021.3067779

Hu, Z., Li, S., Zhang, C., Yi, K., Wang, G., & Manocha, D. (2020). DGaze: CNN-based gaze prediction in dynamic scenes. *IEEE Transactions on Visualization and Computer Graphics*, *26*(5), 1902–1911. doi:10.1109/tvcg.2020.2973473

IBM Corp. (2020). *IBM SPSS statistics for macintosh* (Version 27.0) [Computer software]. Armonk, NY: IBM Corp

IoSR-Surrey. (2023). *IoSR Matlab Toolbox* [Computer software]. Retrieved from https://github.com/IoSR-Surrey/MatlabToolbox

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, *40*(10–12), 1489–1506. doi:10.1016/s0042-6989(99)00163-7

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience*, *2*(3), 194–203. doi:10.1038/35058500

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(11), 1254–1259. doi:10.1109/34.730558

Jiang, M., Chen, S., Yang, J., & Zhao, Q. (2020). Fantastic answers and where to find them: Immersive question-directed visual attention. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. doi:10.1109/cvpr42600.2020.00305

John, B., Raiturkar, P., Le Meur, O., & Jain, E. (2018). A benchmark of four methods for generating 360° saliency maps from eye tracking data. *2018 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*. doi:10.1109/aivr.2018.00028

JSON Org. (1999). *JSON*. Retrieved 12 July 2022, from https://www.json.org/json-en.html

Judd, T., Durand, F., & Torralba, A. (2012). *A benchmark of computational models of saliency to predict human fixations*. Retrieved from https://dspace.mit.edu/handle/1721.1/68590

Judd, T., Ehinger, K., Durand, F., & Torralba, A. (2009). Learning to predict where humans look. *2009 IEEE 12th International Conference on Computer Vision*. doi:10.1109/iccv.2009.5459462

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika*, *39*(1), 31–36. doi:10.1007/bf02291575

Kaldas, A. (2022). Attention: A descriptive taxonomy. *History and Philosophy of the Life Sciences*, *44*(4). doi:10.1007/s40656-022-00538-3

Kassambara, A. (2023). *rstatix: Pipe-friendly framework for basic statistical tests* [Computer software]. Retrieved from https://CRAN.R-project.org/package=rstatix

Kaya, E. M., & Elhilali, M. (2017). Modelling auditory attention. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1714), 20160101. doi:10.1098/rstb.2016.0101

Khronos Group. (2017). *OpenGL 4.6 Core Profile Specification*. Retrieved from https://www.khronos.org/registry/OpenGL/specs/gl/glspec46.core.pdf

Kiepe, F., Kraus, N., & Hesselmann, G. (2023). Virtual occlusion effects on the perception of self-initiated visual stimuli. *Consciousness and Cognition*, *107*, 103460. doi:10.1016/j.concog.2022.103460

King-Smith, P. E., Grigsby, S. S., Vingrys, A. J., Benes, S. C., & Supowit, A. (1994). Efficient and unbiased modifications of the quest threshold method: Theory, simulations, experimental evaluation and practical implementation. *Vision Research*, *34*(7), 885–912. doi:10.1016/0042-6989(94)90039-6

Kleiner, M., Brainard, D., & Pelli, D. (2007). *What's new in Psychtoolbox-3?* [PowerPoint Slides]. Retrieved from https://pure.mpg.de/rest/items/item_1790332/component/file_3136265/content

Koch, C., & Ullman, S. (1987). Shifts in selective visual attention: Towards the underlying neural circuitry. *Matters of Intelligence*, 115–141. doi:10.1007/978-94-009-3833-5_5

Krasovskaya, S., & MacInnes, W. J. (2019). Salience models: A computational cognitive neuroscience review. *Vision*, *3*(4), 56. doi:10.3390/vision3040056

Krüger, A., Tünnermann, J., Stratmann, L., Briese, L., Dressler, F., & Scharlau, I. (2021). TVA in the wild: Applying the theory of visual attention to game-like and less controlled experiments. *Open Psychology*, *3*(1), 1–46. doi:10.1515/psych-2021-0001

Lentz, J. J., & Townsend, J. T. (2022). Merging the psychophysical function with response times for auditory detection of one vs. two tones. *Frontiers in Psychology*, *13*. doi:10.3389/fpsyg.2022.910740

Lima, D. da, & Ventura, D. F. (2023). A review of experimental task design in psychophysical eye tracking research. *Frontiers in Human Neuroscience*, *17*. doi:10.3389/fnhum.2023.1112769

Llanes-Jurado, J., Marín-Morales, J., Guixeres, J., & Alcañiz, M. (2020). Development and calibration of an eye-tracking fixation identification algorithm for immersive virtual reality. *Sensors*, *20*(17), 4956. doi:10.3390/s20174956

Maranes, C., Gutierrez, D., & Serrano, A. (2020). Exploring the impact of 360° movie cuts in users' attention. *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. doi:10.1109/vr46266.2020.00025

Martin, D., Malpica, S., Gutierrez, D., Masia, B., & Serrano, A. (2022). Multimodality in VR: A survey. *ACM Computing Surveys*, *54*(10s), 1–36. doi:10.1145/3508361

The MathWorks Inc. (2023). *MATLAB* (version: 9.14.0, R2023a) [Computer software]. Natick, Massachusetts, United States: The MathWorks Inc. Retrieved from https://www.mathworks.com

Mesgarani, N., & Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, *485*(7397), 233–236. doi:10.1038/nature11020

Meyer, G. F., & Wuerger, S. M. (2001). Cross-modal integration of auditory and visual motion signals. *Neuroreport*, *12*(11), 2557–2560. doi:10.1097/00001756-200108080-00053

Min, X., Zhai, G., Gao, Z., Hu, C., & Yang, X. (2014). Sound influences visual attention discriminately in videos. *2014 Sixth International Workshop on Quality of Multimedia Experience (QoMEX)*. doi:10.1109/qomex.2014.6982312

Min, X., Zhai, G., Gu, K., & Yang, X. (2016). Fixation prediction through multimodal analysis. *ACM Transactions on Multimedia Computing, Communications, and Applications*, *13*(1), 1–23. doi:10.1145/2996463

Min, X., Zhai, G., Zhou, J., Zhang, X.-P., Yang, X., & Guan, X. (2020). A multimodal saliency model for videos with high audio-visual correspondence. *IEEE Transactions on Image Processing*, *29*, 3805–3819. doi:10.1109/tip.2020.2966082

Mishra, E., Nikam, P., Vidhyadharan, S., & Cheruvalath, R. (2022). An affect-based approach to detect collective sentiments of film audience: Analyzing emotions and attentions. *Acta Psychologica*, *230*, 103736. doi:10.1016/j.actpsy.2022.103736

MongoDB Inc. (2022). *MongoDB*. MongoDB. Retrieved from https://www.mongodb.com/

Moore, B. C. J. (2012). *An introduction to the psychology of hearing*. Leiden: Brill. Retrieved from https://books.google.com.tr/books?id=LM9U8e28pLMC

Morgado, P., Li, Y., & Nvasconcelos, N. (2020). Learning representations from audio-visual spatial alignment. *Advances in Neural Information Processing Systems*, *33*, 4733–4744. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2020/file/328e5d4c166bb340b314d457a208dc83-Paper.pdf

Nachmias, J. (1981). On the psychometric function for contrast detection. *Vision Research*, *21*(2), 215–223. doi:10.1016/0042-6989(81)90115-2

O'Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., … Lalor, E. C. (2014). Attentional selection in a cocktail party environment can be decoded from single-trial EEG. *Cerebral Cortex*, *25*(7), 1697–1706. doi:10.1093/cercor/bht355

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, *42*(1), 107–123. doi:10.1016/s0042-6989(01)00250-4

Pelli, D. G. (1997). The videotoolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*(4), 437–442. doi:10.1163/156856897x00366

Pelli, D. G., & Farell, B. (1995). Psychophysical methods. *Handbook of Optics* (2nd ed., Vol. 1, pp. 29.21–29.13). New York: McGraw-Hill.

Pillai, J. S., & Verma, M. (2019). Grammar of VR storytelling: Analysis of perceptual cues in VR cinema. *European Conference on Visual Media Production*. doi:10.1145/3359998.3369402

Python Software Foundation. (2001). *Python* [Computer software]. Retrieved from https://www.python.org/

Qiao, M., Xu, M., Wang, Z., & Borji, A. (2021). Viewport-dependent saliency prediction in 360° video. *IEEE Transactions on Multimedia*, *23*, 748–760. doi:10.1109/tmm.2020.2987682

R Core Team. (2023). *R: A language and environment for statistical computing*. Vienna, Austria. Retrieved from https://www.R-project.org/

Revelle, W.. (2024). *psych: Procedures for Psychological, Psychometric, and Personality Research* [Computer software]. Evanston, Illinois. Retrieved from https://CRAN.R- project.org/package=psych

Rosseel, Y. (2012). lavaan: An r package for structural equation modeling. *Journal of Statistical Software*, *48*(2). doi:10.18637/jss.v048.i02

Ruesch, J., Lopes, M., Bernardino, A., Hornstein, J., Santos-Victor, J., & Pfeifer, R. (2008). Multimodal saliency-based bottom-up attention a framework for the humanoid robot icub. *2008 IEEE International Conference on Robotics and Automation*. doi:10.1109/robot.2008.4543329

Ryan-Warden, L., Ng, E., & Keating, P. (2019). Using background noise to improve sound localization following simulated hearing loss. *BioRxiv*. doi:10.1101/673806

Salvucci, D. D., & Goldberg, J. H. (2000). Identifying fixations and saccades in eye-tracking protocols. *Proceedings of the Symposium on Eye Tracking Research & Applications - ETRA '00*. doi:10.1145/355017.355028

Schubert, T., Finke, K., Redel, P., Kluckow, S., Müller, H., & Strobach, T. (2015). Video game experience and its influence on visual attention parameters: An investigation using the framework of the theory of visual attention (TVA). *Acta Psychologica*, *157*, 200–214. doi:10.1016/j.actpsy.2015.03.005

Schuetz, I., & Fiehler, K. (2022). Eye tracking in virtual reality: Vive pro eye spatial accuracy, precision, and calibration reliability. *Journal of Eye Movement Research*, *15*(3). doi:10.16910/jemr.15.3.3

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423. doi:10.1002/j.1538-7305.1948.tb01338.x

Shinkle, E. (2008). Video games, emotion and the six senses. *Media, Culture & Society*, *30*(6), 907–915. doi:10.1177/0163443708096810

Tavakoli, H. R., Borji, A., Rahtu, E., & Kannala, J. (2019). Dave: A deep audio-visual embedding for dynamic saliency prediction. *arXiv preprint arXiv:1905.10693*.

Tensorflow, Plakal, M., & Ellis, D. (2020). *YAMNet* [Computer software]. GitHub. Retrieved from https://github.com/tensorflow/models/tree/master/research/audioset/yamnet

Tomar, S. (2006). Converting video formats with FFmpeg. *Linux journal*, *2006*(146), 10.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136. doi:10.1016/0010-0285(80)90005-5

Treutwein, B. (1995). Adaptive Psychophysical Procedures. *Vision Research*, *35*(17), 2503–2522. doi:10.1016/0042-6989(95)00016-x

Tsiami, A., Katsamanis, A., Maragos, P., & Vatakis, A. (2016). Towards a behaviorally-validated computational audiovisual saliency model. *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. doi:10.1109/icassp.2016.7472197

Tsiami, A., Koutras, P., Katsamanis, A., Vatakis, A., & Maragos, P. (2019). A behaviorally inspired fusion approach for computational audiovisual saliency modeling. *Signal Processing: Image Communication*, *76*, 186–200. doi:10.1016/j.image.2019.05.001

Tsiami, A., Koutras, P., & Maragos, P. (2020). Stavis: Spatio-temporal audiovisual saliency network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4766-4776).

Van der Burg, E., Olivers, C. N., 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. doi:10.1037/0096-1523.34.5.1053

Vasser, M., & Aru, J. (2020). Guidelines for immersive virtual reality in psychological research. *Current Opinion in Psychology*, *36*, 71–76. doi:10.1016/j.copsyc.2020.04.010

Wada, Y., Kitagawa, N., & Noguchi, K. (2003). Audio–visual integration in temporal perception. *International Journal of Psychophysiology*, *50*(1–2), 117–124. doi:10.1016/s0167-8760(03)00128-4

Watson, A. B., & Pelli, D. G. (1983). Quest: A Bayesian adaptive psychometric method. *Perception & Psychophysics*, *33*(2), 113–120. doi:10.3758/bf03202828

Wickham, H. (2009). *ggplot2*. doi:10.1007/978-0-387-98141-3

Wolfe, J. M. (1994). Guided search 2.0 a revised model of visual search. *Psychonomic Bulletin & Review*, *1*(2), 202–238. doi:10.3758/bf03200774

Wolfe, J. M. (2021). Guided search 6.0: An updated model of visual search. *Psychonomic Bulletin & Review*, *28*(4), 1060–1092. doi:10.3758/s13423-020-01859-9

Xu, H., & Tian, F. (2020). Biological bias of saliency model in VR. *2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*. doi:10.1109/itnec48623.2020.9085055

Xu, J., Jiang, M., Wang, S., Kankanhalli, M. S., & Zhao, Q. (2014). Predicting human gaze beyond pixels. *Journal of Vision*, *14*(1), 28–28. doi:10.1167/14.1.28

Xu, M., Li, C., Zhang, S., & Callet, P. L. (2020). State-of-the-art in 360° video/image processing: Perception, assessment and compression. *IEEE Journal of Selected Topics in Signal Processing*, *14*(1), 5–26. doi:10.1109/jstsp.2020.2966864

Yantis, S., & Hillstrom, A. P. (1994). Stimulus-driven attentional capture: Evidence from equiluminant visual objects. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(1), 95–107. doi:10.1037/0096-1523.20.1.95

yt-dlp, pukkandan, & shirt-dev. (2021). *A youtube-dl fork with additional features and fixes* [Computer software]. Retrieved from https://github.com/yt-dlp/yt-dlp

Zhang, Y. (2021). ASOD60K: An audio-induced salient object detection dataset for panoramic videos. *arXiv preprint arXiv:2107.11629*.

Zhang, Y., Dai, F., Ma, Y., Li, H., Zhao, Q., & Zhang, Y. (2020). Saliency prediction network for 360° videos. *IEEE Journal of Selected Topics in Signal Processing*, *14*(1), 27–37. doi:10.1109/jstsp.2019.2955824

Zhang, Y., Zhang, L., Hamidouche, W., & Deforges, O. (2020). A fixation-based 360° benchmark dataset for salient object detection. *arXiv preprint arXiv:2001.07960*.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2018). Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(6), 1452–1464. doi:10.1109/tpami.2017.2723009

Zhu, D., Chen, Y., Zhao, D., Zhou, Q., & Yang, X. (2021). Saliency prediction on omnidirectional images with attention-aware feature fusion network. *Applied Intelligence*, *51*(8), 5344–5357. doi:10.1007/s10489-020-01857-3

Zhu, Y., Zhai, G., Min, X., & Zhou, J. (2020). The prediction of saliency map for head and eye movements in 360 degree images. *IEEE Transactions on Multimedia*, *22*(9), 2331–2344. doi:10.1109/tmm.2019.2957986

Zhu, Y., Zhu, X., Duan, H., Li, J., Zhang, K., Zhu, Y., … Zhai, G. (2023). Audio-visual saliency for omnidirectional videos. *Lecture Notes in Computer Science*, 365–378. doi:10.1007/978-3-031-46317-4_29

Zotter, F., & Frank, M. (2019). Ambisonics. *Springer Topics in Signal Processing*. doi:10.1007/978-3-030-17207-7