
SAARLAND UNIVERSITY

Faculty of Mathematics and Computer Science
Department of Computer Science



Active Modulation of Audio-Visual Perception to Enhance Human Performance and Experience in Immersive Environments

Dissertation zur Erlangung des Grades des Doktors der
Ingenieurwissenschaften der Fakultät für Mathematik und Informatik
der Universität des Saarlandes

Daniel Jiménez Navarro

Saarbrücken
2026

Date of the defense: 21st September 2026

Dean: Univ.-Prof. Dr. Jan Reineke

Chair: Prof. Dr. Jan Eric Lenssen

Reviewers:

Dr.-Ing. habil. Karol Myszkowski

Prof. Dr. Hans-Peter Seidel

Dr. Rafael Kuffner dos Anjos

Prof. Dr. Ana Serrano

Academic Assistant: Dr. José Luis Ponton Martinez



Eidesstattliche Versicherung

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit selbstständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus anderen Quellen oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet. Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form in einem Verfahren zur Erlangung eines akademischen Grades vorgelegt.

Declaration of original authorship

I hereby declare that this dissertation is my own original work except where otherwise indicated. All data or concepts drawn directly or indirectly from other sources have been correctly acknowledged. This dissertation has not been submitted in its present or similar form to any other academic institution either in Germany or abroad for the award of any other degree.

Saarbrücken, **2026**

gez. / signed

Daniel Jiménez Navarro

Abstract

Human perception determines how sensory information from the external world is interpreted and processed by the brain. Computer graphics techniques and recent advances in display hardware increasingly exploit the limitations of human perceptual systems to generate perceptually compelling content in immersive environments while reducing computational requirements, as exemplified by foveated rendering approaches. In immersive experiences, visual content is delivered through stereoscopic displays alongside omnidirectional spatial audio. This dissertation investigates how our perceptual systems prioritize and integrate information gathered from multiple sensory inputs, examining cases where one modality is weighted more heavily than another, or where competing signals within the same modality are differentially weighted. Specifically, the works presented in this dissertation focus on enhancing visual performance and reducing computational cost by leveraging multisensory interaction and perceptual weighting mechanisms. First, facilitatory effects arising from audiovisual integration are exploited to reduce saccadic latency by manipulating auditory cues associated with visual stimuli. Second, conflicting auditory and visual depth information is introduced in a depth perception task, demonstrating how audio-based depth cues can overdrive visually-driven depth perception, yielding faster vergence responses as a consequence. Third, perceptual assessment of binocular visual inputs reveals dynamic eye dominance behavior in both fovea and periphery, identifying monocularly dominated peripheral regions that can be leveraged to enhance existing foveated rendering techniques. Ultimately, the findings presented in this dissertation advance the understanding of multisensory perception, including audiovisual interaction and visual dominance, with a clear application-driven focus on enhancing visual performance and reducing computational requirements.

Zusammenfassung

Die menschliche Wahrnehmung bestimmt, wie sensorische Informationen aus der Außenwelt vom Gehirn interpretiert und verarbeitet werden. Computergrafiktechniken und jüngste Fortschritte in der Display-Hardware nutzen zunehmend die Grenzen des menschlichen Wahrnehmungssystems, um in immersiven Umgebungen visuell ansprechende Inhalte zu generieren und gleichzeitig den Rechenaufwand zu reduzieren, wie beispielsweise bei Foveated-Rendering-Verfahren. In immersiven Medien werden visuelle Inhalte über stereoskopische Displays zusammen mit omnidirektionalem räumlichem Audio wiedergegeben. Diese Dissertation untersucht, wie unser Wahrnehmungssystem Informationen aus verschiedenen Sinnesquellen priorisiert und integriert. Dabei werden Fälle betrachtet, in denen eine Modalität stärker gewichtet wird als eine andere oder in denen konkurrierende Signale innerhalb derselben Modalität unterschiedlich gewichtet werden. Die in dieser Dissertation vorgestellten Arbeiten konzentrieren sich insbesondere auf die Verbesserung der visuellen Leistung und die Reduzierung des Rechenaufwands durch die Nutzung multisensorischer Interaktion und Mechanismen zur Wahrnehmungsgewichtung. Zunächst werden die förderlichen Effekte der audiovisuellen Integration genutzt, um die Sakkadenlatenz zu verringern, indem die mit visuellen Reizen verknüpften auditiven Hinweise gezielt variiert werden. Zweitens werden in einer Tiefenwahrnehmungsaufgabe widersprüchliche auditive und visuelle Tiefeninformationen eingeführt. Dies zeigt, wie auditive Tiefenhinweise die visuelle Tiefenwahrnehmung überlagern und dadurch schnellere Vergenzreaktionen auslösen können. Drittens zeigt die Untersuchung der Wahrnehmung binokularer visueller Reize, dass die Augendominanz sowohl in der Fovea als auch in der Peripherie dynamisch variiert. Dabei lassen sich periphere Regionen identifizieren, die von einem Auge dominiert werden und sich zur Optimierung bestehender foveierter Rendering-Verfahren nutzen lassen. Die in dieser Dissertation präsentierten Ergebnisse erweitern das Verständnis der multisensorischen Wahrnehmung, einschließlich der audiovisuellen Interaktion und der visuellen Dominanz, mit einem klaren anwendungsorientierten Fokus auf die Verbesserung der visuellen Leistungsfähigkeit und die Reduzierung des Rechenaufwands.

Acknowledgements

I would like to express my gratitude to all the people who, in one way or another, helped me throughout this journey. It has been a great pleasure to experience everything these last few years had to offer, from stressful days in the office and daily meals at Mensa to more relaxed and funny moments during conferences and trips.

First of all, thank you to *Karol* for taking me on as one of his last PhD students and supporting me over these years, through meetings, insightful discussions, and everything in between. I could not imagine a better wall-to-wall supervisor, and I feel very lucky to have had you by my side. I would also like to extend my appreciation to *Hans-Peter* for allowing me to be part of such an amazing environment in which to work and grow as a person. It has been an honor, and I cannot imagine better conditions in which to start a research career than the ones I have had. This is something I will always be particularly grateful for.

Special thanks to *Ana*, who first introduced me to the idea of doing a PhD and continued to be deeply involved in every project I worked on, becoming my third unofficial supervisor from a distance. I would also like to thank all the people I collaborated with during this time, particularly *Qi* and *Colin*, who always brought new and interesting insights. And, of course, thank you to all the amazing colleagues I had the pleasure of working with throughout these years. You made me feel like a very lucky person and researcher.

To everyone in *D4*, both current and past members, thank you for always being willing to put your own work on hold to help me with early user studies whenever I asked. I really enjoyed our countless discussions and the many moments we shared outside the lab as well. A special mention goes to the administrative team: *Sabine*, *Petra*, *Anne*, and *Bertram*, for helping with everything unrelated to research and for being so kind and easy to talk to. I do not want to forget the *IST Team* either, who were always so efficient and professional whenever an issue arose.

Thanks to my lifelong friends, *Team Vinagres*, for all the moments and memories we have shared. I hope many more are still waiting for us in the future. Also, thank you to all the people from *Lituénigo* for your very warm welcomes on every visit I made, and particularly to everybody in *Énfasis*, which truly feels like a second family.

And last but not least, thank you to all my family for their continuous support. A special mention goes to my parents, *Ricardo* and *Mayte*, and my brother *Santi* :>, for being there through the ups and downs, and for those daily video calls that made me feel a little closer to home. You have always given me everything, and I will always be very proud of you.

*Un barquito de cáscara de nuez
Navegando con velas de papel
Va subiendo y bajando las olas
El barquito ya se fue...*

Contents

1	Introduction	1
1.1	Motivation	1
1.2	Contributions	5
1.3	Outline	7
2	Background and Previous Work	9
2.1	Human Visual System	9
2.1.1	Anatomy and Functionality	9
2.1.2	Visual Cues	10
2.1.3	Stereoscopic Perception	13
2.1.4	Oculomotor Responses	14
2.2	Human Hearing System	16
2.2.1	Anatomy and Functionality	16
2.2.2	Auditory Cues	17
2.3	Multimodality: Audiovisual Interaction	19
2.3.1	Spatial Fusion	19
2.3.2	Temporal Fusion	19
2.4	Perception on Immersive Displays	20
2.4.1	Hardware-driven Limitations	20
2.4.2	Vergence-Accommodation Conflict	22
2.4.3	Foveated Rendering	22
3	Accelerating Visual Saccades Latency	25
3.1	Introduction	25
3.2	Related Work	27
3.2.1	Visual Saccades	27
3.2.2	Audiovisual Interactions	28
3.3	Measuring Saccadic Latency Under Cross-modal Misalignments	29
3.3.1	Experimental Setup	30
3.3.2	Results and Discussion	33
3.4	Applications	37
3.4.1	Practical Application: Basketball Game	37

3.4.2	Interactive Application: Farm Game	39
3.4.3	Other Applications	41
3.5	Limitations and Future Work	42
3.6	Conclusion	42
4	Overdriving Visual Depth Perception	45
4.1	Introduction	45
4.2	Related Work	47
4.2.1	Depth Perception and Eye Movements	47
4.2.2	Limitations of Stereovision in VR	48
4.2.3	Depth Perception in Human Auditory System	48
4.2.4	Depth Perception with Multimodal Cues	49
4.3	Method	49
4.3.1	Sound Manipulation for Eye Convergence	50
4.3.2	Sound Manipulation for Eye Divergence	51
4.4	Sound-Based Depth Perception Modulation	52
4.4.1	Experimental Setup	52
4.4.2	Results	55
4.4.3	Discussion	57
4.5	Visual Response Acceleration	57
4.5.1	Experimental Setup	58
4.5.2	Results and Discussion	60
4.6	Application Case	61
4.6.1	Experimental Setup	62
4.6.2	Results and Discussion	63
4.7	Mixed Reality Application Version	64
4.8	Limitations and Future Work	67
4.9	Conclusion	68
5	Dynamic Visual Dominance	71
5.1	Introduction	71
5.2	Related Work	73
5.2.1	Stereoscopic Perception and Fusion	73
5.2.2	Eye Dominance	73
5.2.3	Foveation and Eye Dominance in Rendering	74
5.3	Experiment 1: Dynamic Dominance in Fovea	75
5.3.1	Experimental Setup	75
5.3.2	Results and Discussion	78

5.4	Experiment 2: Dynamic Dominance in Periphery	80
5.4.1	Results and Discussion	82
5.5	Application Cases	84
5.5.1	Guiding Visual Dominance-Contingent Rendering	84
5.5.2	Computational Savings Estimation in a VRS Prototype	87
5.5.3	Combined Dominance Map	88
5.6	Conclusion and Future Work	89
6	Conclusions and Future Work	91
6.1	Visual Saccades Latency Acceleration	91
6.2	Sound-(over)driven Depth Perception	92
6.3	Dynamic Visual Dominance	93
6.4	Concluding Remarks	94
	Bibliography - Own Work	95
	Bibliography	97
7	Appendices	117
7.1	Appendix A	117
7.2	Appendix B	119
7.3	Appendix C	123

List of Figures

1.1	Examples of visual perception tests	4
2.1	Anatomy of the eye	10
2.2	Cones and rods distribution in the retina	11
2.3	Picture showing monocular pictorial cues	12
2.4	Visualization of stereovision and binocular disparity	13
2.5	Types of oculomotor responses	16
2.6	Anatomy of the ear	17
2.7	Inter-aural differences (IAD) representation	18
2.8	Spatiotemporal interaction of audiovisual cues	20
2.9	Comparison between human and HMD field of view	21
3.1	Visual Saccades latency reduction – Teaser	26
3.2	Visual gap-overlap paradigm	28
3.3	Spatiotemporal setup and experimental conditions	30
3.4	Fixation and stimuli visual cues representation	32
3.5	Temporal profile of measured saccades	34
3.6	Saccade latency main results	35
3.7	Saccade latency main results: landing error	37
3.8	Basketball application illustration	38
3.9	Basketball application results: saccade latency and landing error	39
3.10	Saccade latency model and interactive farm game	40
4.1	Overdriving visual depth – Teaser	46
4.2	Sound modulation in convergence and divergence movements	51
4.3	Fixation, T_{ref} , and T_{test} visual appearance	53
4.4	Visualization of the vergence range considered	55
4.5	Overdriving depth perception main results	56
4.6	Fixation (red cross) and stimulus (clustered Landolt C) used as visual cues	58
4.7	Temporal scheme of the response time measured	59
4.8	Spatial representation of the audiovisual cues used in the cases of the Divergence-Far condition	59
4.9	Accumulated response times for the $T_1 + T_2$ sequence	60
4.10	Response times measured for each T_1 case	61
4.11	Response times measured for each T_2 case	61

4.12	Temporal behavior of angular eye gaze and vergence angle	62
4.13	Driving scene used in the VR application case	63
4.14	Spatial location of both scene items and their annotations in the VR driving application	64
4.15	Results obtained in the VR driving application	65
4.16	Right and left headset views during the MR application case	65
4.17	Spatial location of both scene items and their annotations in the MR application case	66
4.18	Results obtained in the MR application case	67
5.1	Dynamic Visual Dominance – Teaser	72
5.2	Overlap region of the head-mounted display	76
5.3	Fixation locations considered in Experiment 1 (fovea) and examples of blurred visual stimuli	77
5.4	Psychometric curves fitted for horizontal fixation locations	79
5.5	Psychometric curves fitted for vertical fixation locations	80
5.6	Foveal blur detection thresholds	81
5.7	Raw data used to fit the psychometric functions used to compute the blur detection thresholds for the fovea (Experiment 1, Section 5.3).	82
5.8	Fixation and stimuli locations considered in Experiment 2 (periphery)	82
5.9	Results of peripheral visual dominance study	83
5.10	Retinal dominance map	84
5.11	Transformation between blur maps and blur kernel	86
5.12	Peripheral dominance masks dichoptically applied in the gaze-contingent application	86
5.13	Results obtained in the dynamic visual dominance application study and the scenes employed	87
5.14	Shading rate image (SRI) used in the prototype desktop-based VRS application	88
5.15	2D representation of visual dominance according to both fixation and stimuli locations	88

List of Tables

3.1	Results obtained in the interactive farm game application	41
4.1	Physical depth and vergence angle values for the conditions considered	54
7.1	Specifications of the Varjo Aero headset	117
7.2	Specifications of the Varjo XR-4 headset	117
7.3	Specifications of the Beyerdynamic DT 770M headphones	117
7.4	Effect of the analyzed factors on saccade latency	119
7.5	Effect of the analyzed factors on landing accuracy	119
7.6	Post-hocs for the ΔT factor on saccade latency	119
7.7	Post-hocs for the ΔT *Eccentricity interaction on saccade latency	120
7.8	Effect of the analyzed factors on saccade latency (basketball scene) . .	120
7.9	Effect of the analyzed factors on landing accuracy (basketball scene) .	120
7.10	Post-hocs for the ΔT factor on saccade latency (basketball scene)	121
7.11	Effects of the analyzed factors on the overdriving success ratio	123
7.12	Post-hocs for the ΔL factor on success ratio	123
7.13	Binomial test between ΔL cases and Chance Rate for $E = 8$	124
7.14	Binomial test between ΔL cases and Chance Rate for $E=16$	124
7.15	Binomial test between ΔL cases and Chance Rate for Convergence . .	124
7.16	Binomial test between ΔL cases and Chance Rate for Divergence . . .	124
7.17	Binomial test between ΔL cases and Chance Rate for Near range . . .	125
7.18	Binomial test between ΔL cases and Chance Rate for Far range	125
7.19	Effects of the analysed factors on response times for T_1	125
7.20	Effects of the analysed factors on response times for T_2	125
7.21	Effects of the analysed factors on response times for $T_1 + T_2$	126
7.22	Effects of analysed factors on annotation association success ratio . . .	126
7.23	Binomial test between the success association ratio of each sound case and Chance Rate	126
7.24	Effects of analysed factors on annotation association success ratio in the MR application case	126
7.25	Binomial test between the success association ratio of each sound case and Chance Rate in the MR application case	127

Chapter 1

Introduction

This thesis investigates multimodal human perception and its applications from different perspectives in computer graphics, virtual and augmented reality. This introductory section presents the core motivation of this work in [Section 1.1](#), summarizes its main findings and contributions in [Section 1.2](#), and outlines the structure of this thesis in [Section 1.3](#).

1.1 Motivation

The Human Visual System (HVS) constitutes one of the primary mechanisms through which humans perceive the external world. Its complexity is the result of years of evolution adapting to real-world scene sensing, during which it has been used to transmit and consume information in several ways. Our HVS encompasses the optical capture of visual information by the eyes and the subsequent signal processing and interpretation by the brain. Such processing triggers not only active visual responses, such as eye movements, but also reactions in early cortical processing stages, such as the involuntary suppression of the nose contour from our field of view or the reflexive altering response elicited by the red color of blood.

Visual information is continuously captured by our eyes, where the retina –a sophisticated neural layer that contains photoreceptor cells– transforms light into electrical signals that enable, for example, color vision. These signals underlie a range of visual functions, including scene configuration understanding from depth cues or lens accommodation to sharply project the fixated point onto the retina. Lens accommodation serves as a monocular cue, as the ciliary muscles adapt lens curvature to bring the fixated object onto the fovea. Therefore, the accommodation process ends when the object edges are perceived as sharp. Another monocular, but dynamic, visual cue is motion parallax. It refers to the apparent displacement of objects at different depths due to head movement, where near and far objects move in opposite or the same direction of head motion, respectively. These and other monocular cues, such as occlusions, size, shading, and perspective cues from converging lines or vanishing points, are conveyed and processed to infer the three-dimensional structure of the external world.

Human binocular vision arises from the combined input of two eyes, each capturing a slightly different region of the visual field. The offset between the different views is a consequence of the eyes' separation, referred to as interpupillary distance (IPD), which typically ranges between 50 and 75 mm [Dodgson 2004]. Consequently, this dual visual input and its inherent geometrical properties form the basis of stereovision and its associated binocular cues. As a consequence of the IPD, the same point in the field of view projects to slightly different locations on each eye's retina, an effect known as binocular disparity. This binocular cue conveys depth information through the relative displacement between the two retinal images, particularly for regions falling outside Panum's fusional area around the horopter [Fender and Julesz 1967]. Even though our brain is able to fuse both views into a coherent, continuous representation of the visual field, both inputs are not weighted equally. In other words, visual dominance locally prioritizes one view over the other, contributing to a seamless percept across the whole field of view.

Maintaining stable binocular vision requires precise coordination of eye movements to compensate for eccentricity-dependent variations in ocular alignment and prevent diplopia arising from binocular disparity. A key factor here is the structure of the retina, where sharp, high-acuity vision is limited to the central fovea, whereas in the retinal periphery, temporal sensitivity takes precedence over spatial resolution. In this context, saccades are rapid eye movements that shift fixation from one point to another, redirecting gaze to project regions of interest onto the fovea. Saccadic execution involves coordinated movements of both eyes in the same direction, and their duration depends on the amplitude, typically lasting less than 100 ms. During saccades, inherent motion blur is mitigated by neural processing known as saccadic suppression, in which temporal fill-in mechanisms help preserve the continuity of visual perception. On the other hand, to resolve disparity misalignments that would otherwise produce diplopia, vergence movements are performed accordingly. Unlike saccadic movements, eye vergence requires the eyeballs to rotate in opposite directions to vary the vergence angle, defined as the angle between the gaze rays of the two eyes. Vergence movements are significantly slower than saccadic responses, with an approximate average duration of 400 ms–500 ms. In natural viewing conditions, such as scene exploration tasks, a combination of both movements is typically employed. Preparatory stage latency –the interval between the visual stimulus onset and the initiation of the oculomotor response– also impacts overall performance. This preparatory stage lasts approximately around 200 ms in pure saccadic motion, while it also becomes slightly slower in eye vergence movements, potentially reaching 250 ms on average. Other eye movements, such as smooth pursuit, further support detailed scene exploration by enabling continuous tracking of moving objects. Across all these eye movements, a coordinated action of both eyeballs is essential for accurately sampling and interpreting the visual environment represented on the display.

Display hardware has considerably advanced over the past few decades, progressively improving its capacity to support the HVS. Particularly, the introduction of stereoscopic head-mounted displays (HMDs) has enabled stereovision support through independent control of the image presented to each eye. Although considerable effort has been invested in designing these devices to closely replicate natural viewing conditions, not all monocular and binocular cues are reproduced with equal fidelity. For example, while visual movements like saccades or vergence can be conveniently measured using integrated eye tracking technology, more dedicated hardware, such as autorefractors, is needed to measure visual accommodation.

Similarly, depth cues are also constrained by the fixed focal distance of the display. Even though binocular disparity can be simulated by controlling the image rendered on each eye, thus allowing stereoscopic depth perception, excessive disparities with respect to the screen position might lead to uncomfortable viewing, commonly referred to as vergence-accommodation conflict (VAC). Ultimately, the perceptual limitations of the HVS reduce the fidelity requirements for reproducing convincing visual experiences, allowing computer graphics techniques to fill the gap between the technical capabilities of the display and actual HVS requirements.

Whereas visual information under real-world conditions is strictly determined by the observed scene, stereoscopic displays provide unprecedented control through direct access to the rendering pipeline and the virtual environment, enabling independent manipulation of the images presented to each eye. This feature not only facilitates advancing the understanding of visual perception but also enables improvements in visual performance that can exceed what is achievable under natural viewing conditions. Realizing this potential requires novel, carefully designed perceptual experiments to characterize the HVS behavior under such non-natural viewing conditions, while ensuring that any induced visual performance improvements remain controllable, stable, and free from adverse effects such as fatigue or discomfort.

Purely visual cues have traditionally been employed to improve both human performance, like in saccade guiding applications, and computational efficiency, such as reduced computational requirements in foveated rendering techniques. However, these approaches generally neglect the potential contribution of auditory modality. Similar to visual content, most immersive systems include spatial audio rendering as standard. While most efforts in computer graphics have traditionally focused on visual methodologies, audio-based approaches have received comparatively limited attention despite being equally available in modern immersive setups.

Auditory modality is considered the second most relied-upon sense for perceiving our surroundings, just after vision, and in many contexts it works as a natural complement to the visual capabilities. Auditory cues provide information about regions outside the field of view –such as areas behind our head or occluded by other objects– where visual information is absent or very limited. Conversely, distant objects may be detected visually while remaining beyond the effective range of auditory perception. The complementary nature of vision and hearing, which operate through partially independent perceptual channels, makes their combined integration in display systems a compelling framework for extending perceptual capabilities and enhancing immersive experiences.

The interaction between sensory channels is the subject of multimodal integration research, a field that has been studied extensively not only in the context of computer graphics applications but also as a means of understanding how the brain understands and fuses multiple sensory inputs into a final unified percept. Within computer graphics, combining visual and hearing modalities can offer novel, traditionally unexplored solutions for improving human performance and guiding perception. In particular, controlled manipulations of temporal or spatial relationships between audiovisual cues can yield faster and more comfortable responses with minimal computational overhead.

External signals (visual, auditory, or both) are rarely perceived with perfect fidelity. The study of how physical stimuli are transformed and interpreted by human sensorial systems falls within the domain of perceptual science. Such perceptual discrepan-

cies arise from the limitations of sensory organs and neural processing mechanisms. Some examples were already provided previously, such as signal suppression occurring during saccades or the nose being filtered from our visual field. Further examples include the imperceptibility of high-frequency light flicker below temporal sensitivity thresholds, or the narrowing of the effective visual field, commonly referred to as tunnel vision, during high-speed driving. Analogous perceptual effects also emerge in multisensory contexts, where multimodal interactions can systematically alter perception. While some perceptual effects can be beneficial as they improve performance, such as imperceptible peripheral color manipulations that reduce power consumption [Duinkharjav et al. 2022b], others can also impair visual detection performance with auditory cues under specific conditions [Jiménez-Navarro et al. 2024b].

Beyond sensory processing, visual perception is closely linked to cognitive interpretation as perceptual tests have long been used to study the psychology of the subject [Crutchfield et al. 1958], behavioral characteristics [Eriksen 1963], or to find out sensory impairments [Minschew and Hobson 2008]. To illustrate some visual tests performed during these studies, three examples are provided in Figure 1.1. The image on the left illustrates the rotating snakes illusion [Kitaoka and Ashida 2003], which arises from peripheral geometric configurations that produce a compelling sensation of rotational motion, triggered by microsaccades or blinking [Otero-Millan et al. 2012]. In the center image, a representative plate of the Ishihara blindness test is used to identify color vision deficiencies, in which the observer is asked about the number embedded within the chromatic dot pattern [Ishihara 1951]. Lastly, on the right image, a card from the Rorschach inkblot test is displayed. Such a test has historically been used to derive psychological profiles based on the observers' interpretations of ambiguous visual stimuli [Rorschach 1921]. More broadly, perceptual responses also play an important role in artistic experiences, where visual interpretation and emotional response may vary substantially across observers [Livingstone 2014].

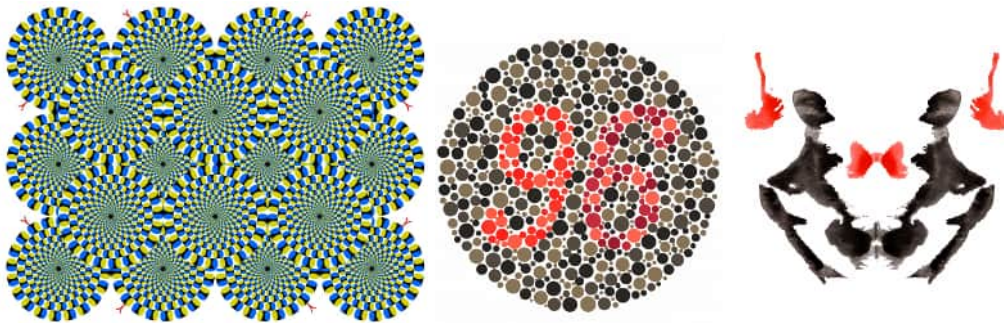


Figure 1.1: *Left*: Rotating Snakes illusion arises from geometric shapes on the peripheral viewing, triggered by blinking or microsaccadic movements. *Center*: Example of Ishihara color blindness test [Ishihara 1951], where chromatic differences must be resolved to see the embedded number. *Right*: Card number III of the Rorschach test used to psychologically evaluate the observer. Original artwork by Rorschach [1921].

In this direction, immersive technologies such as virtual and mixed reality (VR/MR) enable precise control over audiovisual content generation, augmenting the human visual field by adding rendered content to the real-world information. Similar to visual cues, auditory information can be controlled with comparable precision in immersive technologies. Therefore, such controllability enables the systematic study of audiovisual cues interaction and multimodal perception through carefully designed experimental paradigms targeting specific perceptual effects or visual behaviors.

On top of that, recently released devices, such as Meta’s AI-powered glasses¹ or Apple’s Vision Pro², exemplify an industry undergoing rapid expansion, now even further accelerated by advances in AI-driven solutions to optimize content generation.

In this thesis, our primary goal is to further investigate visual perception and performance by leveraging the controllability afforded by immersive environments. VR serves as an experimental platform to further study audiovisual perception regarding how information gathered from different modalities is processed, while finding potential visual performance gains arising from multisensory interaction. We quantify how audiovisual cue interactions can be manipulated to offer significant improvements in visual response performance, proposing direct applications in common experiences. Firstly, we model how spatiotemporal misalignments between visual and auditory cues of audiovisual stimuli can reduce saccade latency, yielding faster human responses than those achievable under natural viewing conditions. Secondly, we examine how visual depth perception can be modulated by sound-driven depth cues, improving eye vergence responses and enhancing the spatial alignment between real-world objects and their virtual annotations in MR applications. Finally, stereoscopic displays are used to investigate how visual inputs are weighted and processed, leading to visual dominance behavior. By independently controlling what is presented on each display, dynamic visual dominance is observed to systematically vary with eccentricities relative to the fixation point. Consequently, monocularly dominated peripheral regions are determined and a gaze-independent retinal dominance map is constructed. In gaze-contingent rendering, applying such findings leads to an improved user experience compared to static eye-dominance methods.

Importantly, the findings presented in this thesis are not limited to VR/MR systems but generalize across display technologies, as they reflect fundamental properties of how the brain processes multisensory integration and visual dominance across the field of view. These principles can therefore be exploited to further improve both human performance and computational efficiency. Additionally, we also contribute novel perceptual experiment designs and perceptual methodologies for studying audiovisual and interocular interactions under controlled settings.

1.2 Contributions

In this dissertation, the presented techniques leverage characteristics of the human visual and auditory systems to enhance user performance and experience in computer graphics, VR, and MR applications.

Visual Saccades Latency Acceleration Under natural viewing, up to 5 visual saccades are typically performed per second [Fischer and Weber 1993]. While the actual eye rotation during a saccade is fast, typically lasting 20–100 ms, the latency required to initiate the movement can be significant, usually around 200 ms [Duinkharjav et al. 2022a]. Since audio warning cues are known to speed up saccade reaction times [Ross and Ross 1981; Van der Stoep et al. 2015], we systematically study and model how much saccade latency times can be reduced with spatiotemporal misalignment at different eccentricities. This work is published as Jiménez-Navarro et al. [2024a] and is presented in Chapter 3. The main contributions are:

¹<https://www.meta.com/de/en/ai-glasses/>

²<https://www.apple.com/de/apple-vision-pro/>

1. We measure visual saccade latency reduction mainly driven by controlled temporal shifts between auditory and visual cues, showing no effect on audiovisual spatial shifts.
2. We provide a model to predict saccade latency according to spatio-temporal shifts on audiovisual cues and visual eccentricity of the saccade.
3. We validate the model in an interactive scenario where saccade latencies for new eccentricity values are estimated, demonstrating its potential application in e-sports or training in time-critical tasks.

Sound-based Modulation of Depth Perception While distance perception is mainly conveyed through visual depth cues, additional cues can also be obtained through auditory modality. We study how opposing both cues infer a multimodal depth percept that can enhance visual performance, while audio cues restore the original depth percept. Audio-based depth overdriving effect is calibrated, and findings are validated in application cases. Overdriving methodology and results are presented in [Chapter 4](#) and the work is published as [Jiménez-Navarro et al. \[2026b\]](#). The main contributions are:

1. We calibrate how audio-based depth cues can overdrive contradictory visual depth cues to bias the final depth perception.
2. We confirm how the depth overdriving effect can enable faster visual performance as visual depth variations are compensated by auditory depth cues, reducing eye vergence movement requirements.
3. We showcase potential application cases in both VR and MR, where sound-based depth cues influence scene understanding and interaction.

Dynamic Eye Dominance on Stereovision Over the years, different perceptual techniques have been proposed to enhance foveated rendering capabilities. Providing recent findings on the dynamic behavior of eye dominance in hand-eye interaction tasks, we study dynamic dominance in purely visual tasks, such as dichoptic blur detection, in both foveal and peripheral regions. Consequently, retinal dynamic dominance behavior is proposed to guide and enhance foveated rendering capabilities. These findings are discussed in [Chapter 5](#), and the work is published as [Jiménez-Navarro et al. \[2026a\]](#). The main contributions are:

1. We verify how visual dominance behaves dynamically in the fovea during a blur detection task, showing detection thresholds for monocular blur to be larger than those traditionally reported for binocular blur and even larger when applied in the non-dominant view.
2. We present a peripheral dominance map where visual dominance is spatially distributed along peripheral regions, with dominance switches originating from eye blind spots and naso-temporal asymmetry.
3. We evaluate our peripheral dominance map in a gaze-contingent application where extra blur applied in non-dominant regions is preferred over foveation methods that rely on static visual dominance, hence increasing user experience.

The author of this thesis is the principal contributor of all the above-mentioned works, which are fully covered in dedicated chapters as listed in the next paragraph. Additionally, another work addressing the perceptual blindness capabilities of carefully designed temporal sound cues was carried out prior to his Ph.D. studies [[Jiménez-Navarro et al. 2024b](#)]. Lastly, during the Ph.D. period, the author has been involved in other research projects related to measuring multimodal saccade latency [[Peng et al. 2024](#)] and speed perception [[Peng et al. 2025](#)]. Those contributions have not been included in this thesis.

1.3 Outline

This thesis is structured as follows. [Chapter 2](#) introduces background information focused on visual perception and a review of relevant literature. [Chapter 3](#) presents the model for reducing saccadic latency through temporally and spatially misaligned audiovisual cues, showing its applicability in a generic scenario where fast visual reaction is needed. [Chapter 4](#) covers the depth overdriving capabilities of audio cues tailored to different vergence visual movements involved in natural visual exploration. Corresponding improvements in visual response are shown, while the effect robustness is validated in VR/MR scenarios. [Chapter 5](#) studies visual dominance in a blur detection task relevant to foveated rendering. Visual dominance is not only found to behave dynamically in foveal regions, where dominance varies according to gaze direction, but also towards the periphery, where monocularly dominated regions are used to construct a retinal dominance map. Applying such a dynamic dominance approach leads to improved user experience compared to fixed eye dominance approaches. Lastly, [Chapter 6](#) synthesizes the findings of each presented work and discusses promising directions for future research.

Chapter 2

Background and Previous Work

In this chapter, we cover high-level concepts required to better understand the scope of this thesis and the systems involved, both human-related and hardware-specific. After introducing visual and auditory senses, how they work and their capabilities, multimodal integration and interaction mechanisms are also discussed. Lastly, hardware limitations inherent to immersive environments are introduced, along with a review of previous works aimed at addressing all these aspects.

2.1 Human Visual System

We first address the sensory system most relevant to computer graphics: the Human Visual System (HVS). After briefly introducing its anatomy and functionality, we discuss both static aspects of visual perception and dynamic processes such as eye movements.

2.1.1 Anatomy and Functionality

Light enters our eyes and is projected on the retina, a thin surface that transforms light rays into electric signals. Therefore, the visual information pipeline starts from the light wavelengths that go through the eye cornea, are transformed by the retina, and are transmitted to the brain, where they are interpreted [Marr 1980]. In addition to the retina, other important eye elements are the pupil, which can adapt its size to incoming light, or the lens, which accurately focuses the retinal projection to get a sharp representation of the focused element by the ciliary muscles, a process referred to as accommodation [Ott 2006]. Imitating this accommodation procedure, cameras adjust their focal length to bring objects at different distances into sharp focus. Also, due to the crystalline lens distortions, light is not uniformly projected to the retina. In this context, the fovea –the retinal region where light is most precisely focused– exhibits the highest visual acuity and spatial detail resolution [Hendrickson and Yuodelis 1984]. Retinal regions farther from the fovea are referred to as the periphery, whose features significantly differ from those of the fovea. Figure 2.1 illustrates the anatomical structure of the human eye.

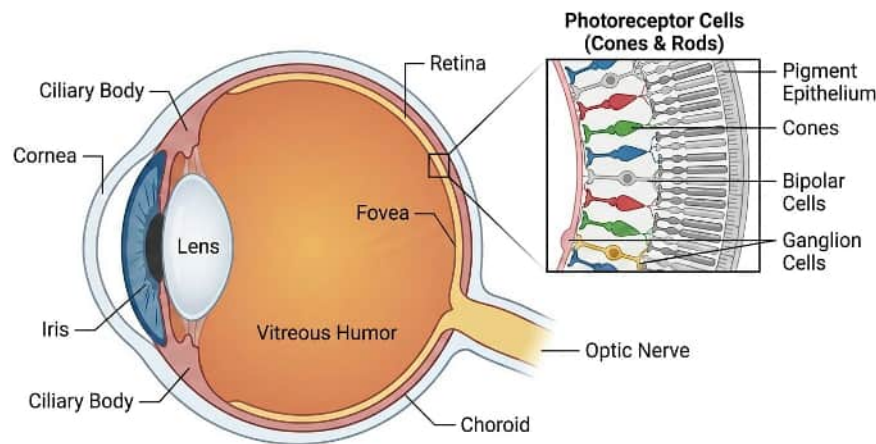


Figure 2.1: Cross-sectional illustration of the eye anatomy. Eye components such as the lens, retina and fovea are referenced, while the distribution of cone and rod photoreceptors in the retina is zoomed in. Graphical elements were generated using *Nano Banana Pro* [Google DeepMind 2025].

The eye retina is composed of light-sensitive cells that react to the incoming light. The two main photoreceptor types in the retina are cones and rods [Hecht 1937]. Cones react particularly well under daylight brightness conditions and enable color vision. Regarding the fovea, cones cover the whole foveal region of the retina as they provide high spatial resolution. Outside the fovea, their density decreases with eccentricity [Provis et al. 2013]. On the other hand, rods are specialized to work under low luminance conditions, therefore complementing cones' functionality. In this case, rods are not present in the fovea but become the majority in peripheral regions, decreasing as we get deeper into the periphery [Kolb 2012]. Consequently, photoreceptors are not uniformly distributed across the retina, with markedly different densities observed between the fovea and the periphery. As a result, the range of resolvable spatial frequencies decreases with increasing eccentricity. A representation of cones and rods densities from the fovea to peripheral regions is shown in Figure 2.2. Another interesting feature of the visual periphery is the physiological blind spot, a retinal region devoid of photoreceptors where the optic nerve exits the eye and transmits retinal information to the brain. Importantly, the visual-field regions corresponding to the blind spots of the left and right eyes are spatially distinct. The blind spot is located approximately at 15° (temporal) from the fovea and can be demonstrated through simple perceptual experiments [Ramachandran 1992].

2.1.2 Visual Cues

Visual information is processed and understood in different ways, and those entities that help to interpret the perceived representation are called visual cues. Visual cues can be classified in different ways and provide information not only related to our field of view but also about our own movement.

Monocular Pictorial Cues Monovision from single-eye viewing can understand the world structure and scene composition through bidimensional (2D) information, as

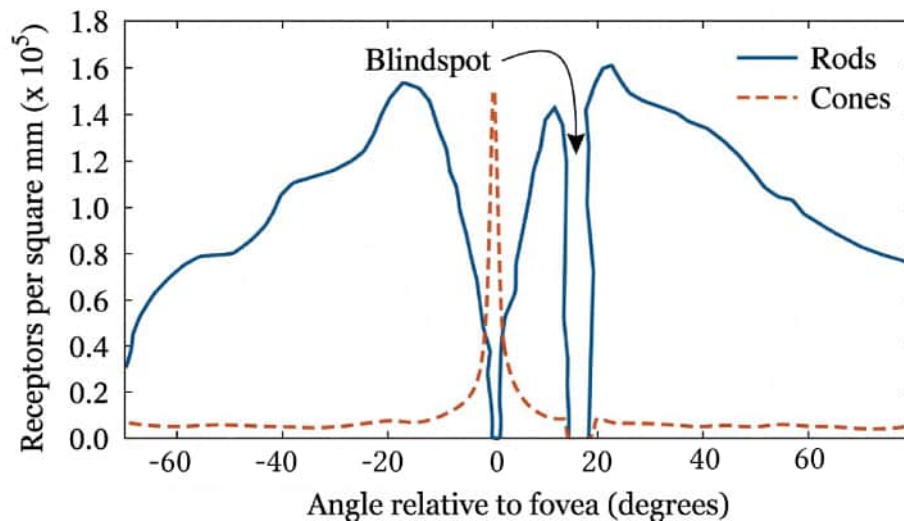


Figure 2.2: Distribution of cone and rod densities as a function of angular distance from the foveal center. While cones dominate foveal regions, rods are mainly spread towards peripheral regions. The figure has been drawn based on [Wandell 1995].

happens when looking at flat photographs. Even though only 2D cues are available, an illusory 3D representation is formed when visually analyzing the realistic image.

While the depth dimension is lost, pictorial cues such as perspective projection recover it to a certain extent by analyzing how the 3D world is bi-dimensionally represented from the specific location from which the picture was taken. Several depth cues arise from perspective projection, such as the convergence of parallel lines towards vanishing points on the horizon in the 2D representation [Palmer 1999]. Similarly, an object's location relative to the horizon line can indicate its distance from the observer.

Other cues, such as relative and familiar sizes [Ittelson 1951], also contribute to depth perception, since objects subtending smaller visual angles are typically perceived as more distant. Occlusion provides an additional monocular cue by establishing ordinal depth relationships, where occluding surfaces are perceived as being closer than the surfaces they partially conceal. Shading and cast shadows within the scene likewise provide important depth information [Kersten et al. 1996].

Additional pictorial depth cues arise from the projection of three-dimensional scenes onto the retinal image. Texture gradients, for example, provide information about surface orientation through systematic changes in the size, density, and shape of texture elements with distance [Gibson 1950]. Atmospheric, or aerial, perspective constitutes another important cue, whereby distant objects exhibit reduced contrast and altered color appearance due to light scattering and absorption in the atmosphere before reaching the observer [O'Shea et al. 1994], typically used by artists when portraying distant landscapes. Several of those monocular depth cues are present in the image displayed in Figure 2.3.

Binocular Cues: Stereovision Humans, as well as most of the animals, have two eyes for survival reasons which extend monocular vision capabilities, such as accurate depth perception, an extended field of view, efficient spatial navigation and task performance [Sheedy et al. 1986]. In the previous section, monocular depth cues such as perspective or occlusions were addressed, which relied on geometrical and con-



Figure 2.3: Photograph showing a shopping street. As can be observed, several pictorial depth cues are present, such as converging lines, relative sizes, texture gradients, shadows, and occlusions.

textual information available from a single viewpoint. Binocular viewing, however, provides additional depth cues, such as binocular disparity or eye vergence, which significantly enhance spatial awareness. This concept relates to an old fundamental limitation in computer vision, as accurate 3D reconstructions cannot generally be recovered from a single 2D image. To get a proper 3D representation and estimate depth reliably, at least two views of the same scene are needed [Moons et al. 2010], which is analogous to the binocular inputs naturally available in the HVS. This forms the basis of stereovision, where slight spatial offsets between the retinal images of the two eyes are used to reconstruct the 3D space through stereopsis [Poggio and Poggio 1984; Cumming and DeAngelis 2001]. The functionality of stereovision is shown in Figure 2.4.

The horizontal separation between the eyes –commonly referred to as interpupillary distance (IPD)– gives rise to binocular disparity, whereby the same object in the visual field is projected onto different retinal locations in each eye. As illustrated in Figure 2.4, binocular disparity can be observed from the relative positional differences of the scene objects between each view. In this example, eyes are fixated on the horse, thus retinal representations of the horse match between eyes as both foveae are aligned. In that case, there is zero retinal disparity. Alternatively, regions at different depths from this focal plane, such as those for the apple or the tree, introduce retinal disparity, since the projections of these objects onto the two retinæ no longer correspond to the same relative positions.

These variations in retinal disparity provide a powerful cue for depth perception, allowing the visual system to infer relative object distances within the scene [Qian 1997]. In contrast, points lying on the horopter produce zero retinal disparity and are projected onto corresponding retinal locations in both eyes. Small disparities around the horopter –typically below 1° of visual angle– can still be fused into a

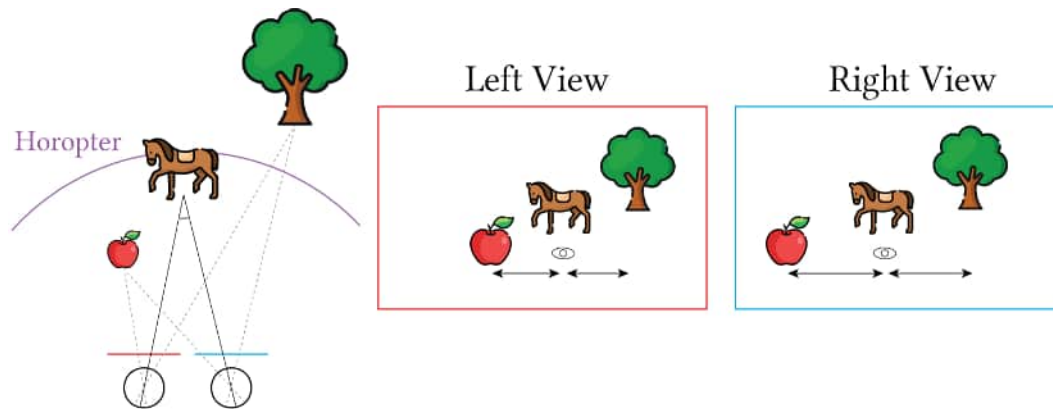


Figure 2.4: Views captured by our eyes slightly vary and introduce binocular disparity differences between the scene elements. The relative separation between the tree and the apple, with the horse fixated on, changes depending on the eye, as shown in the left and right views. Consequently, resolving such disparities allows us to construct a 3D representation of our visual field, therefore introducing information of the depth dimension.

single percept within Panum's fusional area [Fender and Julesz 1967; Schor and Tyler 1981], thereby defining a limited range of depths where binocular fusion remains stable. However, when disparities exceed this fusion range, binocular fusion fails and the observer perceives double vision, a phenomenon known as diplopia [Mitchell 1966]. In this context, points falling in front of or behind the horopter are perceived through stereopsis, which interprets these retinal disparities as depth information. Importantly, binocular disparity is determined not only by scene geometry but also by the fixation distance (focal plane) and eye vergence. By means of vergence eye movements, the horopter can be shifted to different depths, thereby modifying retinal disparity distributions across the visual field during natural viewing.

2.1.3 Stereoscopic Perception

After presenting the benefits of binocular viewing and stereovision in depth estimation and spatial reconstruction tasks, we now address how much the object representation captured by each eye contributes to the final percept of that object. Thresholds exist on how different both representations can be to be fused (binocular fusion) or what happens if they are too different to be fused (binocular rivalry), mentioning when one is suppressed by the other (binocular suppression).

Binocular Fusion, Rivalry and Suppression Once our eyes precisely converge at one object, each monocular image falls properly into each eye's fovea, resulting in an accurate representation due to binocular fusion. Binocular fusion depends on several factors, such as retinal eccentricity [Crone and Leuridan 1973], spatial resolution [Schor et al. 1984], and temporal aspects [Wolfe 1983].

When sufficiently different images are seen by each retina, perceptual fusion fails, and binocular rivalry occurs [Levelt 1965b]. Although this is rare in normal viewing conditions, it can be easily generated in stereoscopic displays. During this rivalry, the image or stimulus that can be seen at a certain timestamp receives the name of

dominant stimulus, while the one not perceived is referred to as suppressed stimulus [Tong et al. 2006]. Extending binocular fusion thresholds, prior works studied how visual features such as brightness and contours [Levelt 1965a] make binocular views non-fusible, triggering binocular inhibition or suppression. Monocular suppression also occurs, for example, when two views present different blurs –thus affecting dichoptic contrast– where the blurred input can be suppressed in favor of the clear, sharp view [Fahle 1982].

Although not directly addressed in this work, dichoptic color has also been widely studied, particularly to reduce headset power consumption [Duinkharjav et al. 2022b; Chen et al. 2024]. In this direction, the same stimulus presented with different colors or textures to each eye does trigger binocular rivalry, if inputs alternate, or color mixing, where a third color is perceived. Tone saturation, color differences, and stimulus size and duration set the thresholds for this mixing to occur [Hecht 1928].

Visual Dominance When two visual inputs are available to humans, it is known that one of them is preferred over the other. The eye that is favored when only a single eye can be used in coordination tasks or when both views are discrepant is defined as the dominant, while the one less relied on is the non-dominant input [Porac and Coren 1976]. This condition also occurs with our hands, as one of them is mostly used to write, use cutlery, or interact with the smartphone or other devices. In such cases, the hand mostly used due to faster and more accurate movements due to practice is referred to as the dominant one, with most people generally being right-hand dominant. Similar to handiness, the same occurs with our visual system. In this context, one visual input is more relied upon than the other in both perceptual [Shneor and Hochstein 2005] and performance tasks, such as aiming and shooting.

Visual dominance in computer graphics has been considered in foveation techniques, mainly to increase foveation levels in the assumed non-dominant view [Meng et al. 2020]. Although visual dominance was traditionally understood as a fixed or static human feature, recent work shows its dynamic behavior under certain visual applications and eye-hand interaction tasks [Prummer et al. 2025a].

2.1.4 Oculomotor Responses

The information presented in the visual field is sequentially explored and processed by different eye responses towards corresponding regions of interest (ROI). While fixations allow for keeping the ROI projected in the fovea, where spatial and chromatic resolutions are the highest, tracking motions are needed to align the fovea with the ROI located at a different position in the field of view. In this section, the main active eye movements performed to explore our field of view are introduced.

Visual Saccade We coordinately move our eyes around 3-5 times per second to change the point we focus on between locations across the visual field [Ross et al. 2001]. In this motion, the main goal is to align our fovea with the corresponding ROI to get the most accurate representation and properly convey visual content. In order to achieve that, both eyes rotate in the same direction to locate the foveae in a position where we can sharply see the ROI. This conjugated eye rotation is performed simultaneously with both eyes and is called a saccade (Figure 2.5, left). Such movements are the fastest ones we can actively perform with our eyes, finishing

just around 20-100 ms after starting [Baloh et al. 1975]. A more restricting temporal aspect is the interval between planning and action, commonly referred to as saccade latency, which is around 200 ms [Fischer and Weber 1993].

Saccades are consequently rapid ballistic movements that enable our fovea to better collect visual information according to its capabilities. When saccades are performed, our fovea is rapidly moved, which results in temporal inconsistencies as the visual representation cannot be formed. Theoretically, this would imply a completely blurry visual representation. However, our brain is able to detect when these saccades are performed and suppresses vision until the saccade is performed [Matin 1974]. This concept is known as saccade suppression and allows us to conveniently perform saccades without any dizziness coming from blurry representations. Such saccadic suppression is used in rendering techniques [Kwak et al. 2024] to alleviate resolution requirements during saccadic motion. Interestingly, saccades cannot be redirected once they have been started. Consequently, efforts have been made to predict saccade landing and adapt quality beforehand, by reducing resolution in the corresponding peripheral regions after landing [Arabadzhiyska et al. 2017].

Visual Vergence When objects are tracked in distance or depth, vergence movements are involved. While an object is being fixated on, the object's projection falls into both foveas, and the angle formed by both gaze rays is referred to as the vergence angle. In order to change our fixation point to a different depth, vergence eye movements are required to properly focus at the new desired depth (Figure 2.5, center). Vergence movements performed toward nearer objects are referred to as convergence, while those toward farther objects are referred to as divergence. Such vergence eye movements are primarily driven by binocular disparity and are coupled with accommodation. As we previously mentioned, binocular fusion can be maintained for a limited range of retinal disparities around corresponding retinal points, commonly referred to as Panum's fusional area.

Conversely to saccadic movements, both eyes now rotate in different directions by performing a disconjugate movement. Vergence completion times are much slower than those for saccades, reaching 500 ms and more [Robinson 1966]. Preparatory phase is also slightly slower, with latencies lasting around 200 – 300 ms. Due to its nature, pure vergence motions are not very common to be performed standalone. Instead, they are usually combined with saccades, commonly referred to as combined movements.

Combined Movement Generally, a combination of saccades and eye vergence movements is mostly what is performed in exploration tasks. To analyze the visual field, visual saccades are combined with vergence motions as ROI are located at different depths in our three-dimensional world (Figure 2.5, right). Due to this synchronized motion, vergence response is facilitated by saccades and its duration is shorter than pure vergence movements [Zee et al. 1992; Enright 1986; Yang et al. 2002].

Other Visual Movements Once a moving object is brought onto the fovea, it can be tracked through smooth pursuit eye movements (SPEMs), which continuously adjust eye velocity to match the motion of the target [Robinson 1965]. Smooth pursuits are also performed in daily tasks like reading, when complete saccades are not

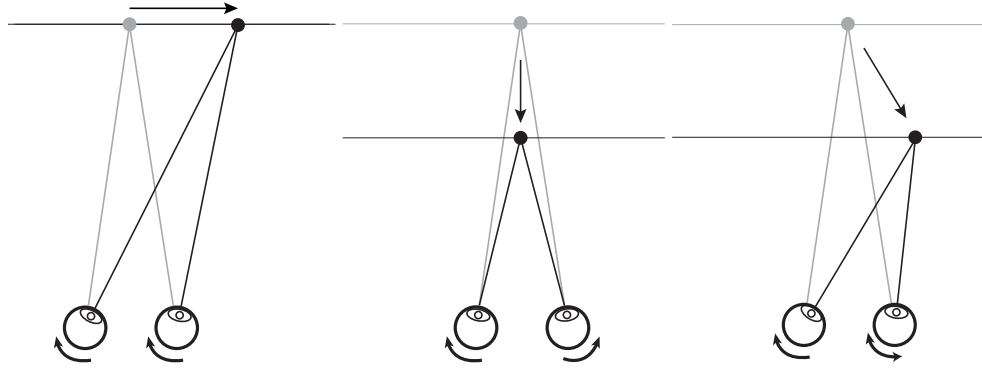


Figure 2.5: Scheme representing oculomotor responses such as saccades (*Left*), vergence (*Center*), and combined movements (*Right*). In this combined movement example, the right eye motion depends on the relative size of the saccade and vergence.

needed due to the limited range involved in the visual motion. Unlike saccades, which are ballistic and cannot be modified once initiated [Arabadzhiyska et al. 2017], smooth pursuit movements rely on continuous visual feedback. When target motion exceeds the tracking capability of smooth pursuit, catch-up saccades are intermittently triggered to re-foveate the object. Additionally, involuntary movements such as microsaccades and tremors also exist. The function of those small, jerky microsaccadic motions involved during fixation is to counteract visual fading [Martinez-Conde et al. 2006], and they have been reported to indicate attentional shifts [Gu et al. 2024].

All eye movements discussed so far are performed without involving head motion, which offers an alternative to explore the visual field other than pure eye movements. In this line, vestibular eye movements allow us to steadily fixate on an object when the head or body moves, while also being faster and more accurate than visual pursuit [Leigh and Zee 2015]. Eye motion in this case is able to accurately compensate for body motion via vestibular information in our inner ears. To easily observe how these vestibular movements work, one could fixate on one point while tilting the head in every possible direction. In this case, the vestibular movements cancel out the head motion performed, ensuring a fixed visual representation of the fixated object.

2.2 Human Hearing System

Similar to the HVS functionality of transforming light rays into electric signals, the human hearing system (HHS) transforms air vibrations into electric signals that are processed and interpreted by the brain. While the visual system detects content presented within the visual field limits, the auditory system works almost omnidirectionally, meaning that spatialized audio signals can be detected from every direction. This sense is very twined with speech skills, needed for verbal communication.

2.2.1 Anatomy and Functionality

Our ears are organs that are sensitive to sound pressure being transmitted by the corresponding medium, being generally air. This requirement sets another difference from the visual system, where photons that make light wavelengths do not need any specific medium to travel.

With sound, the lack of media means no differences in pressure level can occur, thus no sound can be effectively heard. This occurs, for example, in the space where pressure differences have no medium to travel. Another example of this is how different our voice sounds in normal conditions when speaking, compared to when we try to speak underwater. The change of medium from air to water reflects how sound is transmitted and, therefore, how it is perceived by our ears. Anatomically, the structure of the ears can be divided into three separate sections. The external or outer ear is the visible ear and collects external sound waves with the eardrum located in the middle ear. In the middle ear, eardrum vibration is amplified by a chain of small bones and transmits the signal to the inner ear. Lastly, in the inner ear, mechanical vibrations are transferred to fluid movements that are transformed into electric signals by the hair cells. In this transformation, external signals are encoded into neural information. A representation of the human ear anatomy is shown in Figure 2.6.

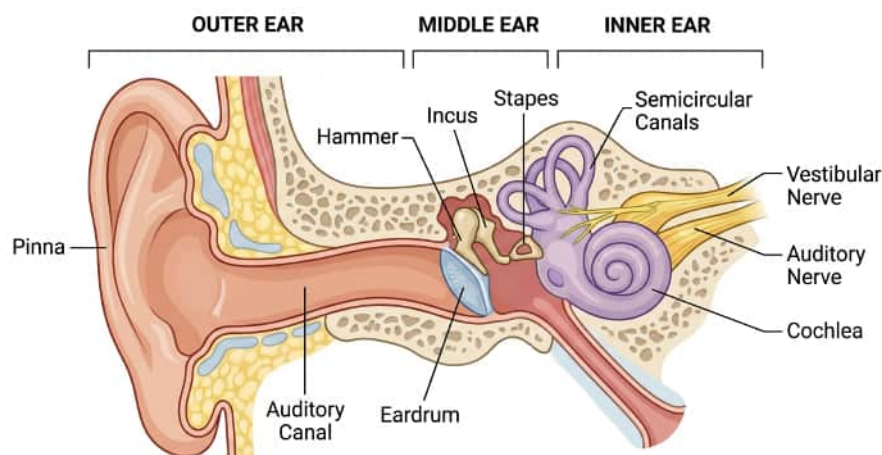


Figure 2.6: Cross-sectional illustration of the ear anatomy. Ear components such as pinna, eardrum and auditory nerve are referenced, while outer, middle and inner ear regions are separated. Graphical elements were generated using *Nano Banana Pro* [Google DeepMind 2025].

2.2.2 Auditory Cues

Similar to the HVS, some auditory cues can be detected by a single ear, while others are obtained by comparing information received by both ears, defining monaural or binaural hearing cues, respectively.

Monaural Cues Distance perception through monaural hearing is achievable by means of several auditory cues collected by a single ear channel. The main monaural cue is sound loudness [Warren et al. 1958], a psychological property related to the air pressure reaching the listener's eardrum. The intensity of sound signals is measured in decibels (dB) and indicates the sound pressure perceived by the listener. This perceived intensity is related to the power of the sound source and decreases approximately logarithmically with distance from the listener. Consequently, analogous to relative size comparisons or size familiarity in monocular pictorial cues, relative sound differences can also be used to estimate distance to the audio sources.

Sound distance estimation also benefits from the frequency content present in the audio signal, as there is a distinction in how high-frequency and low-frequency content attenuate with the atmosphere. This media absorption of high frequencies allows our hearing system to assess sound distance based on the signal frequency content [Naguib and Wiley 2001].

Also, sound reverberation especially informs about the room acoustics [Ribeiro et al. 2010]. Although we are often not consciously aware of it, the human brain is able to process these echoic cues to assist auditory scene interpretation. For example, echo-based localization is highly developed in animals such as bats, which perform active echolocation by emitting acoustic pulses and analyzing the returning echoes.

Binaural Cues HHS also benefits from the dual input offered by our ears, conveniently located at each side of our head. Similar to how our binocular vision enables stereoscopic vision and its benefits in depth perception or a larger field of view, for example, binaural cues also offer better distance estimation and sound localization capabilities. Our brain detects intensity and temporal differences between the signals processed on each ear, commonly referred to as Interaural Intensity Differences (IID) [Gaik 1993; Grantham 1984] and Interaural Temporal Differences (ITD) [Wightman and Kistler 1992; McFadden and Pasanen 1976]. Such cues are used to spatially localize sound sources, particularly allowing a better estimation of distance and lateralization. A visualization of these interaural differences (IAD) is shown in Figure 2.7.

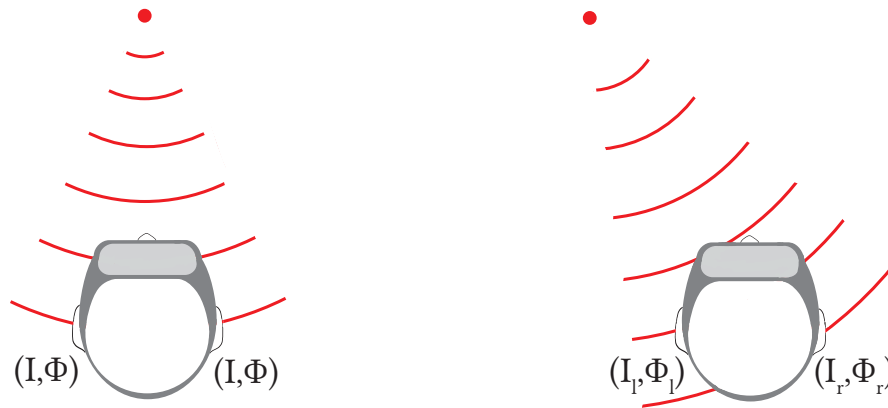


Figure 2.7: Inter-Aural differences (IAD) Visualization. *Left*: The picture represents an audio source (red dot) located at some distance in front of the listener and its corresponding wavefront. In this case, intensity (I) and temporal (phase, Φ) differences between left and right ears do not exist as the travel path is the same. *Right*: In this case, the audio source is located towards the left side of the listener. Accordingly, the travel distance from the audio source to each ear is not the same, thus introducing intensity and temporal differences between ears: $(I_l, \Phi_l) \neq (I_r, \Phi_r)$. This can be visualized with the corresponding wavefront. Our brain detects and interprets those differences in intensity and temporal delays to spatially localize the audio source.

2.3 Multimodality: Audiovisual Interaction

External information rarely involves only a single human sense or modality. Generally, entities are perceived by more than one sensory input that provides complementary information. For example, ambulances turn on both lights and warning alarms when on duty, or the strong burst that occurs when a glass breaks into small pieces.

Involving more than one sensory input has been found to influence human capabilities in several ways. Several facilitatory effects arise from multimodality support [Martin et al. 2022], such as a better environment perception when including a correct sound spatialization [Chen et al. 2020] or proprioceptive cues [Poupyrev et al. 1998], improving experience realism and sense of presence with olfactory cues such as scent [Gonçalves et al. 2019], or enhancing user performance by the combination of hearing and haptic information [Hecht et al. 2006].

Alternatively, multimodality can also be used to generate illusory or inhibitory effects, such as body ownership illusions [Petkova and Ehrsson 2008] or visual performance degradation effects driven by auditory cues under specific spatiotemporal circumstances [Malpica et al. 2022]. In Chapters 3 and 4 of this thesis, we focus on the facilitatory interaction between vision and hearing modalities, both in spatial and temporal domains.

2.3.1 Spatial Fusion

Focusing on auditory and visual cues, the spatial location of each cue and the resulting interaction affect the final percept of the audiovisual stimuli. Regarding the spatial domain, visual localization cues tend to dominate the perceived spatial position of audiovisual stimuli, biasing the perceived auditory location towards the visual cue [Bischoff et al. 2007]. This spatial binding phenomenon is known as the ventriloquism effect, pioneered by Edgar John Bergen (1903-1978) and his dummy Charlie McCarthy [Encyclopædia Britannica [n. d.]]. The ventriloquist synchronizes the movements of his puppet's mouth with his own speech, causing the sound to be perceived as originating from the puppet [Slutsky and Recanzone 2001]. In this case, audiovisual fusion is performed spatially, as sound localization is moved towards the visual cue –the puppet– as depicted in Figure 2.8 (left).

2.3.2 Temporal Fusion

While visual cue controls spatial perception of the audiovisual information, the auditory cue controls the temporal perception. The term temporal binding window (TBW) stands for the time interval between both audiovisual cues that can still be resolved to perceive both stimuli as integrated [Wallace and Stevenson 2014]. In this case, temporal delays between visual and auditory cues can affect the timestamp when each cue is perceived, where broader TBW are found when the auditory cue precedes the visual information [Cecere et al. 2016]. This means that the auditory cue dominates over the temporal dimension, biasing when the audiovisual stimuli is perceived. Therefore, the timestamp when a visual cue is perceived is temporally driven towards the auditory cue. Another example of this behavior is found in the well-studied sound-induced flash illusion (SIFI), where multiple beeps control the perception of visual flashes, even though they do not match [Hirst et al. 2020; Shams et al. 2005]. A representation of the SIFI effect is shown in Figure 2.8 (right).

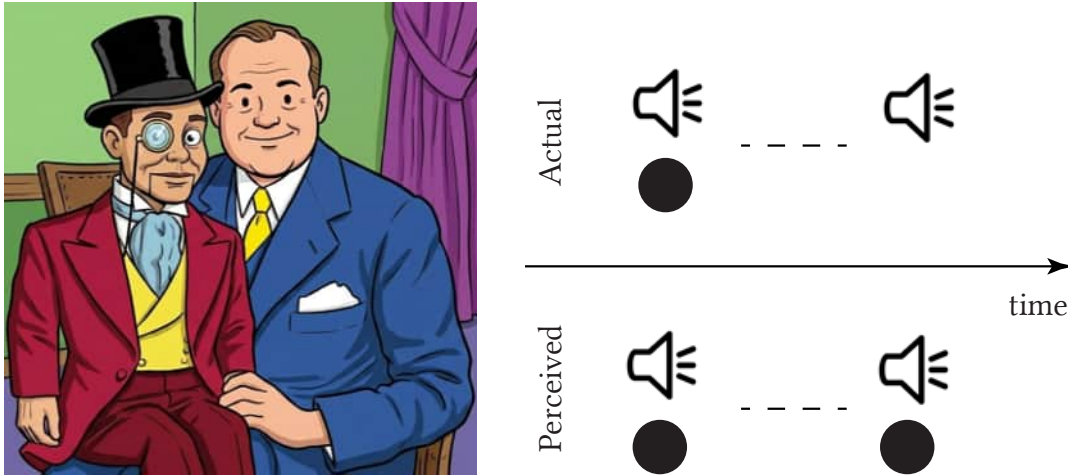


Figure 2.8: *Left*: Illustration of the ventriloquism effect showing the ventriloquist and his puppet, stylized using *Nano Banana Pro* [Google DeepMind 2025]. By making sneaky sounds with his own mouth at the same time the puppet’s mouth is articulated, it gives the impression that the puppet is actually speaking. *Right*: Representation of the sound-induced flash illusion (SIFI) in the temporal dimension. While the visual input is only truly generated in the first audiovisual stimulus, the second beep makes the flash be perceived twice, once per beep heard.

2.4 Perception on Immersive Displays

Immersive experiences are commonly delivered through stereoscopic head-mounted displays (HMDs) together with spatial audio reproduction. While these devices provide substantial control over the generated audiovisual content, they remain subject to hardware-specific limitations that must be considered when designing such immersive experiences. In addition to the perceptual constraints of the visual and hearing systems, which motivate techniques such as foveated rendering, prior work has also investigated limitations introduced by the hardware itself, including discrepancies between the perception of audiovisual cues in real and virtual environments.

2.4.1 Hardware-driven Limitations

Head-mounted devices continue to evolve toward higher resolutions, improved portability, reduced weight, and lower power consumption. At the same time, integrated AI solutions enable new forms of virtual and immersive experiences by improving content generation and facilitating more natural interaction. Nevertheless, both perceptual and hardware constraints must still be considered to optimize perceived content fidelity while maintaining efficient system performance.

Visual Cues and Depth Perception Despite being under constant development during the last decades, current stereoscopic (head-mounted) displays’ capabilities cannot represent visual information faithfully, failing to completely replicate natural viewing conditions. Although displays offer good resolution and pixel density, they still significantly differ from how good our eyes are and cannot match eye acuity.

These differences from real-world viewing conditions are also evident in the absence of correct accommodation cues. In conventional stereoscopic displays, accommo-

dation remains fixed at the physical focal distance of the display, while vergence responses are driven by binocular disparity cues that simulate objects at varying virtual depths [Wann et al. 1995]. As a consequence, vergence and accommodation cues become decoupled, creating an unnatural mismatch that can reduce visual comfort and potentially induce symptoms such as eye strain and fatigue. This phenomenon is known as the Vergence-Accommodation Conflict (VAC) and is discussed further in Section 2.4.2, together with techniques proposed to mitigate it. Another example could be the difference between the rendered visual field, regarding both monocular and binocular (stereoscopic) regions, and our natural field of view (FOV) covered by our eyes. Such a reduction in the rendered visual field limits, especially, peripheral capabilities and features. A schematic comparison between the real FOV of our eyes and the FOV offered by HMDs is shown in Figure 2.9.

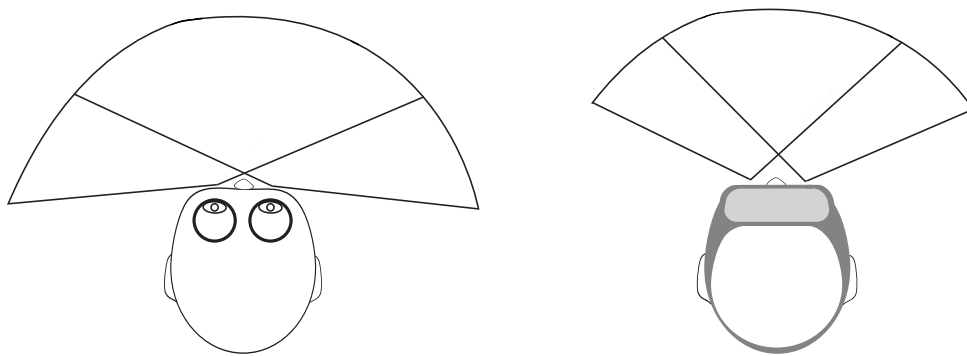


Figure 2.9: Visual comparison among human field of view (FOV), on the left side, and the FOV covered by head-mounted displays, on the right side. Human FOV (*Left*) naturally covers $200^{\circ} - 220^{\circ}$, with a binocular (overlapping) region of 110° [Chang et al. 2020; Stidwill and Fletcher 2017]. In the case of the HMD (*Right*), both monocular and binocular regions are smaller and change between devices.

Another important limitation is color representation, which relates to another major issue of HMD, as is power consumption. Previous work relies on peripheral color perception to reduce the high power requirements of blue-colored pixels without introducing visual artifacts [Duinkharjav et al. 2022b]. Additionally, efforts have been made to include absent visual cues such as motion parallax, which helps to increase realism in 360 video playback [Serrano et al. 2019].

Perceived visual quality and comfort depend not only on static spatial characteristics, such as color fidelity and display resolution, but also on temporal aspects, including refresh rates. In stereoscopic HMDs, the high-resolution imagery required for both eyes must be rendered at sufficiently high refresh rates to avoid visible temporal artifacts, making real-time rendering performance one of the main technological challenges of immersive systems. Modern HMDs typically operate at refresh rates of 90 Hz or higher, up to 180 Hz, becoming especially beneficial for latency-sensitive applications like e-sports or to reduce simulator sickness [Wang et al. 2023c].

Auditory Cues Spatial audio reproduction in immersive systems is commonly achieved through binaural rendering over stereo headphones. In spatial audio rendering, head-related transfer functions (HRTF) are used to model how sound waves are affected by the listener's anatomy, including the shape of the head and ears.

By applying this filtering, virtual sound sources can be better perceived at different spatial locations. Since anatomical characteristics vary across individuals, HRTFs can also be personalized to better match a specific listener's anatomy and improve localization accuracy. In this thesis, generic HRTFs are employed, as they are more widely available and commonly used in consumer applications, facilitating reproducibility and broader applicability of the obtained results. However, considerable effort has been made towards building HRTF datasets [Algazi et al. 2001; Engel et al. 2023] and HRTF modeling, such as synthesizing personalization using anthropometric measurements [Zotkin et al. 2003; Bilinski et al. 2014] or sound projection via spherical harmonics [Schissler et al. 2016].

Integrated Eye-Tracking While eye-tracking (ET) technologies are often integrated in the HMD and allow for tracking the visual behavior, their capabilities are much more limited than desktop-based eye trackers due to size requirements in both refresh rate and accuracy. Due to the nature of the work performed in this thesis, we use high-end headsets equipped with the fastest ET system in the market, offering a refresh rate of 200 Hz and sub-degree accuracy³. In comparison, desktop-based eye trackers can achieve sampling rates up to 1000 Hz, such as the Eyelink 1000 Plus⁴.

2.4.2 Vergence-Accommodation Conflict

As introduced in the previous section, the Vergence-Accommodation Conflict (VAC) has been one major issue in recent years for VR devices. The basis of this conflict is the lack of accommodation cues in the immersive display. As a result, the VAC hinders visual performance and causes visual fatigue [Hoffman et al. 2008].

Different methods have been developed to reduce the impact of the VAC in user experiences [Kramida 2015], from hardware-based solutions like accommodation-free displays or varifocal devices [Maimone et al. 2013] to eye-tracking-based approaches [Kellnhofer et al. 2016a]. Intentional depth compression techniques to alleviate the mismatch between accommodation and vergence have also been studied as a potential solution [Pajak et al. 2014].

2.4.3 Foveated Rendering

Gaze-contingent rendering techniques have been developed to alleviate hardware requirements, such as refresh rate limitations for real-time rendering, while accounting for HVS-specific features [Wang et al. 2023b]. Regarding visual content, decreased visual acuity in peripheral regions reduces the rendering resolution requirements on regions far from the fovea, which allows for reducing computational rendering cost. Several alternatives have been proposed for applying visual foveation, including log-polar transformations to represent resolution variations in a more perceptually suitable space [Meng et al. 2018], neural radiance fields to jointly optimize visual quality and rendering performance [Deng et al. 2022], or AI-driven enhanced gaze tracking solutions to improve the accuracy and robustness of foveated rendering systems [Liu et al. 2025].

³<https://varjo.com/products/xr-4>

⁴<https://www.sr-research.com/eyelink-1000-plus/>

On the other hand, foveation methods not only rely on eccentricity-based resolution decay, but also on other variables such as scene local luminance contrast [Tursun et al. 2019], attention [Krajancich et al. 2023], or eye dominance [Meng et al. 2020], which relies on applying extra foveation levels in the fixed non-dominant eye. However, as we will see in Chapter 5, this solution becomes suboptimal due to the local dominance shifts in peripheral regions.

In these works, computational requirements are reduced according to the HVS limitations, allowing for a shading rate reduction [Patney et al. 2016] or ray-tracing reprojection [Weier et al. 2016]. Nevertheless, foveation techniques go beyond visual perceptual limitations, as some of them also consider audiovisual saliency regions [Shi et al. 2025] or audio sources clustering, which introduces a "foveation" process on the audio modality where relevant sounds rendering is prioritized [Peng et al. 2025].

Chapter 3

Accelerating Visual Saccades Latency

This chapter investigates the temporal and spatial relationship of audiovisual interactions, focusing on how auditory cues that precede visual stimuli influence saccadic latency. Auditory cues are observed to significantly accelerate visual responses, but this effect plateaus beyond certain temporal thresholds. Additionally, while the spatial positioning of visual stimuli influences the speed of these eye movements, we find that the location of auditory cues with respect to their corresponding visual stimulus does not have a comparable effect. Findings are evaluated in two scenarios: a basketball training task and an interactive farm game that explores previously untested values of our key factors.

3.1 Introduction

Human beings continuously interact with the external world via their senses, gathering and integrating information to construct their perception. This cross-modal integration, particularly between sight and hearing, profoundly influences our perception of the world. However, the relationship between these modalities is not always straightforward. For instance, some studies [[Malpica et al. 2022](#); [Hidaka and Ide 2015](#)] have shown that in certain scenarios, auditory stimuli can unexpectedly suppress visual perception, demonstrating the complexity and sometimes counter-intuitive nature of these interactions. In this chapter, we focus on the *spatial* and *temporal* aspects of the relationship between visual and auditory stimuli. In typical real-world situations, visual and auditory cues occur almost simultaneously, or sound is slightly delayed due to differences in the speed of light and sound travel. Existing literature often reflects these scenarios accordingly [[Arnold et al. 2005](#); [Vroomen and Keetels 2010](#)]. In contrast, we benefit from the flexibility of immersive digital environments and virtual reality (VR) setups to investigate a variety of temporal and spatial shifts between visual and auditory stimuli, with an emphasis on sound preceding visual information. This focus is motivated by evidence showing that auditory cues preceding visual stimuli can significantly reduce the latency of saccadic eye movements [[Vidal and Vitu 2022](#); [Ross and Ross 1981](#)]. Saccades are rapid, simultaneous movements of both eyes in the same direction, which are crucial for directing gaze

and attention, impacting reaction times in diverse situations, including gaming and e-sports [Koposov et al. 2020]. A saccade, once triggered, follows a pre-programmed, close-to-ballistic motion whose velocity and duration cannot be easily changed [Bahill et al. 1975a; Arabadzhiyska et al. 2017]. Therefore, we focus on the latency of the preparatory stage when the oculomotor mechanisms are preparing for the saccade launch, which is guided both by visual and auditory cues [Kowler 2011]. In particular, we investigate for the first time how retinal eccentricities and spatio-temporal shifts between visual and auditory stimuli can influence saccadic latency in a virtual environment, as explained in Section 3.3. We conduct our experiments systematically, considering those dimensions in the ranges relevant to VR applications. Our results show that temporal shifts in auditory cues can significantly accelerate saccadic responses, but this acceleration reaches a plateau beyond certain temporal intervals. Moreover, while eccentricity consistently influences saccadic latency, as observed in previous work [Duinkharjav et al. 2022a], spatial shifts in the auditory cues do not seem to significantly impact this latency. These findings not only advance our understanding of cross-modal interactions but also have practical implications across various domains, such as the gaming scenario shown in Figure 3.1.

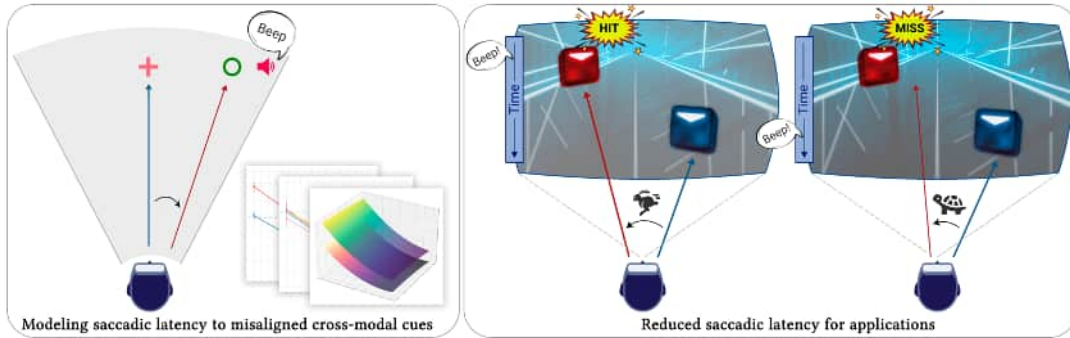


Figure 3.1: *Left*: Simplified diagram of our experiment exploring how the *spatial* and *temporal* relationships between visual and auditory stimuli affect saccadic latency. The results derived from this experiment provide insights into how these factors modulate saccadic latency. *Right*: Illustration of a practical application of our findings in a game similar to *Beat Saber*. In the depicted scenario, user’s ability to react to a target is influenced by the timing of auditory cues. When the sound precedes the visual stimulus, following our insights, the user experiences reduced saccadic latency, enabling them to hit the target on time. Conversely, if the sound and visual stimulus are synchronized, the user’s reaction time is slower, resulting in missing the target.

Reducing saccadic latency through finely tuned audiovisual timings is useful in a variety of applications. For user interfaces, reduced saccadic latency can lead to faster navigation and data retrieval, streamlining workflows. In augmented reality (AR) environments, faster saccadic responses can enhance real-virtual interactions, making information assimilation more intuitive. Within videogames, especially those relying on quick reflexes, our insights can inform the design of cues (visual-auditory timings) that improve players’ response to in-game events, leading to a more engaging and competitive experience [Kim et al. 2019; Koposov et al. 2020]. Digital multimedia, which often relies on carefully sequenced audiovisual information to achieve specific effects, such as the perception of simultaneity or attention-redirecting effects [Serafin et al. 2018; Ogawa et al. 2023], also stands to benefit from these insights. In online multimedia, latencies and delays are common issues that can deteriorate the user

experience, sometimes to the point of discomfort [Hopkins et al. 2022]. By applying our understanding of saccadic latency, these challenges can be mitigated, enhancing the overall experience for users.

In this context, we validate these concepts through two proof-of-concept applications. The first, a basketball training task set in a more realistic environment (Section 3.4.1), whose results demonstrate how saccades can be effectively accelerated, extending our findings beyond simple experimental setups. The second, an interactive farm game (Section 3.4.2), explores the effects of previously untested values of our main factors, further illustrating the versatility of our model.

Our collected anonymized data, model, and code are publicly available at <https://avsaccadeaccel.mpi-inf.mpg.de/>

3.2 Related Work

In this section, we discuss human visual saccades and the factors that affect them, focusing on their importance in graphics applications. Then, we introduce audiovisual spatio-temporal interactions and discuss their connection with the behavior of visual saccades.

3.2.1 Visual Saccades

Saccades, the fastest eye movements, align the fovea with the region of interest for optimal visual acuity. Understanding the dynamics of visual saccades has been beneficial for many graphics applications, informing key areas such as foveated rendering [Albert et al. 2017; Franke et al. 2021], optimization of 3D VR streamings [Chen et al. 2022], or the development of strategies to stimulate the human visual system beyond real-world capabilities [Dunn et al. 2020]. The relationship between saccade amplitude, peak velocity, and duration, typically ranging between 20-80 ms, is well-understood [Bahill et al. 1975a; Arabadzhiyska et al. 2017; Anliker 1976]. However, saccade latency, the time from the appearance of a target stimulus to the execution of the saccade, varies more significantly and is longer, between 200-400 ms [Duinkharjav et al. 2022a]. During this time, a number of tasks must be performed, such as releasing attention from the previous fixation point, sensory registration of a new target, decision-making on active attending, and programming of oculomotor mechanisms for saccade launching [Zambarbieri et al. 1995].

Saccade latency can be influenced by various visual features of the target, such as saliency [Yamagishi and Furukawa 2020], contrast [Carpenter 2004], intensity [Bell et al. 2006], and spatial frequency [Duinkharjav et al. 2022a]. This means that saccade latency can be effectively reduced by enhancing the contrast of potential targets (this effect saturates for strongly suprathreshold contrasts), increasing their intensity, or assuring that spatial patterns include frequencies over 4 cycles-per-degree (while higher frequencies have not been measured following the contrast sensitivity function [Barten 1999], one can assume a latency drop over 10–20 cpd [Duinkharjav et al. 2022a]). Additionally, saccade target direction and eccentricity play roles in latency variation, with horizontal saccades typically being faster than vertical ones [Dafoe et al. 2007], and different effects observed based on the target positioning in the visual field. Lower saccade latency can be expected for the target in the upper hemifield compared to the lower hemifield due to more visual processing in the latter case,

as required for navigation in the environment [Dafoe et al. 2007]. This pattern is complemented by findings that accuracy improvements in saccades are more pronounced in elevation than azimuth, especially in audiovisual contexts [Corneil et al. 2002]. Such improvements are greater when the stimuli include auditory information, highlighting the significant role of sound in enhancing saccadic accuracy. The impact of target eccentricity on saccade latency remains ambiguous. Dafoe et al. [2007] observed no significant changes in latency related to eccentricity, while Zambarbieri et al. [1995] reported an increase in latency as eccentricity rose. Saccades, traditionally studied from a purely visual perspective, have been shown to suppress visual sensitivity [Ross et al. 2001], with visual space compression occurring just before the onset [Ross et al. 1997]. These effects relate to neurophysiological anticipatory procedures and decision processes in saccade generation.

The gap-overlap paradigm, illustrated in Figure 3.2, is a key concept in understanding visual engagement and attention [Pierrot-Deseilligny et al. 1995; Fischer et al. 1997]. It has been widely used to measure changes in saccade latency, relating these to target luminance and anti-saccadic tasks (looking away from the target) [Kingstone and Klein 1993; Reuter-Lorenz et al. 1991]. While this paradigm has been extensively used in visual studies, it has also been applied in purely auditory studies to detect disorders such as tinnitus [Fournier and Hébert 2013; Boyen et al. 2015]. However, its potential relationship with auditory cues in cross-modal conditions has been less explored.

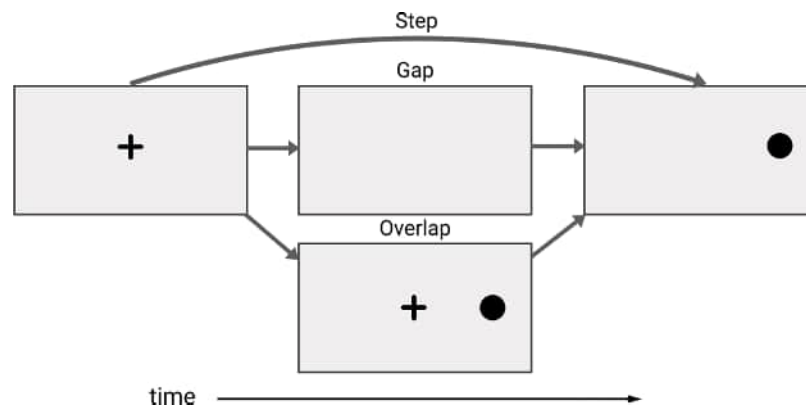


Figure 3.2: Visual gap-overlap Paradigm. This paradigm involves three distinct scenarios. In the *step* case, the target (represented by the filled circle) appears immediately after the offset of the fixation point (denoted by the cross). In the *gap* case, there is a delay between the disappearance of the fixation point and the appearance of the target, which typically results in faster saccade responses. Conversely, in the *overlap* case, the target appears before the fixation point disappears, leading to slower saccades. The gap scenario facilitates quicker saccades by reducing attention engagement at the fixation point [Kingstone and Klein 1993], while the overlap condition holds attention longer on the fixation point, requiring active disengagement and thus delaying the saccade initiation [Kalesnykas and Hallett 1987].

3.2.2 Audiovisual Interactions

The perceptual interaction between visual and auditory signals is a topic of extensive research, with findings indicating both facilitatory and inhibitory effects. The exact nature of this audiovisual interaction is complex, with some studies showing that

auditory signals can even cause visual suppression under certain conditions [Hidaka and Ide 2015; Malpica et al. 2022]. Conversely, facilitatory effects have also been found for this audiovisual integration. For instance, reaction times to bimodal (audiovisual) stimuli are often faster than to unimodal stimuli [Teder-Sälejärvi et al. 2005], and phenomena such as the McGurk effect (blending between non-consistent audio and visual cues in speech) demonstrate the powerful illusory potential of audiovisual integration. Multisensory interaction is the physiological mechanism through which two different sensory stimuli can affect each other and overall perception. Focusing on the audiovisual modality, different factors have been found to affect this multisensory binding [Spence and Driver 2004], including the temporal window [Vatakis and Spence 2007; Chen and Spence 2017], and the meaning and content of stimuli for semantic congruency perception [Doehrmann and Naumer 2008].

Regarding visual saccades, the limits on the advantages of including the auditory modality remain unclear. Audiovisual integration has been observed to be beneficial in reducing visual saccade latencies [Wang et al. 2017; Diederich et al. 2012; Vidal et al. 2020] and improving overall performance. Thus, auditory cues have been shown to alter how saccades are performed [Zou et al. 2012]. Research has shown that auditory cues aligned with visual signals in time and space can reduce saccadic latency. Frens et al. [1995] observed that, while this latency reduction diminishes with spatial misalignments, the increase in latency is relatively minor, maintaining stability even with misalignments of approximately 20° . They also observed that when auditory cues lag behind visual signals, there is an increase in saccadic latency. Nevertheless, it has been shown that behavioral responses to bimodal stimuli generally outperform those to unimodal stimuli, both in speed and accuracy, regardless of the audiovisual pairing locations [Teder-Sälejärvi et al. 2005; Corneil et al. 2002; Van Wanrooij et al. 2009].

Closest to our approach is Vidal and Vitu [2022], where visual latency is modulated by manipulating the audiovisual delay (temporal shift) and the temporal interval between fixation offset and target onset (gap-overlap paradigm). As highlighted in their findings, combining these methods is crucial for achieving maximum modulation of saccades. Our research takes this a step further by providing a comprehensive analysis of visual saccade latency, including both spatial and temporal shifts between audiovisual cues. Our contributions extend beyond this analysis, as we also validate our findings in more complex scenarios, thus bringing our insights closer to practical applications. Additionally, we conduct our tests in VR environments, incorporating stereo and spatial sound and offering a more immersive and realistic context for understanding the facilitative role of auditory information in saccadic movements.

3.3 Measuring Saccadic Latency Under Cross-modal Misalignments

In this section, we describe our main experiment, which focuses on measuring saccadic latency under various spatio-temporal misalignments of audiovisual cues. We describe our experimental setup and methodology in detail (summarized in Figure 3.3), followed by an in-depth analysis of the results.

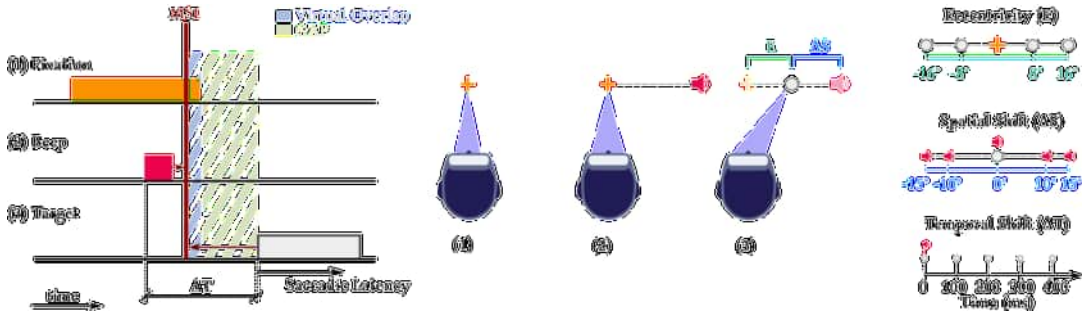


Figure 3.3: *Left*: Temporal Dynamics in the Experiment: The Multisensory Integration (MSI) process, influenced by the temporal shift ΔT , alters the perception of the target onset to be closer to the beep onset, thereby effectively integrating these two elements into a single event. This creates a perceived virtual overlap of the fixation and the visual target, which is effectively compensated with a real gap as inspired by [Vidal and Vitu \[2022\]](#). *Center*: User Study Task. Participants initially fixate on the central cross (1). When they confirm readiness, the experiment starts. After a random time to avoid learning effects, the beep sounds (2), followed by the visual target apparition (3) after the ΔT interval. *Right*: Visual representation of the stimuli cases with different levels of eccentricity (E), spatial shift (ΔS), and temporal shift (ΔT).

3.3.1 Experimental Setup

Hardware We used a Varjo Aero HMD and Beyerdynamic DT 770M headphones with passive noise cancellation. This HMD is equipped with an eye-tracker running at 200 Hz and with an accuracy of 1° of visual angle. Additional hardware details are shown in [Section 7.1, Appendix A](#). Audio rendering was handled by the Oculus Native Spatializer plugin (current Meta XR Audio SDK⁵), with sound spatial propagation being reproduced using a generic head-related transfer function (HRTF). Existing research supports the notion that generic HRTFs provide satisfactory azimuth localization accuracy within our spatial shift range, compared to individualized HRTFs [[Berger et al. 2018](#); [Wenzel et al. 1993](#); [Rummukainen et al. 2019](#)]. Therefore, we believe these minor variations in spatial perception are unlikely to impact our results, while also simplifying the experimental procedure. Participants remained seated and were free to move their heads. Notably, all studies and virtual environments referred to in this dissertation were designed using Unity⁶.

In our pipeline, two main measurement error sources exist: visual stimulus latency (max 11.11 ms at 90 Hz) due to Unity’s render to display delay, and eye-tracker logging latency (6-7 ms), from eye-tracker sampling (max 5 ms at 200 Hz) plus Unity processing. Combined, these contribute to a potential 18.11 ms error. Wired headphones introduce 1-10 ms latency, potentially creating a 10 ms visual-auditory mismatch. However, maximum delays are rare across trials, reducing their overall impact despite occasional occurrences. Highly specialized equipment like high-speed cameras and oscilloscopes would be needed for precisely tracking this error. Our GLMM analysis effectively manages random variations across subjects and trials. Although this specific error isn’t explicitly modeled as a factor, it’s included within the trial variability we account for, thus expected to have minimal impact on systematic bias in our results. These values were also confirmed by the headset manufacturer.

⁵<https://developers.meta.com/horizon/documentation/native/audio-intro/>

⁶<https://unity.com/>

Stimuli The target was a white circle of 0.36° placed at 6 meters from the participant, with a lifetime of 2 s [Vidal and Vitu 2022]. The Weber contrast was set to 0.3, above detection thresholds, such that it would not affect saccade latency [Duinkharjav et al. 2022a]. The fixation cross was of size 0.12° , located in the middle of the screen [Vidal and Vitu 2022]. These are illustrated in Figure 3.4. The auditory cues were beep-like sounds, centered at 880 Hz and 60 dB [Zambarbieri et al. 1995; Zambarbieri 2002] that stop after 150 ms. We employ the *beep and flash* scenario, a simple experimental setup involving auditory beeps and visual flashes, for its ease of control and limited cognitive interference. This scenario has been shown to tap into fundamental neurophysiological processes related to temporal resolution and integration in the brain, which are relevant across various types of more complex tasks [Shams et al. 2000]. Additionally, this scenario is particularly relevant in practical applications such as gaming environments where quick perception of simple, synchronized audiovisual cues is crucial.

Simple stimuli are widely used to establish foundational understandings of sensory processing. They are highly controllable and are less likely to be influenced by higher cognitive processes than more complex stimuli like speech reading, which engage specific language processing areas and require the integration of more complex, meaningful information. Therefore, simple stimuli are thought to engage basic sensory integration mechanisms that are more universally present across different individuals and contexts, making findings more generalizable across a wider range of situations and populations [Stein and Stanford 2008]. The *beep and flash* setup is thought to tap into fundamental neurophysiological processes related to temporal resolution and integration in the brain, which are relevant across various types of more complex tasks [Shams et al. 2000]. Simple auditory and visual stimuli can be precisely controlled and manipulated in virtual settings, allowing for a clearer interpretation of results and the underlying processes. *Beep and flash* was selected due to its potential applicability to the audiovisual cues prevalent in interactive gaming environments. These environments often rely on the rapid perception of simple, meaningless synchronized auditory and visual signals to inform user actions and reactions. Such stimuli are analogous to the *beep and flash*, providing an experimentally sound basis that is highly relevant to the temporal processing demands encountered in gaming applications. This approach allows us to explore the fundamental neurophysiological mechanisms of multisensory integration with direct implications for enhancing user experience in virtual settings.

Nevertheless, further research using more complex, naturalistic tasks is necessary to fully understand how these principles extend to real-world multisensory experiences. Our study can be positioned as a crucial step in building toward a more comprehensive understanding of temporal integration across a spectrum of stimulus complexities. In Section 3.4, where the application cases are proposed, a more realistic basketball scene is tested to overcome *beep and flash* limitations.

Conditions The variables considered in this study were visual *eccentricity* (E), *temporal shift* (ΔT), and *spatial shift* (ΔS) between the visual target and its corresponding auditory cue. For eccentricity, we tested 8 and 16 visual degrees in both left (-) and right (+) peripheries ($E = [-16^\circ, -8^\circ, 8^\circ, 16^\circ]$), aligning with common saccade displacements [Bahill et al. 1975a]. The spatial shifts were $\Delta S = [-15^\circ, -10^\circ, 0^\circ$ (co-located), $10^\circ, 15^\circ]$ visual degrees, considering the simultaneity fusion window limits based on eccentricity levels [Godfroy et al. 2003].

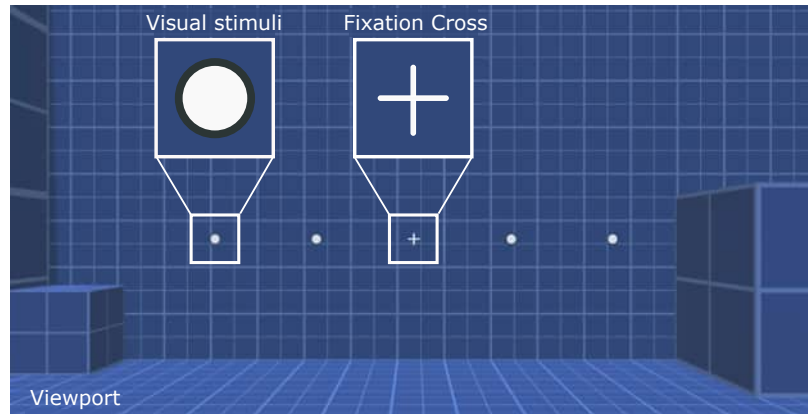


Figure 3.4: Illustration of the experimental view from the participant's perspective: Central to the view is a fixation cross with a 0.12° diameter, surrounded by visual targets measuring 0.36° in diameter, positioned at eccentricities of $E=[-16^\circ, -8^\circ, 8^\circ, 16^\circ]$. These elements are visually enlarged for visualization purposes. The purpose of the surrounding virtual environment is to provide some visual cues for depth perception and a higher feeling of immersion (adapted from *Speed Tutor in Unity Asset Store*).

For temporal shifts, we considered $\Delta T=[0$ (no delay), 100, 200, 300, 400] ms, accounting for the temporal binding window limits in literature [Spence 2011], with audio cues always generated before the visual target onset. We selected these values to achieve effective audiovisual integration, taking relevant literature into account.

According to literature, audiovisual cues are perceived as temporally simultaneous as long as the temporal shift (ΔT) between their onset times is located within the limits of the Temporal Binding Window (TBW) [Stevenson and Wallace 2013]. This TBW has been reported to vary between individuals [Stevenson et al. 2012], time instants due to recalibration procedures and training [Powers et al. 2009] [Van der Burg et al. 2013], and the task asked to the participants [García-Pérez and Alcalá-Quintana 2015].

This temporal shift tolerated between visual and auditory cues such that the stimuli are recognized as a single event has been measured for different tasks under diverse conditions. Several works have addressed temporal audiovisual integration in several speech-related tasks and object-based videos [Vatakis and Spence 2006], reporting an asymmetric TBW when this audiovisual integration keeps taking place. It is also presented how this TBW is not centered at the zero delay point ($\Delta T=0$), but it is a bit shifted towards the visual leading part instead. However, it is not difficult to tell how these results can be biased by our prior knowledge. It is expected in normal conditions, and even more in human speech, to perceive the visual cue before the auditory effect: our experience tells us that it is not common to hear the human voice before seeing the lips' motion.

In VR, tolerance limits to audiovisual temporal binding are studied and its effect on presence and immersion feelings [Kim and Lee 2022]. The average minimum temporal shift noticed by participants in their setup was reported to be 320 ms. A similar limit value is reported in [Hopkins et al. 2022] as the temporal threshold from which the temporal shift starts to be perceivable, thus the audiovisual musical cues were not perceived as simultaneous anymore. Based on these studies, we set our maximum ΔT to 400 ms. These factors are depicted in Figure 3.3 (right).

We considered a fully factorial design resulting in $4 (E) \times 5 (\Delta S) \times 5 (\Delta T) = 100$ conditions. We additionally included four baseline cases without auditory cues (one for each eccentricity). Each participant was presented with each condition four times, totaling 416 trials.

Participants The study was conducted with twelve participants (six male, six female, none non-binary; ages 22-26). All reported normal or corrected-to-normal vision and hearing. From the participant pool, 58.3% reported having no experience with VR before, while 41.7% reported having used VR equipment 5 times or less. For every user study performed throughout this thesis, participants provided written consent, and study protocols were approved by the Ethical Review Board of the Department of Computer Science at Saarland University (Application No. 23-11-3).

Procedure The experiment was conducted in four sessions of ten minutes each, to prevent fatigue. The order of trials was randomized across sessions. For each session, the eye-tracker was calibrated following a five-dot calibration. In each trial, participants started by looking at the fixation cross located in the middle of the screen (starting point). Afterward, participants pressed the *space* on the keyboard to indicate that they were ready. After some random time (between 750 and 1750 ms), used to avoid learning and prediction effects, the fixation cross disappeared and audiovisual cues were generated according to the corresponding E , ΔS and ΔT in the trial. Participants were asked to look at the visual target (make a saccade towards it) as soon as they perceived it. After some time (2 s), the visual target disappeared and the fixation cross spawned again, starting a new trial. This procedure is represented in [Figure 3.3](#) (left and center).

3.3.2 Results and Discussion

Gaze Data processing We provide the methodology to detect visual saccades and the procedure followed to discard spurious cases. We also show in [Figure 3.5](#) a sample of all recorded saccades for one of our participants to illustrate the saccadic profiles obtained during our experiment.

Saccades are obtained for all sessions using the algorithm explained in [[Arabadzhiyska et al. 2017](#)] ([Section 3.2](#)). Velocity thresholds used are both theoretically and empirically tuned. Saccades are detected using a threshold of $120^\circ/s$, which is enough to detect the typically surpassed value of $100^\circ/s$ [[Arabadzhiyska et al. 2017](#)]. To detect the visual saccade starting point (anchor threshold) and the saccade end (fixation threshold), a conservative value of $60^\circ/s$ is selected so fixation movements are not included [[Boghen et al. 1974](#)]. Once all the saccades are detected, the first detected saccade after each target onset is identified and checked to be a correct user response by comparing the target's real eccentricity location and the saccade landing point. Saccades are discarded if they were performed too early, the starting or ending points could not be detected in the specified range, or gaze tracking was lost during the saccade. Following previous literature [[Ross and Ross 1981](#)] [[Frens et al. 1995](#)] [[Vidal and Vitu 2022](#)] we filter out saccade latencies lower than 100 ms, associated with predicting cases, and higher than 500 ms, corresponding to clear error cases, for further analysis. Saccades detected during one experiment session can be found in [Figure 3.5](#). Undershoot saccades are observed for closer eccentricities [[Becker and Fuchs 1969](#)] while eye-tracker limitations can also be observed for higher eccentricity values.

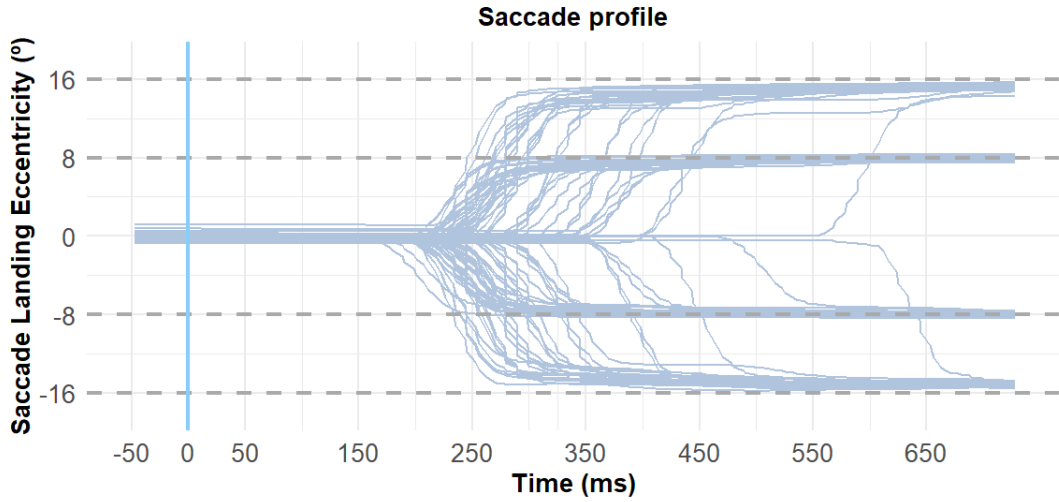


Figure 3.5: All saccades detected and measured in a single experiment session. At $t=0$, the visual target spawns, starting the latency measurement. Different saccade onset times are appreciated thus varying the final measured latency. Different landing eccentricities for each E condition are also observed, finding natural undershoot for a closer horizontal distance and also eye-tracker limitations for larger eccentricities.

Analysis We used a two-step velocity threshold method to accurately identify the start (onset) and end (offset) of the saccades, accounting for eye-tracking noise [Arabadzhiyska et al. 2017]. Conservative thresholds of $120^\circ/s$ and $60^\circ/s$ were set to detect the saccade and its onset. We compute the *saccadic latency* as the difference between the target stimulus onset and the saccade onset. Following previous literature [Ross and Ross 1981; Frens et al. 1995; Vidal and Vitu 2022], we filter out saccade latencies lower than 100 ms (prediction cases), and higher than 500 ms (distraction cases). For computing the *landing error* of the saccade, we compute the distance in degrees of visual angle between the landing point (related to the saccade offset) and the location of the target [Arabadzhiyska et al. 2017]. Given the eye-tracker’s accuracy of 1° , the target size of 0.36° , and typical saccadic undershoots ranging from 10% to 20% [Kapoula and Robinson 1986; Lisi et al. 2019], deviations up to 4.2° for the largest saccades are expected. Consequently, we discard saccades with a landing error exceeding 5° to filter out inaccuracies, such as those from distracted participants.

We combine positive and negative eccentricities (merging -8 with 8 and -16 with 16) after verifying their similarity and in line with previous studies that consider symmetry [Charbonneau et al. 2013; Duinkharjav et al. 2022a], resulting in two eccentricity levels for the analysis: 8 and 16. We use a generalized linear mixed model (GLMM), as it offers a flexible and robust framework for analyzing data in the presence of random effects [Bolker et al. 2009]. Our model includes three fixed effects: *temporal shift* (ΔT), *eccentricity* (E), and *spatial shift* (ΔS), as well as their interactions. We consider the participant as a random effect. The dependent variables are the *saccadic latency* and the *landing error*. We establish the significance level at $\alpha = 0.05$. After applying the exclusion criteria outlined above, each unique combination of main factors appears 32 to 48 times (4 trials $\times$ 12 users) in our analysis, with trial count variations effectively managed by the GLMM. The main results of the analysis are discussed in the following and depicted in Figure 3.6. All the statistical tests and post-hocs can be consulted in Section 7.2, Appendix B.

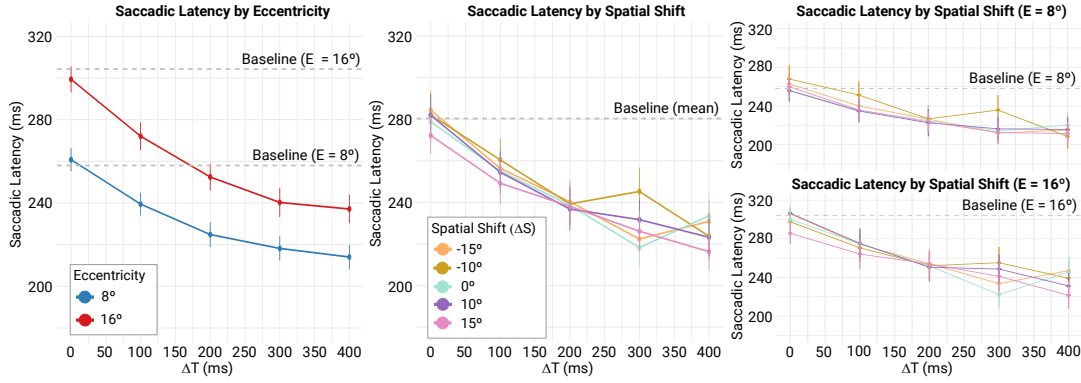


Figure 3.6: Saccadic latency in the main experiment for different temporal shifts depending on eccentricity (*left*), spatial shifts (*center*), and their interaction (*right*). Baseline cases (no sound) are represented with dashed lines and error bars indicate 95% CI.

Temporal shift (ΔT) The effect of ΔT is statistically significant ($p < 0.001$), indicating that varying ΔT levels effectively modulate saccade latency. When ΔT increases, saccade latency decreases. Post-hoc analysis shows no significant difference between ΔT intervals 200-300 ms ($p = 0.09$) and 300-400 ms ($p = 0.44$). This suggests a trend towards stabilization and a potential upper limit in saccade acceleration at higher ΔT levels. This finding overturns the assumption that earlier auditory cues always lead to proportionally faster saccades, suggesting a more complex relationship where the efficiency of auditory cues plateaus beyond a certain point. These results are in line with conclusions reported in [Vidal and Vitu \[2022\]](#). They reported a decrease in latency for $\Delta T = 0$ compared to the no-sound case, and latencies were found to be shorter for beeps preceding the target onset $\Delta T = 60, 120$ ms with gap compensation, showing how the modulation effect was larger with gap compensation compared to not compensated cases, reducing latency even more. [Frens et al. \[1995\]](#) (experiment 3) also reported how ΔT can affect the saccadic latency distribution. Using different ΔT cases ($-100, -50, 0, 50$), the lowest latency was found when the auditory preceded the visual $\Delta T = 50$ instead of when they were synchronous, justifying this on different speeds generating different arrival moments in receptor organs. Different onset and offset conditions for visual or auditory warning signals were tested by [Ross and Ross \[1981\]](#) (experiment 1), changing also ΔT to trigger saccades. For the auditory warning, they found that both onset and offset warnings could facilitate saccade latency at $\Delta T = 100$ and 300 ms but not when the audio onset was after the target onset.

Eccentricity (E) This factor significantly influences latency ($p < 0.001$), with larger eccentricities showing increased latency. Interestingly, our results reveal that this effect appears to be uniform across all ΔT levels, reinforcing and complementing previous studies. This suggests that the spatial positioning of visual stimuli plays a crucial role in saccadic responses, regardless of auditory cue timing, emphasizing the importance of visual cue placement in multisensory integration. Our results align with prior research. [Duinkharjav et al. \[2022a\]](#) report a similar trend in their preliminary study, where they varied eccentricity (with frequency and contrast) to compare saccade latency durations towards a Gabor patch stimulus. They observed that the latency-eccentricity relationship exhibited a U-shaped curve, with latency being higher at both the foveal (0°) and peripheral (20°) regions, and lower at a mid-

peripheral point (10°). The same bowl shape was found for the latency-eccentricity function by [Kalesnykas and Hallett \[1994\]](#). They measured 38 different retinal eccentricities and observed how color and intensity have small contributions to the central peak (5-15 ms). Their central robust minimum latency peak ranged from 0.75° to 12° (our 8°) and reported a height of 35-75 ms. After that, latency gradually increases towards the periphery (our 16°), becoming larger and more erratic for the largest periphery values (around 50° - 60°).

Spatial shift (ΔS) Our data indicates no significant effect of this factor on saccadic latency ($p = 0.18$). However, it is important to interpret this with caution, as the absence of a statistically significant impact does not definitively rule out an effect. This finding suggests that if there is an influence of precise sound localization (co-location) on saccadic latency, it may not be consistent or robust across different scenarios. This could have significant implications for how auditory cues are used in various applications, allowing for more flexibility in sound design.

Our findings align with those of [Frens et al. \[1995\]](#), who studied the effects of vertical, horizontal, and diagonal misalignments from the central fixation point. They found that spatially coincident visual and auditory stimuli reduced visual latency by approximately 50 ms compared to visual-only scenarios. However, this latency reduction diminished with increasing spatial shifts (completely vanishing at large misalignments of about 54°), and was very subtle for small shifts like those we considered. Similarly, [Van Wanrooij et al. \[2009\]](#) reported that spatially aligned auditory events produced shorter reaction times, but these effects lessened with greater spatial shifts, leading to a breakdown in multisensory integration and resulting in bistable saccades. Lastly, [Colonius and Arndt \[2001\]](#) reported how audio cues indeed help to reduce saccadic latencies towards visual targets up to large spatial shifts. Without temporal shift ($\Delta T = 0$), there were almost no latency differences for $\Delta S = +/ -15$, while very small differences were reported again with $\Delta T = 30$. Our results, showing no significant impact of spatial shifts within our explored range on saccadic latency, are consistent with the subtle differences observed in other studies for comparable spatial misalignments. This complements our approach of restricting spatial shift values to ensure effective multisensory integration and maintain stimuli within the field of view, essential for audiovisual integration applications.

Interaction of ΔT and E This is the only significant interaction ($p < 0.001$). Our results suggest that smaller eccentricities ($E = 8^\circ$) reach a point of stabilization and maximum acceleration at an earlier ΔT compared to larger eccentricities ($E = 16^\circ$). Specifically, for $E = 8^\circ$, the differences in latency become smaller and statistically non-significant starting from an ΔT of 200 ms. For $E = 16^\circ$, a similar pattern emerges only from a ΔT of 300 ms, as the interval between ΔT 200-300 ms still shows significant differences. This finding, particularly the stabilization points for different eccentricities, provides valuable insights for optimizing multisensory cues in environments where both spatial and temporal factors are critical.

Saccades accuracy Landing errors are shown in [Figure 3.7](#). Although eccentricity variations were significant ($p < 0.001$), differences might be partly attributed to the eye-tracker resolution, which is less accurate at larger eccentricities [[Stahl 2001](#)]. Despite this, all the observed errors consistently remained under 1° , which is roughly the accuracy of the eye-tracker. In the context of audiovisual saccades, previous

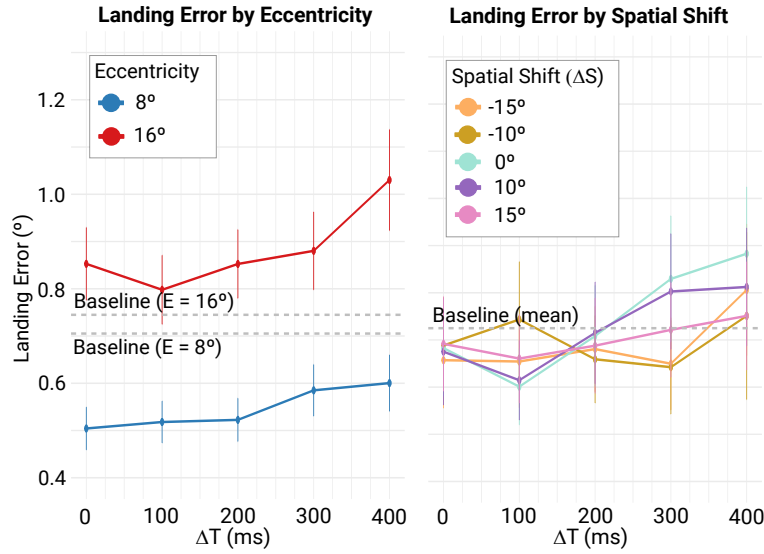


Figure 3.7: Saccade landing error in the main experiment for different temporal shifts depending on eccentricity (*left*) and spatial shifts (*right*). Baseline cases (no sound) are represented with dashed lines, and error bars indicate 95% CI. The increased error at larger eccentricities can be partly due to reduced eye-tracker accuracy, yet all errors stay below 1° .

research by [Corneil et al. \[2002\]](#) suggests that these are generally faster and more precise than saccades driven solely by visual or auditory stimuli. Our study did not record bistable unimodal responses, which are typically absent with small perceived ΔS where multisensory integration (MSI) occurs. Consequently, given the small ΔS levels employed in our experiment, we did not anticipate observing notable effects on saccade accuracy.

3.4 Applications

In this section, we present two main applications of our study. First, we validate our findings through a realistic use case, linking experimental results to a practical scenario. Second, we introduce a more interactive validation using a game-like experience that tests previously unmeasured values of our main factors, demonstrating the model's broader applicability. Finally, we discuss other potential applications of our insights and model.

3.4.1 Practical Application: Basketball Game

Our findings regarding latency acceleration are validated in a scenario that features audiovisual stimuli with enhanced semantic content, closer to real applications than the *beep and flash* paradigm used in our main experiment.

Scene and stimuli conditions The designed task simulates a basketball training application ([Figure 3.8](#)). Participants were located in the middle of a basketball court, looking at one basket. At each side of the basket, two teammates were looking towards the participant. Behind them, four referees were placed at different locations.

As visual stimuli, basketballs of a size of 3° appeared in front of the teammates as they were holding them. For simplicity and to avoid fatigue, a single eccentricity level was considered in both directions $E=[-8^\circ, 8^\circ]$. Auditory cues were a whistling noise of 60 dB, generated by the referees with two possible spatial shifts $\Delta S=[0^\circ$ (co-located), $15^\circ]$. A red cross in the middle of the whiteboard in the basket pole was used as a starting fixation point. Audiovisual temporal shifts were again $\Delta T=[0$ (no delay), 100, 200, 300, 400] ms (audio-leading). This resulted in $2 (E) \times 2 (\Delta S) \times 5 (\Delta T) = 20$ conditions plus 2 baseline conditions with no sound. Each condition was experienced 10 times, resulting in a total number of 220 trials per user.

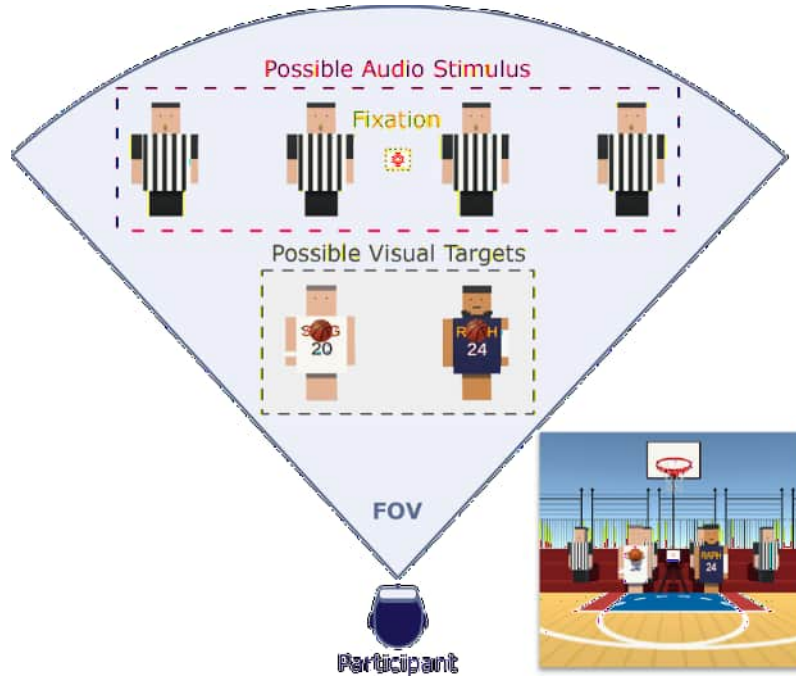


Figure 3.8: Illustration of our basketball training application. The fixation point is a red cross located under the basket in the middle of the whiteboard. The visual target is a basketball being held by teammates, while the auditory cue is a whistle-blowing generated by the referees on the back. The basketball court and basketball assets are obtained from the *Unity Asset Store*.

Participants and procedure A total of seven participants who did not participate in the main experiment carried out this study (six male, one female, non non-binary; ages 24-29). All reported being experienced using VR. The procedure was identical to that of the main experiment, except for the scene and the audiovisual cues. The task of the participants was to react as quickly as possible to the ball appearing and gaze towards it (perform a saccade).

Results After merging eccentricities according to horizontal symmetry [Charbonneau et al. 2013], we obtain a total of 140 samples per condition. Following the same data processing and analysis as in the main experiment (details in Section 7.2, Appendix B), results are shown in Figure 3.9. In summary, results in this more realistic scenario confirm our key findings from the main experiment: the persistence of the effect of ΔT on latency acceleration with consistent modulation of latency with

temporal shifts ($p < 0.001$), and the potential upper limit in acceleration efficiency at higher temporal shifts. Furthermore, results for the tested eccentricity case ($E = 8^\circ$) align with the main study, demonstrating substantial latency acceleration with temporal shifts, whereas spatial shift (ΔS) remains non-significant ($p = 0.08$). Saccade accuracy, quantified through landing error, exhibits no significant differences ($p = 0.91$ for ΔT and $p = 0.28$ for ΔS), maintaining errors around 1° , consistent with eye-tracker accuracy. These results confirm the robustness and generalizability of our findings across diverse scenarios.

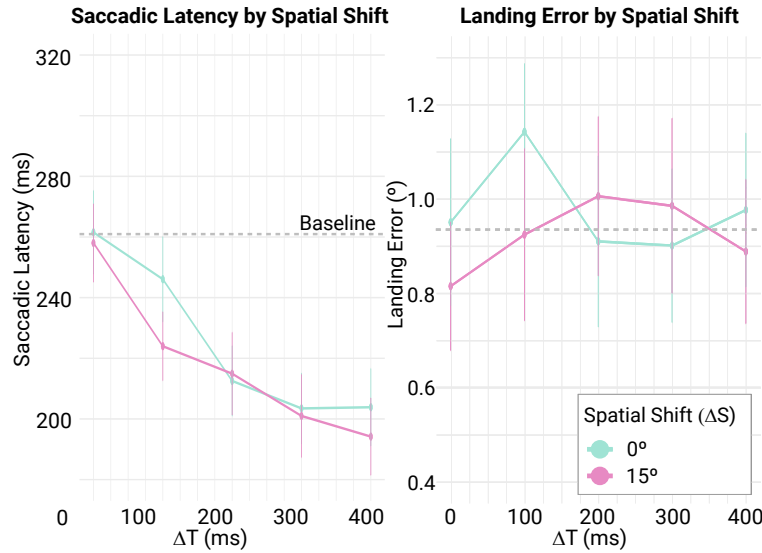


Figure 3.9: Saccadic latency (*left*) and landing error (*right*) in the basketball application for different temporal shifts depending on the spatial shifts. Baseline cases (no sound) are represented with dashed lines, and error bars indicate 95% CI.

3.4.2 Interactive Application: Farm Game

Our study progresses to a more interactive validation with a virtual farm game, testing our model in new conditions. For this purpose, we first fit a model using eccentricity, spatial shift, and temporal shift as inputs and saccadic latency as the output. Choosing a polynomial based on the main experiment results, we perform a 10-fold cross-validation and select a 2-degree polynomial with an R-squared value of 0.95. The model is illustrated in Figure 3.10 (left). The model equation can be found in Equation (3.1), with details on each factor meaning below.

$$\begin{aligned} \text{Latency} = & (0.008) * E + (-0.174) * \Delta S + (-0.206) * \Delta T + (0.198) * E^2 \\ & + (-0.004) * E * \Delta S + (-0.006) * E * \Delta T + (-0.009) * \Delta S^2 + 249.197 \end{aligned} \quad (3.1)$$

1. Squared Terms

- (a) E^2 : The positive coefficient (0.198) indicates that as Eccentricity increases, its effect on *Latency* increases in a quadratic manner.
- (b) ΔS^2 and ΔT^2 : These terms have very small or zero coefficients, suggesting that their quadratic relationships with *Latency* are negligible.

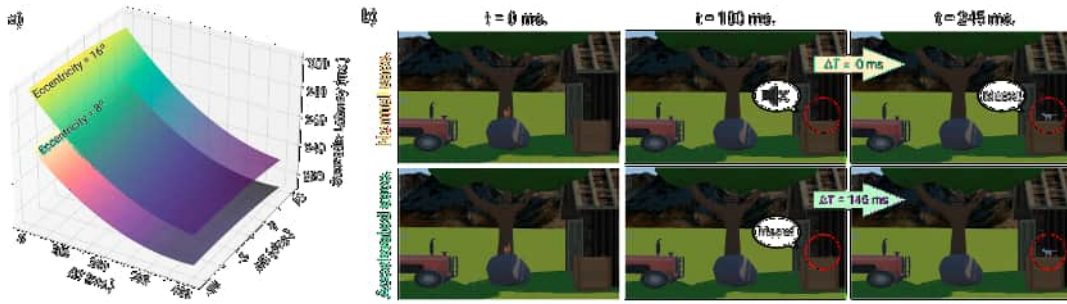


Figure 3.10: Latency acceleration in the interactive farm game. *Left*: Model fitted to our data as described in the text. Two representative eccentricities are depicted. *Right*: Model applied to our interactive farm game to influence saccade latency. In normal conditions (top), the sound and the visual event are presented simultaneously. In the accelerated response condition using our model (bottom), a temporal shift is employed to play the sound associated with a key game event before the visual presentation. As a result, differences in saccade latency are observed, achieving the acceleration as predicted by our model and prompting a faster saccadic response.

2. Linear Terms

- (a) E : The coefficient 0.008 suggests that a one-unit increase in Eccentricity is associated with an increase of 0.008 units in *Latency*, all else being equal.
- (b) ΔS : The coefficient -0.174 indicates that a one-unit increase in ΔS is associated with a decrease of 0.174 units in *Latency*.
- (c) ΔT : Similarly, the -0.206 coefficient for ΔT suggests that increasing ΔT by one unit decreases *Latency* by 0.206 units.

3. Interaction Terms

- (a) $E - \Delta S$, $E - \Delta T$ and $\Delta S - \Delta T$: These coefficients represent how the interaction between these variables affects *Latency*. For example, a coefficient of -0.004 for Eccentricity Disparity implies a combined effect on *Latency* when both Eccentricity and Disparity vary together. However, such small coefficients suggest that these interactions minimally impact *Latency*.

- 4. Intercept (249.197): This is the predicted value of *Latency* when all independent variables are 0.

In summary, the model suggests that *Latency* is primarily influenced by a significantly strong effect of Eccentricity (E^2), followed by linear terms of Eccentricity, Spatial shift (ΔS), and Temporal shift (ΔT).

Scene and stimuli conditions Participants' task was to protect a rooster (fixation point) from approaching cats (visual targets) at new eccentricities ($E = 10^\circ, 12^\circ, 14^\circ$), with co-located auditory cues (meows, $\Delta S = 0$), as shown in Figure 3.10 (right). Different temporal shifts (ΔT) were computed using our model with the goal of accelerating saccades by 20, 40, and 60 ms. The game featured a timing bar that filled if participants failed to maintain gaze on the rooster, with the rooster escaping if the bar filled completely. Players used a controller to interact with and deter the cats, creating a dynamic gaming environment typical of action sequences.

Participants and procedure Seven new participants (five male, two female, ages 22-34) were involved in the study, following the same procedures for eye-tracking calibration and trials as in the main experiment and the basketball application.

Results The average error across conditions between the measured latencies and those predicted by our model is 8 ms, confirming the robustness of our model. The more interactive dimension of this application corroborates the model's consistency and practical relevance, as similarly demonstrated in the basketball scenario. Results obtained for each condition considered in the interactive farm game can be found in [Table 3.1](#).

Table 3.1: Results of our farm game application. For each eccentricity, we test two target saccadic latency accelerations (no acceleration vs. 20, 40, 60 ms). Based on the desired saccadic latency, we estimate with our model the ΔT that needs to be applied. Then, we show the predicted latency by our model for those conditions and the measured latency (mean $\pm$ standard deviation) in the study. Predicted latency by our model is very close to that measured in all cases. The average error across conditions is 8 ms.

Eccentricity (°)	Target acceleration (ms)	Applied ΔT (ms)	Predicted latency (ms)	Measured latency (ms)
10	0	0	269	268 $\pm$ 10
10	20	75	249	242 $\pm$ 11
12	0	0	277	270 $\pm$ 9
12	40	145	237	222 $\pm$ 7
14	0	0	288	281 $\pm$ 9
14	60	135	228	214 $\pm$ 4

3.4.3 Other Applications

In this section, we discuss potential scenarios in which our insights could offer valuable guidelines. Our two applications demonstrate the direct feasibility of our approach as content design guidelines for triggering faster saccadic responses and overall reaction times, suggesting adaptability across various scenarios. For instance, in AR-based navigation, auditory cues can precede visual markers, reducing cognitive load and enhancing quick visual localization and responses. Furthermore, regarding high-stakes user interface design, such as air traffic control or complex data visualization, multisensory integration principles can optimize alert timing by design, thus aiding natural, rapid focus shifts to critical information without overwhelming users.

Additionally, the feasibility of audio-visual decoupling, demonstrated in platforms like Teams and Zoom, suggests potential adjustments for varying network latencies in live streaming and esports, facilitating synchronous experiences in competitive gameplay. While the reported model may not be immediately deployable in these scenarios, further research could refine and adapt these concepts for practical applications.

3.5 Limitations and Future Work

While we selected the ranges of our factors based on previous literature and with the goal of maintaining multisensory integration [Vidal and Vitu 2022; Hopkins et al. 2022; Godfroy et al. 2003], larger ranges could be measured. Nevertheless, we have already observed how acceleration stabilizes within the 300-400 ms delay range. Further, the ranges of temporal shifts we measure already surpass the simultaneity perception threshold [Kim and Lee 2022; Hopkins et al. 2022]. Expanding these delays further could disrupt multisensory integration, causing the visual and auditory cues to be perceived as separate, isolated events, with the sound acting as a distractor [Berti and Schröger 2001; Tellinghuisen and Nowak 2003]. Similarly, larger spatial shifts could lead to the same issue, resulting in even visual suppression effects [Malpica et al. 2022; Hidaka and Ide 2015].

Depth perception in virtual scenarios remains an important area of research, particularly in the context of the vergence-accommodation conflict. We employed a single depth value for all stimuli, but different distances may influence sound perception and subsequent saccade latency. Additionally, variations in depth regarding audiovisual mismatch, which we only considered horizontally, could yield different results. The influence of saccades on eye accommodation and binocular vision also remains to be further explored. Previous research suggests that saccades can accelerate eye focusing [Schor et al. 1999] and disparity vergence through neural mechanisms [Enright 1984, 1986; Duinkharjav et al. 2023]. Investigating how saccade latency reduction affects these processes in the presence of temporally shifted audio could provide valuable insights.

In our experiment, we consider the *beep and flash* paradigm. Additionally, we validate our insights in a more complex scene with semantic audiovisual signals. However, future work could further investigate semantic correspondences between audio and visual cues (speech, playing an instrument, object-based videos), as well as explore how participants' emotional states, induced through various means, might influence the saccade acceleration process.

In future research, additional factors such as motion, cognitive load, contextual information, and task complexity could be explored. Regarding complex scenarios with multiple visual and auditory targets, principles like the cocktail party effect [Kaya and Elhilali 2017; Mangun 1995] show that we can prioritize primary stimuli, suggesting that our observed effect may hold in more intricate environments. In this direction, the reported findings might help to establish a foundational framework and set the stage for future investigations in these directions.

3.6 Conclusion

In this chapter, we focus on investigating the effect of *temporal* and *spatial* alignment of visual and auditory stimuli on saccadic latency in immersive environments. Making use of VR, we study various temporal and spatial shifts between these stimuli, specifically when auditory cues precede visual information. In particular, we explore three main factors: visual eccentricity, spatial shift, and temporal shift to achieve saccade latency acceleration.

The main insights from our research reveal that temporal shifts in auditory cues can significantly accelerate saccadic responses, although this acceleration reaches a

plateau beyond certain temporal intervals. Interestingly, while visual eccentricity consistently affects saccadic latency as described in previous work [Duinkharjav et al. 2022a; Kalesnykas and Hallett 1994], spatial shifts in auditory cues do not seem to significantly impact this latency. Our findings align with previous research while extending our understanding of these cross-modal interactions by systematically considering these dimensions in VR environments and for different retinal eccentricities and sound source localizations. In our proof-of-concept game experiences, we demonstrate the practicality of our insights, validating them in more complex and interactive scenarios and showing that saccades can be effectively accelerated beyond experimental setups.

In conclusion, our findings contribute significantly to the understanding of cross-modal interactions and their influence on saccadic latency. By investigating both spatial and temporal factors in audiovisual cues, we have provided new insights into this complex interplay. The applications of these findings are broad, with potential benefits in user interfaces, AR, video games, digital multimedia, and online experiences.

Chapter 4

Overdriving Visual Depth Perception

This chapter investigates how depth perception relies on integrating audiovisual cues. In stereoscopic displays, visual depth arises from disparity and vergence movements, while auditory depth cues are conveyed through loudness, interaural differences, and spectral content. We investigate how contradictory auditory cues can overdrive visual depth perception to counteract depth compression in VR. A visual target is shifted closer to a prior fixation, while its sound cue is spatially or spectrally manipulated to bias audiovisual fusion toward the intended depth. This reduces required vergence changes and enables faster gaze retargeting. Psychophysical studies demonstrate effective depth overdriving, with findings further validated in VR/MR applications.

4.1 Introduction

Virtual reality (VR) devices support for multimodal content shows how the combination of visual, auditory, and sometimes haptic cues, also referred to as multisensory integration, enhances realism and improves the perception of our surroundings [Martin et al. 2022; Groth et al. 2023b]. Focusing on audio and visual modalities, a higher sense of immersion and presence can be achieved when sound is combined with visual content [Wu et al. 2024; Latif et al. 2021; Naef et al. 2022]. Benefits of audiovisual content are also reported in guiding navigation by drawing attention to off-screen events [Wu et al. 2024; Kudnick et al. 2025], supporting interactivity by reinforcing feedback from user actions [Latif et al. 2021], and promoting stress relief in immersive interventions [Naef et al. 2022]. In such cases, audio does not merely complement visuals but actively shapes perception and behavior.

Multimodal sensory data, both visual and auditory, are also unconsciously integrated to perceive depth in the physical world. Audiovisual depth cues improve our ability to estimate distances and accurately interpret spatial layouts [Battaglia et al. 2003; Ernst and Bühlhoff 2004; Clark and Yuille 1990]. In virtual content, visual-based depth perception in stereoscopic displays has been shown to be influenced by real-world auditory sources [Turner et al. 2011]. For fully immersive setups (e.g., AR), complementing visual depth cues with depth-scaled sound spatialization enhances relative depth perception and reduces completion times [Zhou et al. 2007].

In this chapter, we investigate *misplaced* (non-colocated) visual and auditory cues in VR and show that, when such conflicting cues are fused, the perceived depth can be overdriven from its visually defined position toward the sound’s spatial location. We focus on newly appearing targets, which typically trigger a gaze shift from the current fixation, involving an ocular movement sequence that combines a saccade to the new target’s screen position and vergence to its stereoscopic depth. A representation of the method is shown in Figure 4.1.

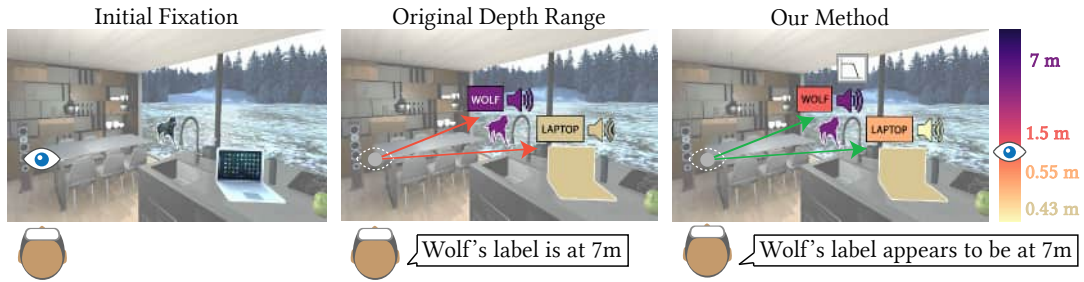


Figure 4.1: This chapter investigates the effect of spatially decoupling auditory and visual cues on depth perception in virtual reality (VR). *Left*: The user begins by fixating on some random object within a VR scene. *Center*: A new annotation (e.g., related to the wolf) appears in the distance, accompanied by colocated audio (indicated by the same purple color). Visual response and fixation on the wolf annotation requires a significant eye vergence angle change ($\Delta\theta$). *Right*: Using our method, the disparity variation of the visual annotation relative to the initial fixation point is reduced, thereby decreasing $\Delta\theta$. To compensate for this visual depth reduction, we manipulate the sound’s depth cues to increase perceived distance. Due to the fusion of visual and auditory cues, the user perceives the audiovisual annotation shifted toward the wolf’s depth. In contrast, for the laptop case, the visual reduction ($\Delta\theta$) is now compensated by adjusting the sound’s depth cues to reduce perceived distance. Note how different strategies are applied depending on the visual motion performed: sound source spatial location is altered for the laptop (convergence eye movement), while the sound frequency spectrum is modified for the wolf (divergence eye movement). In both scenarios, our method preserves accurate depth perception while reducing $\Delta\theta$, which accelerates gaze retargeting (illustrated with green arrows). The right bar visualizes physical depth.

We experiment with reducing the depth change between the current fixation and the new target by shifting the target along its line of sight without changing its screen position (e.g., the annotations in Figure 4.1). The resulting depth compression may be desirable, as it helps control potentially uncomfortable depth changes and enables faster reaction times by reducing the magnitude of vergence required.

To compensate for the compressed depth, we displace the corresponding sound location in the opposite direction with respect to the depth manipulation. We manipulate the sound sources depth by adapting their loudness and spectral frequency content accordingly, but we also displace sound laterally to act on interaural differences (IAD). Across all investigated conditions, we ensure that visual and auditory cues remain integrable, allowing them to fuse into a consistent percept that is biased toward the intended target depth location. In this context, the audiovisual fusion mechanism offers an alternative means of supporting perceived depth in scenarios where depth accuracy is reduced as a consequence of the measures required to alleviate the vergence–accommodation conflict (VAC) [Hoffman et al. 2008].

We design a perceptual experiment to systematically assess the efficiency of such perceived depth overdriving for various VR-relevant combinations of initial fixations and shifted target object locations under the described audio condition variations. Our results show that, if optimally designed, success probability in judging depth comparisons can decrease from about 80% with visual-only cues towards chance level (50%) with conflicting overdriving audio. Based on the outcome of the first experiment, we select the most successful scenarios of depth overdriving and demonstrate up to 50 ms speedup in gaze retargeting. Finally, we demonstrate the benefits of sound-based depth perception enhancement for dynamic interfaces in natural VR applications.

Our contributions throughout this chapter include:

- We show that non-colocated auditory and visual cues can be used to manipulate perceived depth, and we provide guidelines for controlling this effect in VR.
- For response-time-critical scenarios, our approach enables faster visual reactions while the underlying vergence reduction manipulation is masked by overdriving sound.
- We demonstrate practical applications of the sound overdriving method in a naturalistic VR application, with consistent results further shown in a proof-of-concept mixed reality (MR) version.

Our collected anonymized data and code are publicly available at <https://avdepthoverdrive.mpi-inf.mpg.de/>

4.2 Related Work

In this section, we discuss visual depth cues, particularly for stereovision (Section 4.2.1), their limitations in VR (Section 4.2.2), the role of auditory cues in depth perception (Section 4.2.3), and finally the integration of both audio and stereovision (Section 4.2.4).

4.2.1 Depth Perception and Eye Movements

Human depth perception relies on both monocular and binocular cues [Cutting and Vishton 1995]. Monocular cues such as perspective, occlusion, shading, and motion parallax convey depth through experience and context [Palmer 1999], while binocular stereopsis provides a direct sense of depth from retinal disparities [Howard and Rogers 2012]. Whereas monocular cues benefit from familiarity, stereopsis functions as a low-level, pre-attentive mechanism independent of prior knowledge [Cutting and Vishton 1995; Howard and Rogers 2012]. Depth cues may conflict, but the brain resolves inconsistencies through cue fusion, modeled as strong fusion (winner-takes-all), weak fusion (Maximum Likelihood Estimation weighting by cue reliability), or modified weak fusion (Bayesian inference with priors) [Landy et al. 1995; Knill and Richards 1996].

We focus on stereovision, where saccades and vergence are the primary eye movement mechanisms during free scene exploration. Referring to the previous chapter, saccadic

movements are preceded by a 200 ms latency period and last 20–100 ms depending on amplitude [Bahill et al. 1975b; Arabadzhiyska et al. 2017; van Beers 2008].

On the other hand, vergence movements involve opposite-direction eye rotations driven by depth changes between fixations, with latencies of 180–250 ms and rotations lasting 200–1000 ms depending on depth magnitude [Semmlow and Wetzel 1979; Templin et al. 2014]. In scenarios with fixation shifts at different depths, combined saccade–vergence movements can occur, producing up to 20% vergence speedup at the cost of saccades slowing by up to 50% [Yang et al. 2002; Yang and Kapoula 2004; Zee et al. 1992; Collewijn et al. 1995; Duinkharjav et al. 2023].

4.2.2 Limitations of Stereovision in VR

Standard VR displays lack the eye accommodation cue, leading to the vergence–accommodation conflict (VAC) [Hoffman et al. 2008; Shibata et al. 2011]. This mismatch between accommodation, set by the display distance, and vergence, aligned with the fused object, causes discomfort [Hoffman et al. 2008; Shibata et al. 2011; Groth et al. 2023a] and prolongs vergence duration [Spiegel and Erkelens 2024]. A common mitigation strategy compresses scene depth by manipulating disparity, i.e., the horizontal pixel shift between left- and right-eye images [Lang et al. 2010; Didyk et al. 2011; Groth et al. 2024]. However, disparity manipulation reduces perceived depth and can conflict with other monocular depth cues. To address excessive disparity compression and VAC, fixated regions can be shifted closer to the display while expanding the disparity range around the display (zero disparity) [Jones et al. 2001; Oskam et al. 2011; Bernhard et al. 2014; Kellnhofer et al. 2016a]. For controllable elements such as annotations, disparity changes relative to fixation should be minimized, and when re-vergence is unavoidable, inducing saccades can reduce vergence duration [Duinkharjav et al. 2023]. A similar issue arises in stereoscopic film cuts, where depth differences at likely fixation points are reduced in post-production [Templin et al. 2014].

Since visual adjustments alone may not resolve excessive depth compression, we explore auditory depth cues. By shifting a new visual target closer to fixation while displacing the associated audio source in the opposite direction, we overdrive depth perception and reduce the vergence demand.

4.2.3 Depth Perception in Human Auditory System

Audio is an important cue for localizing the distance of audiovisual targets, providing semantic information and spatial cues such as loudness, interaural differences (IAD), frequency spectrum, and reverberation [Zahorik et al. 2005]. Loudness is usually the primary cue, being interpreted based on contextual comparisons rather than an inherent scale of sound intensity [Strybel and Perrott 1984]. In addition, interaural time differences (ITD) and interaural intensity differences (IID), the components of IAD, jointly contribute to spatial localization but are less effective for absolute depth estimation [Brungart and Rabinowitz 1999]. Moreover, the frequency spectrum changes with distance, as high frequencies attenuate more than low frequencies due to air absorption [Little et al. 1992]. Reverberation also serves as an additional cue, where closer sounds have a higher direct-to-reverberant ratio and distant sounds contain more reflected components [Bronkhorst and Houtgast 1999]. Ultimately, these cues are integrated to support more precise localization of sound sources in 3D space.

4.2.4 Depth Perception with Multimodal Cues

Colocated audio and visual cues significantly improve location judgment and reaction time compared to visual-only conditions [Grohn et al. 2005; Bolia et al. 1999; Kim et al. 2019], an effect also confirmed for depth-scaled 3D sound in AR environments [Zhou et al. 2007]. The focus of this chapter, however, is on the integration of *non-colocated* audio and visual cues. A prominent example is the ventriloquism effect, or “visual capture,” in which spatially misaligned auditory cues are perceptually pulled toward the visual signal in a winner-takes-all manner. Importantly, such audio–visual integration is constrained by the limits of tolerable misalignment, both in the angular domain, as directional separation between cues [Jack and Thurlow 1973; Godfroy et al. 2003], and in the spatial domain, as distance misalignment [Zahorik 2003; Gorzel et al. 2012]. More generally, misaligned cues interact to produce a perceived cross-modal location, analogous to visual depth cue fusion (Section 4.2.1) [Berry et al. 2014; Turner et al. 2011; Landy et al. 1995; Knill and Richards 1996], with visual capture [Jack and Thurlow 1973], weak fusion [Clark and Yuille 1990], and Bayesian models accounting for visual bias [Battaglia et al. 2003; Ernst and Bühlhoff 2004] describing aspects of audiovisual depth integration. While feedback-driven recalibration can adjust perceived depth via cue reweighting [Lin et al. 2024; Kramer et al. 2020; Kumpik et al. 2019], this mechanism is outside the scope of the present chapter, which focuses on fast-performance applications without feedback.

The impact of temporal and spatial misalignment between auditory and visual cues has been studied in monoscopic and stereoscopic displays [Van der Stoep et al. 2015; Ross and Ross 1981; Jiménez-Navarro et al. 2024b]. Audio preceding visual onset can reduce saccade latency and enhance performance in time-critical tasks [Jiménez-Navarro et al. 2024a]. Closest to our goals is the work of Turner et al. [Turner et al. 2011], who demonstrated that spatially misaligned audio can bias perceived depth. Using a stereo 3D desktop display with a pair of loudspeakers positioned at different distances to produce a stereo effect, they found that sounds from the closer pair caused the stereoscopic display’s apparent depth to be perceived as shifted toward that speaker location.

In this chapter, we employ a VR setup with headphone-based sound spatialization to systematically investigate how loudness, IAD, and spectral content can overdrive visual depth perception across different vergence angles and eccentricities during free scene exploration while combining saccade and vergence movements.

4.3 Method

Our goal is to study how sound cues dislocated from their visual counterparts can overdrive stereoscopic depth perception. To this end, our methodology exploits manipulations of sound loudness based on spatial shifts, IAD, and frequency spectrum, with different procedures for convergence and divergence eye movements.

Theoretical Background Key components in visual depth perception are stereoscopic disparity and eye vergence [Mon-Williams et al. 2000]. The *vergence angle* θ , is defined by the convergence of both eyes’ gaze rays on a fixation point. For an object at distance d , it is given by $\theta = 2 \cdot \arctan(\text{IPD}/2d)$, where IPD is the interpupillary distance. At close range, even small changes in distance result in significant ver-

gence shifts. Objects positioned too close to the observer cannot be fused, leading to diplopia, while objects far away make gaze rays parallel, limiting the vergence range. On the other hand, audio-driven distance perception is primarily influenced by sound intensity that reaches the listener I , which follows the inverse square law $I = I_0/4\pi d^2$ where I_0 is the sound source intensity and d the distance between both. However, humans perceive intensity differences as relative changes [Strybel and Perrott 1984], denoted as I_r in decibels (dB):

$$I_r = 20 \cdot \log_{10} \left(\frac{R}{R_{\text{ref}}} \right) \quad (4.1)$$

where R is the distance of the sound location and R_{ref} the reference distance.

Hypothesis We hypothesize that sound-driven depth cues can become strong enough to overdrive contradictory visual depth perception arising from disparity and vergence. Under this assumption, we deliberately move the newly appearing visual target closer to a previous fixation point to reduce the required vergence adjustment. We refer to this newly appearing target as $\mathbf{T}_{\text{test}}$ (Figure 4.2). Our technique then aims to reproduce the depth percept associated with the full vergence change for the originally intended target ($\mathbf{T}_{\text{ref}}$) by manipulating the auditory cues of $\mathbf{T}_{\text{test}}$. To achieve the intended stereoscopic depth overdriving, we build on prior work [Strybel and Perrott 1984; Zahorik et al. 2005] that extensively analyzes the effectiveness of relative differences in auditory depth cues. For sound sources located close to the listener, distance-based loudness (Equation (5.1)) and interaural differences (IAD) are the most effective depth cues [Kolarik et al. 2016]. This situation corresponds to the convergence eye movements, in which the visual target approaches the observer, and our overdriving strategy places the sound source even closer. In Section 4.3.1, we discuss the convergence motion condition, where, along with distance-based loudness modulation (Equation (5.1)), we appropriately manipulate the spatial location of the audio source to enhance IAD (Figure 4.2, left). At larger sound source distances, however, distance-based loudness and IAD cues become progressively less effective, making them unsuitable for divergence eye movements, where the distance to the visual target increases. In this case, our method aims for the fused audiovisual cue to be perceived as even farther away, reflecting a behavior distinct from the convergence condition. Therefore, we replace these cues with spectral cues, which are known to remain informative at greater distances [Kolarik et al. 2016]. In Section 4.3.2, we detail how low-pass filtering of the sound frequency spectrum (Figure 4.2, right) produces frequency-based loudness variations that are employed in the divergence motion condition.

4.3.1 Sound Manipulation for Eye Convergence

In the convergence motion condition, we aim to perceive $\mathbf{T}_{\text{test}}$ closer towards $\mathbf{T}_{\text{ref}}$'s depth. The simplest scenario to be tried is to shift $\mathbf{T}_{\text{test}}$'s sound source closer to the observer along the line of sight. In that case, the sound intensity naturally increases according to Equation (5.1). However, to further enhance the overdriving effect, we also introduce lateral shifts, separating the audio source from the line of sight towards the periphery, which induce strong IAD cues. Acting on IAD alters the perceived distance of lateral audiovisual stimuli, replicating how sound behaves when objects are closer [Hartmann 2021]. Our approach is depicted in Figure 4.2 (left).

However, to maintain audiovisual coherence, the angular (lateral) offset between visual and auditory cues (β) must be restricted to the fusion limit $\beta_{\max} = 16^\circ$ to prevent multimodal integration from breaking down into distinct auditory and visual stimuli [Godfroy et al. 2003]. For example, in Figure 4.2 (left), the positions p_{1-3} are constrained to the maximum lateral shift $\beta = \beta_{\max}$, thus the audio source can only move toward the observer.

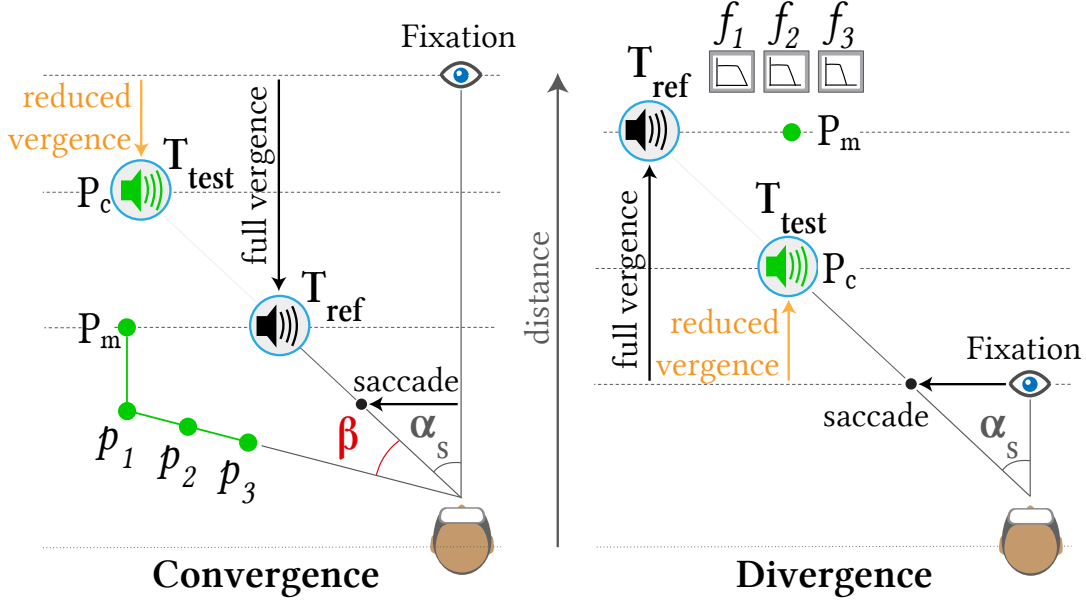


Figure 4.2: Illustration of our method. Given the current fixation, the originally intended location (T_{ref}) requires a large vergence change while a reduced change is required for our proposed visual location (T_{test}). For convergence (left), we strengthen spatial-based loudness variations with IAD by detaching the audio signal from the visual target. For divergence (right), we consider a static spatial offset between sound and visual cues and purely alter the auditory frequency spectrum. In both cases, the magnitude of the required vergence movement can be reduced due to the closer visual target (orange arrow) from fixation. Labels P_c , P_m , p_{1-3} and f_{1-3} indicate the different loudness variations considered for T_{test} in the main experiment in Section 4.4.

4.3.2 Sound Manipulation for Eye Divergence

On the other hand, in the divergence motion condition, we aim to perceive T_{test} farther away towards T_{ref} 's depth. Since distant sounds are characterized by a higher proportion of low-frequency content due to atmospheric absorption [Little et al. 1992], this effect is replicated by applying a low-pass filter to the sound signal, which naturally reduces audio loudness. In this case, loudness variations based only on spatial shifts would end up breaking the spatial limits of audiovisual fusion due to the logarithmic nature of sound loudness variations (Equation (5.1)) [Gorzel et al. 2012]. Consequently, the filtered sound is always presented matching T_{ref} 's depth (P_m) (see Figure 4.2, right). For even a further intended perceived depth, we intensify the filtering effect by modifying the cut-off frequency (f_1, f_2, f_3). For example, for the beep-like audio signal centered at 880 Hz used in Section 4.4, the cut-off frequencies are selected within the fifth octave (740 or 660 Hz) to obtain the

desired loudness variations. On the lower end, cut-off frequencies with excessive signal degradation, which causes the sound to be barely audible, are prevented. The highest cut-off frequency is set to 880 Hz, yielding a loudness reduction of 3 dB (f_1), which is comparable to the p_1 sound overdriving level (Figure 4.2; left) in the convergence case (Section 4.3.1). A more detailed discussion of the specific audio settings and conditions used in evaluating depth perception overdriving is provided in Section 4.4.

4.4 Sound-Based Depth Perception Modulation

In this section, we present our main psychophysical experiment, which examines how modifications in auditory cues influence visual depth perception. Applying the approach outlined in Section 4.3, we determine the levels of audio signal modulation that conceal visual depth shifts while maintaining audiovisual fusion.

4.4.1 Experimental Setup

Hardware As in the previous chapter, user studies were conducted using a Varjo Aero HMD (200 Hz eye-tracking, accuracy of 1° of visual angle) and the sound was played by Beyerdynamic DT 770M headphones with passive noise cancellation. The experiment was designed using the Meta XR Audio Spatializer, handling audio spatial propagation. Directional audio was rendered using a generic head-related transfer function (HRTF), which offers greater generalizability and has been reported to provide accurate sound localization in similar studies [Jiménez-Navarro et al. 2024a]. Regarding the spatializer parameters, scene reverberation was removed by applying a -60 dB intensity reduction, the pitch of audio sources remained unmodified, and the overall scene volume was reduced by -5 dB to maintain comfortable listening levels.

Audiovisual Stimuli and Depth Cues Control As visual cues, we employed 3D traffic cones with a visual angle size of 0.7° at eye level, featuring high-contrast orange and white stripes. Conversely, the background 3D scene was gray and mostly empty. With a lifetime of 2 s, targets were shown well above the typical visual response times [Duinkharjav et al. 2023]. Simultaneous with the onset of the visual cue (traffic cone), a short beep-like sound centered at 880 Hz was played for 0.5 s, following the sound design of previous works [Jiménez-Navarro et al. 2024a; Zambarbieri 2002]. Such audiovisual stimuli were employed as both fixation and targets. Refer to Figure 4.3 for more details on the sequential presentation of the stimulus.

In order to ensure experimental control over the existing visual cues, audiovisual stimuli are attached to the user's head [Prummer et al. 2025a] while the background scene is world-referenced. Since target locations are fixed in the user's field-of-view (FOV), motion parallax from depth-related motion drifts on the retina is canceled [Kellnhofer et al. 2016b; Rogers and Graham 1979]. This ensures controlled saccades, as head motion does not alter target presentation. Simultaneously, head jitter and uniform motion-drift between all stimuli and the background are enabled by 6 DoF head-tracking. This design preserves vestibular feedback cues and enhances VR immersion, compared to other head-fixed setups such as chin-rest [Mikawa and Fukiage 2024] or head tracking cancellation [Konrad et al. 2020]. Additionally, the target scale is made depth-independent, always displaying it with 0.7° size horizontally.

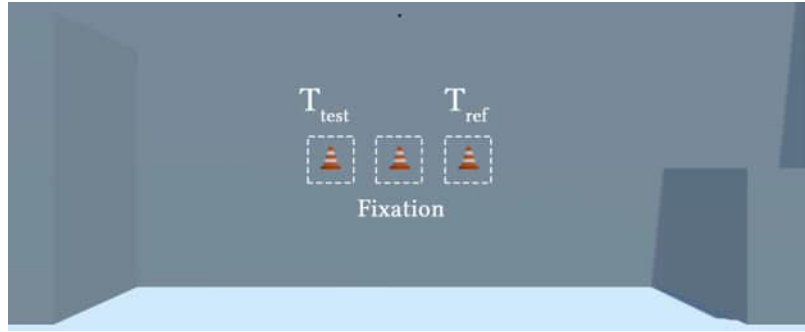


Figure 4.3: User point of view during the main user study. Visual stimuli (traffic cones) appeared sequentially and only one was visible at the same time (dashed lines). All of them (T_{ref} , T_{test} and Fixation) have their size increased by a factor of 4 for visualization purposes. In this case, Fixation is placed at 0.8 m, T_{ref} is at 1.5 m and T_{test} is at 7 m, corresponding to Divergence-Far condition (Table 4.1). Note that the stimulus size is fixed irrespective of distance to eliminate undesired pictorial monocular depth cues.

In our case, due to the small size of the visual cue, undesired depth indications based on texture gradients are minimal and not expected to affect our measurements. Targets' shadow-casting is disabled, and directional light ensures uniform shading of the targets in the entire depth range, leaving disparity and vergence as the only visual depth cues. Lastly, regarding audio depth cues, spatializer parameters allow us to disable audio reverberation to remove its influence on the measured overdriving effect. Since inner scene geometry changes are needed to study reverberation differences, our study is limited to audio source depth cues for a broader generalization, such as spatial location and frequency content.

Procedure All experiment sessions started with a five-dot eye-tracking calibration. During the experiment, both T_{ref} and T_{test} audiovisual stimuli were sequentially presented to the user who selected **which one was closer in depth**. Each trial began with the fixation stimulus at the center of the screen. When the fixation stimulus disappeared, the first audiovisual target was presented, eliciting the participant's visual response. The fixation stimulus then reappeared, followed by the second audiovisual target, and the trial concluded with the fixation stimulus shown once more. After all stimuli appeared, participants were instructed to indicate the target that was closer to them by button pressing, relying on both sound and visual cues. Then, the next trial starts. The order (first or second) and side (left or right) of the T_{ref} and T_{test} cases was randomized, while opposite sides for each case were always ensured. The experiment was conducted in three sessions of 20 minutes to prevent fatigue, while the order of the trials was also randomized across sessions. Over the three sessions, each participant completed a total of 240 trials with five minutes rest between sessions. The entire experiment lasted 1.5 hours.

Conditions and Sound Modulation The variables we consider are *eccentricity* (E), *loudness shift* (ΔL), *vergence range* (R), and *vergence direction* (V). The *eccentricity* (E) describes the horizontal visual angle between the fixation and the visual targets. We choose $E = \{8^\circ, 16^\circ\}$ to align with common saccade displacements [Bahill et al. 1975a]. When visually shifting the focus point at different depths, the vergence

movement can either diverge from the observer or converge to a closer target, setting the values for *vergence direction* (V) as $V = \{C, D\}$. Near and far distance ranges are investigated separately, setting $R = \{N, F\}$ for *vergence range* (R). All Direction-Range combinations can be seen in Table 4.1, where distances and vergence angles of visual cues are shown.

Condition (Direction - Range)	Fixation	T_{ref}	T_{test}
Convergence - Near	0.8 m–4.5°	0.43 m–8.5°	0.55 m–6.5°
Convergence - Far	7 m–0.5°	0.8 m–4.5°	1.5 m–2.5°
Divergence - Near	0.43 m–8.5°	0.8 m–4.5°	0.55 m–6.5°
Divergence - Far	0.8 m–4.5°	7 m–0.5°	1.5 m–2.5°

Table 4.1: Physical depth and corresponding vergence angle (θ) related to fixation, reference target (T_{ref}), and test target (T_{test}) visual cues for the vergence Direction-Range combinations considered. Importantly, the placement of visual cues is adjusted according to the corresponding eccentricity for a correct vergence angle θ .

Note how vergence angle variations ($\Delta\theta$) for reference (T_{ref}) and test (T_{test}) cases from fixation are respectively $\Delta\theta = 4^\circ$ and $\Delta\theta = 2^\circ$. The whole vergence range considered in this chapter is displayed in Figure 4.4 for a direct visualization.

Regarding auditory modality, different sound modulation levels for the test case (T_{test}) relative to the reference case (T_{ref}) are considered within *loudness shift* (ΔL). For convergence (Figure 4.2, left), we consider $\Delta L = \{\Lambda \text{ dB } (P_c), 0 \text{ dB } (P_m), +3 \text{ dB } (p_1), +6 \text{ dB } (p_2), +9 \text{ dB } (p_3)\}$. For divergence (Figure 4.2, right), we consider $\Delta L = \{\Lambda \text{ dB } (P_c), 0 \text{ dB } (P_m), -3 \text{ dB } (f_1), -8 \text{ dB } (f_2), -13 \text{ dB } (f_3)\}$. Note that $\Lambda \text{ dB}$ denotes case-dependent loudness variations (Table 4.1). Next, we explain each of these conditions:

Colocated, P_c : The audio source is located at the same spatial position as the visual target T_{test} , so the loudness variation with respect to T_{ref} ($\Lambda \text{ dB}$) depends on each case (Table 4.1).

Matching depth, P_m : T_{test} 's audio source is located at the same depth as T_{ref} . Therefore, sound cues associated to the T_{test} and T_{ref} targets have minimal differences in loudness, resulting in $\Delta L = 0 \text{ dB}$. In this first overdriving attempt, the auditory depth shift tries to compensate for the visual depth change, testing if users mainly rely on the auditory cue for deciding on the depth of an object.

Overdriving Level 1, 2 and 3: In convergence, it covers the increased loudness shifts of +3 dB (p_1), +6 dB (p_2) and +9 dB (p_3), achieved by spatially moving T_{test} 's audio source closer to the user according to Equation (5.1). We chose steps of +3 dB as they represent doubling the power of the sound [Marks 1979], resulting in a small but noticeable change in loudness by human observers. To enhance the effect, we also maximize IAD while ensuring audiovisual fusion (Figure 4.2, left). In the case of divergence, since no IADs are introduced, the method is boosted by modifications of the frequency spectrum and consequent loudness shifts of -3 dB (f_1), -8 dB (f_2) and -13 dB (f_3) (Figure 4.2, right).

We employed a fully factorial design with $2(E) \times 2(R) \times 2(V) \times 5(\Delta L) \times 2(\text{left/right}) = 80$ conditions. Each condition was repeated three times, yielding 240 trials per user.

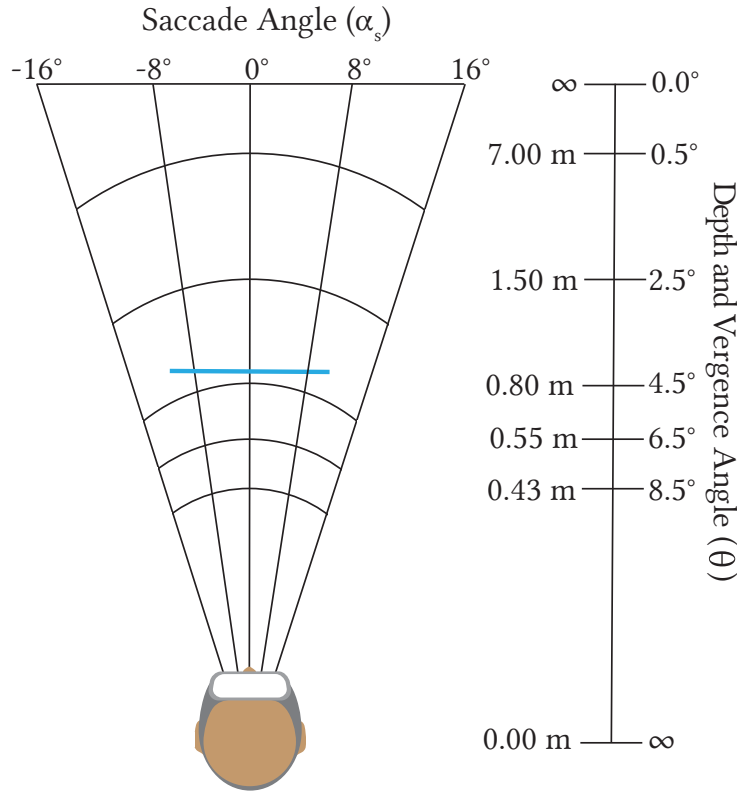


Figure 4.4: Vergence range considered in this chapter. The vergence angle formed between both gaze rays depends on the depth of the visual target that is being fixated on. The closest distance is set to 0.43 m ($\theta = 8.5^\circ$) and the farthest limit is set to 7 m ($\theta = 0.5^\circ$), creating a vergence range of 8° . The focal distance of the Varjo Aero HMD at 0.85 m is highlighted (light blue). Scheme adapted from [Duinkharjav et al. 2023].

Baseline Study Before running the main experiment, a pilot study was run with ten participants (four female, six male, ages 21-33) to test depth perception performance for every condition considered without any sound cue (no audio, only visual stimuli). The pilot was designed and performed to validate users' ability to properly distinguish visual depth changes above the stereoacuity threshold used in our experimental setup [Didyk et al. 2011]. The procedure explained in the previous section also applies in this study, and the results are shown together with the main study for better comparison (Figure 4.5).

4.4.2 Results

A total of fourteen participants took part in the main experiment (six females, eight males, ages 21-33). All reported normal or corrected-to-normal vision and hearing, and written consent was provided. User study results are shown in Figure 4.5. A Generalized Linear Mixed Model (GLMM) with a logit link function was applied to analyze the binary dependent variable (success at selecting the closest target), including fixed effects (*eccentricity*, *loudness shift*, *vergence range*, *vergence direction*, and *side* of the test target) and a random intercept to account for variability across participants. Post-hoc pairwise comparisons using estimated marginal means were conducted to check differences between all levels of sound loudness shift condition.

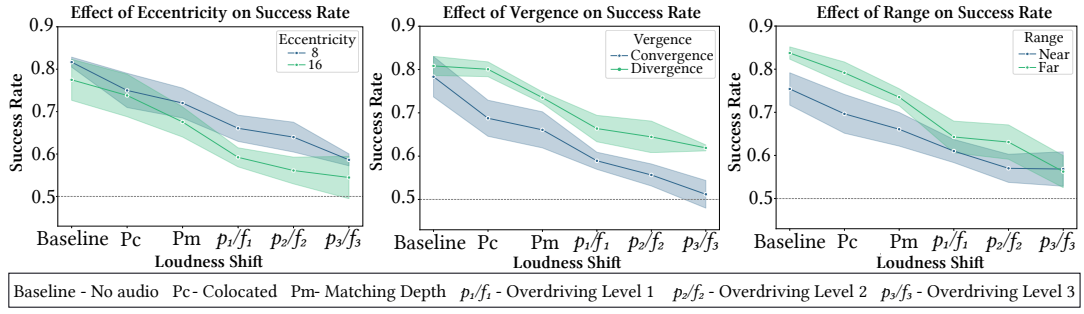


Figure 4.5: Results of the main experiment on overdriving depth perception. The success rate indicates the probability of correctly selecting the stimulus visually closer to the participant: T_{ref} in Convergence and T_{test} in Divergence. We separately visualize the effect of eccentricity (*left*), vergence direction (*center*), and vergence range (*right*) on success rate. Data points indicate mean success rates across participants for each sound condition (linear trends for visualization only), and shaded regions indicate SEM error. The width of the shaded areas reflects variability across participants. The Baseline condition shows correct stereo acuity (success rate around 80%) without any sound influence. By increasing sound modulation, success rates decrease toward chance level (dashed line), suggesting an intensified perceptual bias.

Additionally, binomial tests compared participants' performance against chance level (50%) across all sound conditions and experimental subgroups ($E = 8^\circ$, $E = 16^\circ$, convergence, divergence, near range and far range). Refer to [Section 7.3](#), [Appendix C](#) for detailed numerical results of the analyses. The analysis revealed significant effects for *loudness shift*, *eccentricity*, *vergence direction*, and *vergence range*, while the *side* did not significantly influence success rates ($p = 0.69$), suggesting that eye dominance does not influence depth discrimination. Participants showed a high success ratio of approximately 80% in the Baseline condition, consistent with expectations that vergence differences of 2° between T_{ref} and T_{test} are well above the stereoacuity threshold [[Didyk et al. 2011](#)].

Higher *loudness shift* (ΔL) led to lower success rates ($p < 0.001$). Post-hoc pairwise comparisons confirmed significant differences among several sound conditions, particularly between small (P_c and P_m) and large (p_2/f_2 and p_3/f_3) loudness shifts, supporting the notion of auditory-induced overdriving of visual depth perception. Importantly, individual differences, while moderate (group-level variance of 0.9369), were not the primary source of the effect, emphasizing the findings' generalizability.

A significant effect of *eccentricity* was observed ($p < 0.001$), with lower accuracy for higher eccentricities, reflecting a reduction in stereoacuity as stimuli moved further into the periphery. This aligns with the well-documented decrease in depth perception accuracy at larger eccentricities [[Siderov and Harwerth 1995](#)].

The effect of *vergence direction* was also significant ($p < 0.001$), showing better performance for convergence than divergence. Such a difference shows how effective interaural differences (IAD) are processed as a depth cue at near distances, being more reliable than frequency-based cues for far distances even with increased loudness variations (e.g., 8–12 dB with f_2 and f_3).

For *vergence range* ($p < 0.001$), there is a stronger decrease in accuracy in near-range conditions, where distance ranges from 0.43 m to 0.8 m (corresponding to θ ranging from 8.5° to 4.5°), compared to far-range conditions, where distance ranges from 0.8 m

to 7 m (corresponding to θ ranging from 4.5° to 0.5°). This may be associated with the VAC, where closer stimuli align better with natural viewing conditions than farther stimuli, although further investigation is needed to disentangle the contributions in convergence or divergence movements. Nevertheless, no difference is visible for the largest loudness shift (Overdriving Level 3) used in the following sections.

Binomial tests also confirm what is visually apparent in [Figure 4.5](#). In most conditions, performance was statistically similar to chance when considering the strongest manipulation (Overdriving Level 3), indicating that overdriving the visual depth perception through sound made the target's depth indistinguishable from the reference. Only in conditions with the lowest eccentricity ($E = 8$, $p = 0.0018$) and under divergence ($p < .0001$), performance was statistically distinguishable from chance level. This suggests a reduced, but still measurable, influence of the audiovisual manipulation.

4.4.3 Discussion

The results of the main experiment confirm our hypothesis that auditory cues can effectively modulate perceived depth and fully overdrive stereoscopic visual depth perception. As shown in [Figure 4.5](#), increasing the magnitude of the sound manipulation progressively reduces participants' ability to distinguish between T_{test} and T_{ref} targets, with performance approaching chance level. This indicates that the audiovisual fusion process successfully masked the underlying visual disparity manipulation. Thereby, the effect was observed consistently across both near and far vergence ranges, covering depths in the full operational depth range of standard VR systems from 0.43 m to 7 m. The effectiveness of the overdriving varies depending on eccentricity and vergence direction. For lower eccentricity ($E = 8^\circ$), the overdriving effect was weaker compared to the $E = 16^\circ$ condition, which may be attributed to the reduced strength of IAD at smaller lateral displacements. At higher eccentricity, IAD becomes a stronger cue due to increased lateralization of the sound source, especially in near-range convergence conditions [[Kolarik et al. 2016](#)]. Similarly, in divergence motion, the method is less effective than in convergence. Again, the IAD impact is reduced given the limited horizontal angle displacement for larger distances. Overall, the results support our hypothesis that spatially misaligned sound cues can override visual depth judgments in stereoscopic displays, provided the manipulation is strong enough and well-tuned to the perceptual sensitivity of the respective cue type (IAD vs. spectrum).

4.5 Visual Response Acceleration

The results of the main experiment ([Section 4.4](#)) show that moving a visual target closer to a former fixation can remain indistinguishable if the shift is properly compensated by sound modulation. In this section, we evaluate the effect of this depth reduction on the vergence response time. To do so, we apply Overdriving Level 3 sound modulation from [Section 4.4](#) to test how the time needed to fixate on a new target significantly decreases for smaller vergence variations. Consequently, depth-overdriven targets would significantly increase user performance, as visual information is recovered faster.

4.5.1 Experimental Setup

To optimally measure the response behavior, we specifically place the targets' visual cues at the limits of the vergence range. This means, we focus on Convergence-Near and Divergence-Far conditions (Table 4.1), which allow us to shorten the vergence range from 0.43 m to 7 m in distance (with θ ranging from 8.5° to 0.5° , $\Delta\theta_{max} = 8^\circ$) to a new compressed range from 0.5 m to 1.5 m in distance (with θ ranging from 6.5° to 2.5° , $\Delta\theta_{max} = 4^\circ$). Meanwhile, sound manipulation allows for overdriving perceived depth as shown in the previous section.

Stimuli and Procedure Compared to the main experiment, the same scene (Figure 4.3, background) and Overdriving Level 3 loudness shift (Section 4.4) were used. Regarding visual cues, here the initial fixation displayed a centered red cross, while visual targets displayed a clustered Landolt C shapes of 0.7° , as shown in Figure 4.6. Such a visual target ensures peripheral masking, thus enforcing proper fixation [Spiegel and Erkelens 2024].



Figure 4.6: Red cross (*left*) used as initial fixation and clustered Landolt C shapes (*right*) used as target to measure vergence time. Surrounding Landolt C's pattern is used to avoid orientation discrimination using peripheral vision, ensuring correct foveation [Spiegel and Erkelens 2024].

The participant's task was to react to the newly appearing audiovisual targets and provide **the orientation of the central Landolt C as fast as possible** by button-pressing after appearance. In each trial, participants started by looking at the fixation cross, which eventually disappeared, showing the first audiovisual target T_1 . After the first response was introduced, the second target T_2 was sequentially generated at the other extreme of the vergence range. We measure the response time between the onset of both audiovisual targets, T_1 and T_2 , and the manual response provided for each of them. These times include saccade and vergence motions and cognitive processing. Notably, if the response provided for T_1 or T_2 did not match the orientation of the central Landolt C, the trial was rejected and the response time was discarded. Throughout the visual response acceleration experiment, a step scenario is always employed between the offset of the previous visual anchor and the onset of the new audiovisual cue. This setup avoids variations in visual response latency typical of other conditions, such as gap or overlap modes between visual stimuli [Kingstone and Klein 1993; Kalesnykas and Hallett 1987]. Therefore, saccade latency remains unaffected in our conditions as well as saccade response time since they are associated with visual eccentricities [Jiménez-Navarro et al. 2024a]. Manual response delay time acts as an offset for all measurements, so the response time variations measured are associated with temporal differences in vergence motion.

The temporal scheme of the total response time measured, considering visual saccade, vergence and manual response, is depicted in Figure 4.7.

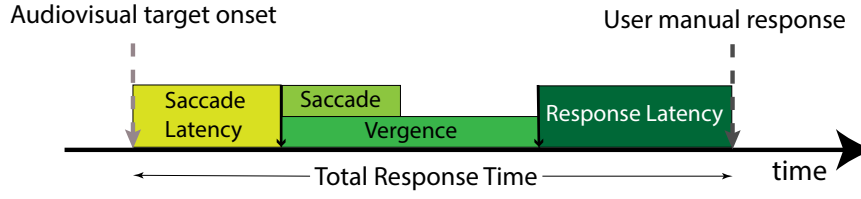


Figure 4.7: Temporal representation of response time measured. Target onset indicates the start of the visual response, covering saccade latency and saccade execution time, as well as vergence execution time [Duinkharjav et al. 2023]. Additionally, button press latency is also considered in the total response time.

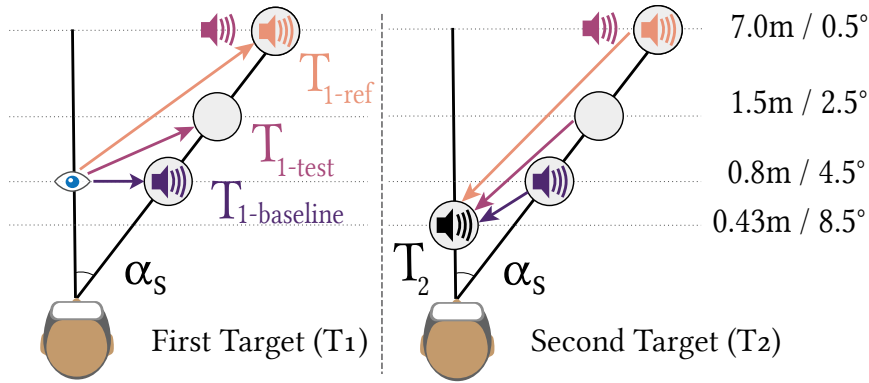


Figure 4.8: Scheme showing the audiovisual cues location used in Divergence-Far condition cases. (left) After the initial fixation at 0.8 m, the subsequent fixation at the first target T_1 requires different divergence changes ($\Delta\theta_{T_1} = 0^\circ, 2^\circ$, or 4°) depending on the case. (right) Consequently, corresponding convergence motions to fixate at the second target T_2 at 0.43 m ($\Delta\theta_{T_2} = 4^\circ, 6^\circ$, or 8° , respectively) are required. Note that the $T_{1\text{-baseline}}$ and $T_{1\text{-ref}}$ cases have audio co-located with the visual target. In the $T_{1\text{-test}}$ case, overdriving sound is applied toward the original position at $T_{1\text{-ref}}$ following main experiment results (Section 4.4). $T_{1\text{-baseline}}$ is included for further comparison. A representation of the recorded gaze trajectories on angular displacements for each T_1 case is shown in Figure 4.12.

Conditions and Participants Aside from the two vergence Direction-Range (V-R) conditions considered (Convergence-Near and Divergence-Far), we also test three different cases (C) for the depth of the first target (T_1) from fixation: $T_{1\text{-baseline}}$ ($\Delta\theta_{T_1} = 0^\circ$, pure saccade); $T_{1\text{-test}}$ ($\Delta\theta_{T_1} = 2^\circ$, T_{test} with overdriven depth referred in previous sections); and $T_{1\text{-ref}}$ ($\Delta\theta_{T_1} = 4^\circ$, T_{ref} in previous sections), as displayed in Figure 4.8 (left). Consequently, the second target T_2 ($E = 0^\circ$) at the opposite extreme of the range with colocated audio requires corresponding revergence motions according to each T_1 case, being respectively $\Delta\theta_{T_2} = 4^\circ$, $\Delta\theta_{T_2} = 6^\circ$ and $\Delta\theta_{T_2} = 8^\circ$ (Figure 4.8, right).

See Figure 4.8 for an illustration of the Divergence-Far condition. Notably, while Divergence-Far condition requires a divergence-convergence motion sequence to fixate at T_1 and T_2 , the opposite convergence-divergence motion sequence is required for the Convergence-Near condition. Lastly, 8° and 16° eccentricity values (E) are also considered.

Each of the 24 combinations – $2(E) \times 2(V - R) \times 3(C) \times 2(left/right)$ – is presented 8 times, resulting in a total of 192 trials per participant. The 192 trials are conducted in two sessions with five minutes of rest in between, and a total experiment length of one hour. A total of five participants took part in the experiment (two females, three males, ages 21-28).

4.5.2 Results and Discussion

Results of the accumulated response $T_1 + T_2$ are shown in Figure 4.9. A Linear Mixed Model (GLMM) was used to analyze the dependent variables (response times for T_1 , T_2 , and accumulated $T_1 + T_2$), including fixed effects (E , T_1 side, $\Delta\theta$, vergence direction) and a random intercept to account for variability across participants.

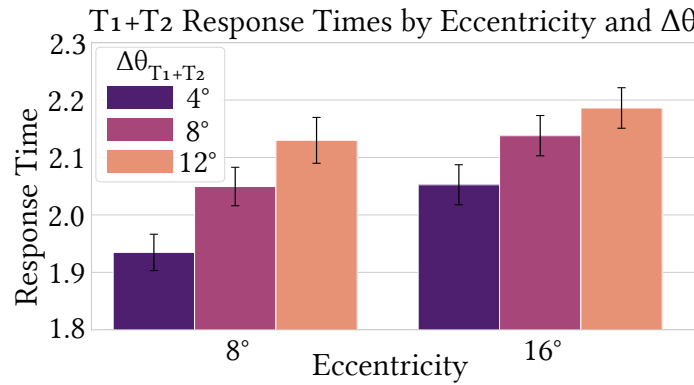


Figure 4.9: Accumulated response times for the $T_1 + T_2$ sequence. Response times increase with larger variations of vergence angle $\Delta\theta_{T_1+T_2}$ and larger eccentricity, while error bars show standard error of the mean (SEM).

A significant effect of eccentricity ($p < 0.001$) shows that larger eccentricities increase saccade latency, showing a slower reaction time towards the periphery [Jiménez-Navarro et al. 2024a]. We also observe differences for $\Delta\theta$ ($p < 0.001$), where larger response times are found for vergence motions with larger $\Delta\theta$, a consequence of larger disparity variations [Templin et al. 2014]. No significant differences were observed when comparing convergence vs. divergence motion ($p = 0.948$), in line with previous work [Duinkharjav et al. 2023]. The side of appearance (*left/right*) did not significantly influence response times ($p = 0.585$), suggesting that ocular dominance does not affect performance in this task either. In Figures 4.10 and 4.11, we can respectively find the measured timing differences for T_1 and T_2 targets employed in the experiment. Same findings can be observed for these cases as for the reported $T_1 + T_2$ combination. Larger variations in vergence angle cause slower response times, while higher eccentricities also slow down the visual response due to longer times in saccade latency [Jiménez-Navarro et al. 2024a].

For a better understanding of the visual response, different visual performance profiles can be found in Figure 4.12, where each graph corresponds to the different cases of T_1 ($T_{1-baseline}$, T_{1-test} and T_{1-ref}) in the Divergence-Far condition. To visualize the real vergence angle variations corresponding to different fixation depths, we can analyze the recorded eye tracking data accordingly. Since we focus on purely horizontal visual motion, both saccades and vergence, left and right gaze rays are projected and normalized to the horizontal plane. Consequently, the vergence angle (θ) can be computed as the difference between the left and right gaze horizontal

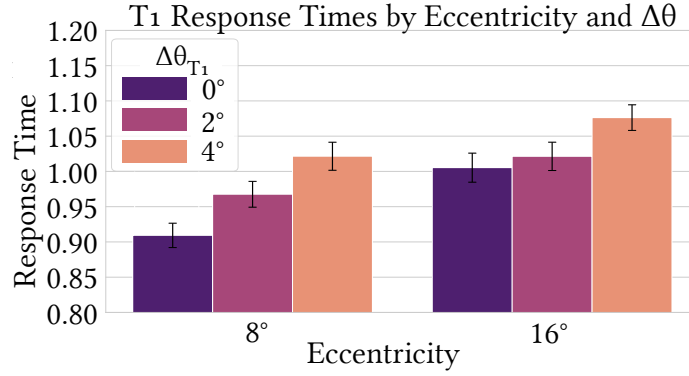


Figure 4.10: Manual response times for T_1 . Response times increase with larger variations of vergence angle $\Delta\theta_{T1}$ and larger eccentricities, while error bars show standard error of the mean (SEM).

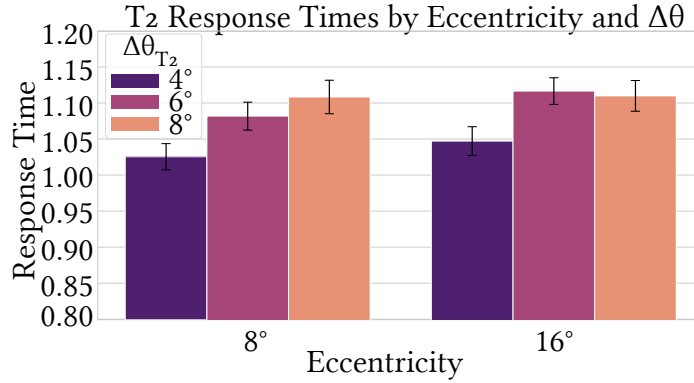


Figure 4.11: Manual response times for T_2 . Response times increase with larger variations of vergence angle $\Delta\theta_{T2}$ and larger eccentricities, while error bars show standard error of the mean (SEM).

angles: $\theta = \alpha_l - \alpha_r$ [Duinkharjav et al. 2023]. Angles are computed from the corresponding eye gaze direction vector: $\alpha = \text{atan2}(x, z)$; being x and z the horizontal and forward components, respectively, in a left-hand coordinate system and relative to head pose. Similarly, the saccade angle (α_s) is obtained from the combined gaze direction vector. In Figure 4.12, a moving average filter (kernel size of 10 samples) is applied to smooth the signal for visualization reasons. Combined with our main experiment showing effective sound-based visual depth overdriving, these results suggest that our method can not only reproduce the depth percept of the full vergence change but also improve response times in VR scenarios.

4.6 Application Case

We evaluate our method in a naturalistic application scenario resembling a driving scene and an annotation task, where text descriptions appear to provide information about objects in the environment. Here, we can apply our methodology to shift the perception of one annotation at a fixed visual location to be associated with multiple distant objects.

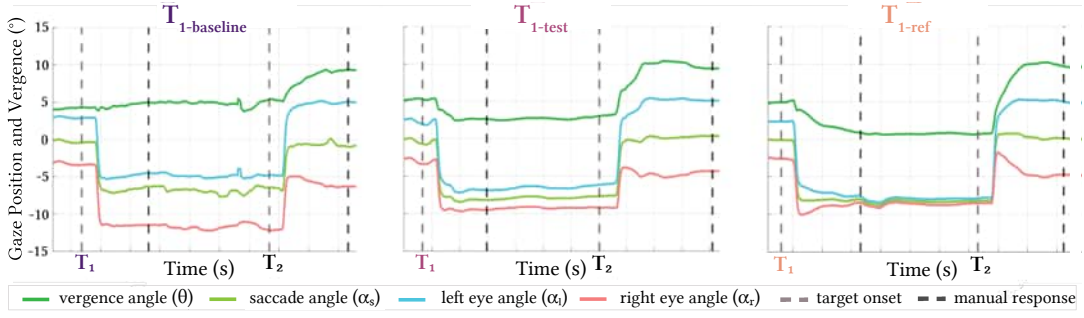


Figure 4.12: Visualization of angular eye gaze and vergence trajectories over time, as recorded by eye tracking after applying a moving average filter on the raw data (kernel size of 10 data samples). Each graph shows one experiment trial, covering the sequence of T_1 and T_2 (targets’ onsets trigger the visual response and press-button action) as illustrated in Figure 4.8 for the Divergence-Far condition. Here, the three different cases ($T_{1\text{-baseline}}$, $T_{1\text{-test}}$, $T_{1\text{-ref}}$) of the first audiovisual target T_1 ($E = -8^\circ$) are compared. In all cases, the initial fixation point is placed at 0.8 m ($\theta = 4.5^\circ$). As $T_{1\text{-baseline}}$ is located at the same depth as the initial fixation, no vergence motion is required. On the other hand, $T_{1\text{-test}}$, at 1.5 m, and $T_{1\text{-ref}}$, at 7 m, require respective divergence motions of -2° and -4° ($\Delta\theta_{T_1}$) to focus correctly at the new depths. Afterward, all cases require the corresponding convergence to focus on T_2 ($E = 0^\circ$) located at 0.43 m ($\theta = 8.5^\circ$). Note the relative differences between left and right eye angular positions showing disparity changes related to the depth being fixated on each T_1 case. Each tick in the horizontal axis represents 0.5 s.

4.6.1 Experimental Setup

Stimuli and Procedure The scene displays a low-poly city scene where the user takes the perspective of a driver with a multitude of items in the street (see Figure 4.13). The location of the scene items was chosen to cover the full visual vergence range conditions of former experiments (Table 4.1): bicycles at Divergence-Far ($E = -16^\circ$), kids at Convergence-Far ($E = 16^\circ$), cats at Divergence-Near ($E = -8^\circ$), and toys at Convergence-Near ($E = 8^\circ$). In this case, a policeman or a bee (only for very close fixations) acted as a fixation item. We chose 2D annotations as audiovisual targets (2° size), which describe the items they refer to. Auditory cues were chosen in context: a ring bell for **BIKE**, a “hello” call for **KID**, a meow sound for **CAT**, and a squeak for **TOY**. At the beginning of each trial, the corresponding fixation item was shown. Afterward, the visual annotation spawned with a lifetime of 2 s and the audio cue was played for 0.5 s. We asked participants to indicate **which of the same two scene items the annotation was referring to** by button pressing.

Conditions and Participants The visual part of the audiovisual annotation always spawns at a fixed depth, the same depth as the scene item closer to the respective fixation. For the sound, two conditions are tested: audio colocated (reference condition) and audio modulation based on the results of Section 4.4 (overdriving condition). Scene items spatial distribution, annotations visual depth, and an example of these two sound conditions related to **BIKE** category are shown in Figure 4.14. Selected Overdriving levels (ΔL) are +6 dB for **TOY**, +9 dB for **KID**, -8 dB for **BIKE** and -13 dB for **CAT**. Ten new participants (four females, six males, ages 21-34) completed all 64 randomized trials (8 conditions x 8 repetitions) in a single 15-minute session.



Figure 4.13: Representation of the driving scene from the user perspective. Annotations used as targets correspond to different item categories present in the scene (**BIKE**, **CAT**, **TOY**, **KID**). During the experiment, annotations are shown one after the other, while participants choose the scene item the annotation is likely paired with.

4.6.2 Results and Discussion

Again, we use a GLMM with a logit link function to analyze the binary dependent variable (success at associating the annotation to the item at the same visual depth), including fixed effects (E , ΔL , and vergence direction) and a random intercept to account for variability across participants. As shown in Figure 4.15, sound modulation significantly impacts user decisions when associating annotations with scene objects ($p < 0.001$). Binomial tests are computed comparing participants' performance against the chance level (50%) for each experiment condition. Details can be found in Section 7.3, Appendix C. Despite observing a clear effect of sound modulation on associating annotation with the scene item, different scenarios are found for each scene item category. For the **CAT** and **KID** annotations, participants correctly indicated the associated objects ($> 80\%$ accuracy, being both significantly different from chance level) with colocated sound. Here, the perceived depth of the annotations could be overdriven by displaced sound, thus not being significantly different from chance level ($t_{stat} = 0.6685, p = 0.5058$ for **CAT**; $t_{stat} = -0.8933, p = 0.3744$ for **KID**), meaning that the annotation is equally associated with the scene item at a different visual depth and farther from fixation. Although the **CAT** condition (Divergence, $E = -8^\circ$) is the most critical condition according to the main experiment results (Figure 4.5), it shows the expected perceptual depth overdriving behavior. Note that annotations' visual depth does not vary between colocated and overdriven conditions. Hence, sound compensation replicates the main experiment results for these two cases. Conversely, for the **TOY** and **BIKE** annotations, colocated audio reinforced position ambiguity since ratios are not significantly different from chance level ($t_{stat} = -2.0525, p = 0.0434$ for **TOY**; $t_{stat} = 0.2223, p = 0.8247$ for **BIKE**), while sound modulation allows to perceptually recover the initial larger vergence range. In the **TOY** case, the short distance between the items seems to diminish the effect of colocated sound and obscure the clearly visible depth differences provided by stereovision (in the absence of sound), as also observed in the main experiment. In the **BIKE** case, the annotation's specific on-screen placement may have biased participants, causing them to overlook consistent visual and auditory depth cues.

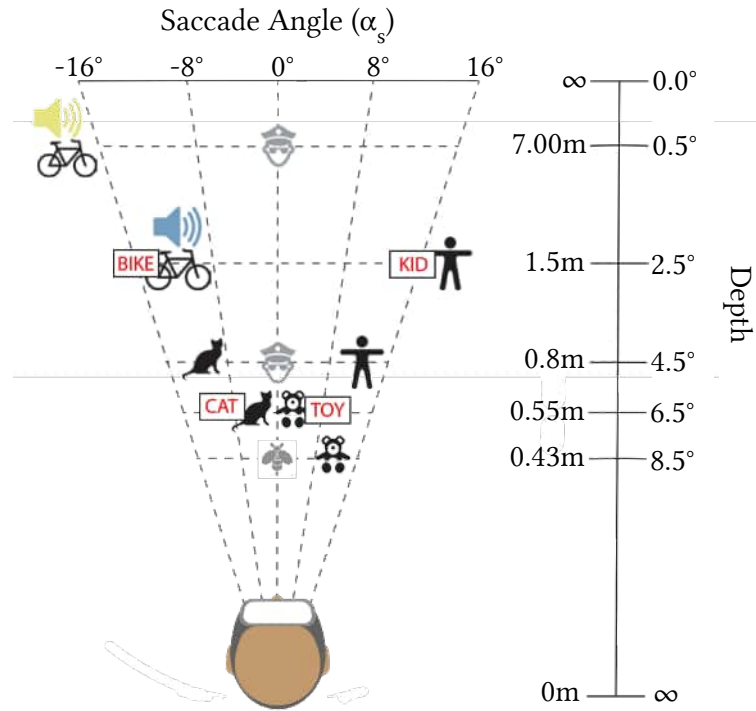


Figure 4.14: Spatial placement of the scene items and their corresponding annotations. Each annotation is positioned at the same depth as the scene item that is closer to the previous fixation, while two sound conditions are tested. For example, both the colocated (blue) and overdriving (yellow) sound conditions are represented for the bike item annotation case. As observed, annotations fall within a vergence range of 4° ($[2.5^\circ, 6.5^\circ]$), while the scene items span the initial 8° range ($[0.5^\circ, 8.5^\circ]$). Initial fixation items, such as the central policeman (Figure 4.13), are marked in gray.

In conclusion, sound modulation has been reported to be effective in two different scenarios: hiding clear visual vergence thus confusing item association as in **CAT** and **KID** cases and the main experiment; but also resolving item association when colocated sound makes it confusing, as in **BIKE** and **TOY** cases. Therefore, we observe how the overdriving effect can be influenced by the scene configuration and depends on the relative strength of audiovisual cues as postulated by depth cue fusion models [Clark and Yuille 1990; Battaglia et al. 2003; Ernst and Bühlhoff 2004].

4.7 Mixed Reality Application Version

Additionally, a similar application was implemented in Mixed Reality (MR) to explore further use cases. New scene items are used to cover the whole vergence range such as birds, toys, phones and speakers. Regarding audio cues, they are chosen according to the scene item: a bird song for **BIRD**; a squeak for **TOY**; a phone vibration for **PHONE**; and a vinyl needle skip for **SPEAKER**. As in the VR application, annotations are presented for each scene item and participants need to choose the scene item to which the annotation refers. The same audio conditions and experimental procedure are considered as explained in the VR application case (Section 4.6).

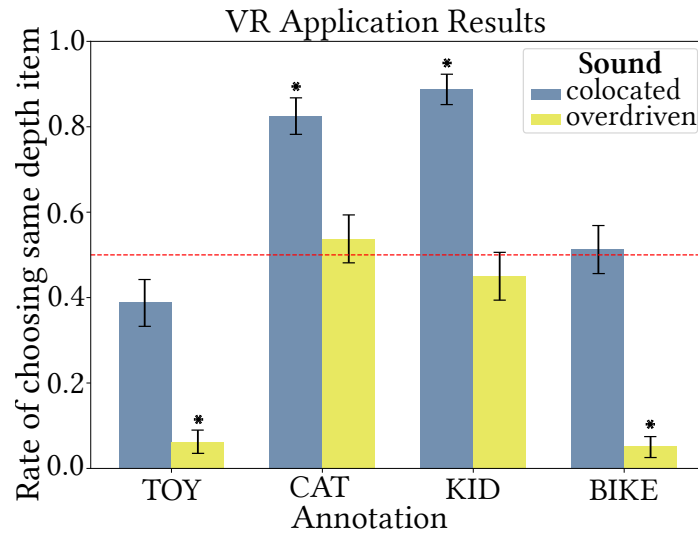


Figure 4.15: Results obtained in the application experiment. The Y-axis indicates the ratio of selecting the scene item at the same visual depth as the annotation. Low ratios indicate the tendency to perceive the annotation to be related to the scene item farther away from the fixation. Error bars show SEM and stars show statistical significance against chance level, represented by the horizontal red line.

A table head-chin rest was used to avoid users performing head rotations during the experiment, making sure that annotations appeared at the intended angles, aligned with the real-world objects. The chin-rest height was fixed to 35 cm from the scene items surface for all users, equalizing perspective and height of the annotations across participants. Different scene items were used as shown in Figure 4.16. The experimental procedure is the same as the one explained in the VR version and the new scene setup is represented in Figure 4.17.

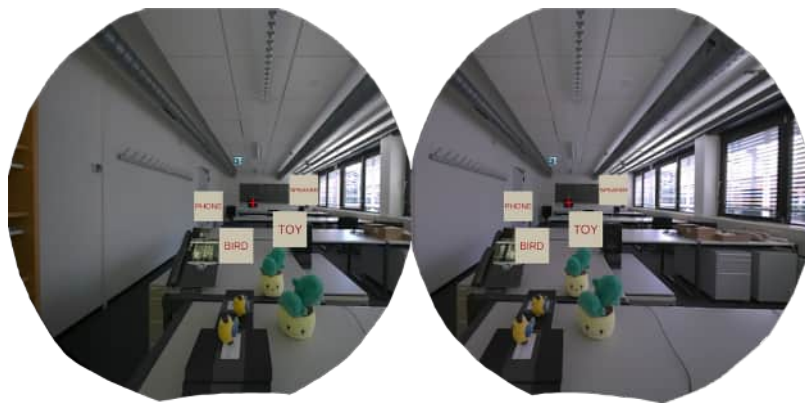


Figure 4.16: MR application version from the participant's point of view, showing left and right views. Virtual annotations' visual cues, scaled by a factor of 3 for visualization purposes, are shown within the real scene for each scene item (**BIRD**, **TOY**, **PHONE**, and **SPEAKER**), as well as the central red cross used as fixation.

In this case, five new participants (male, ages 25-33) performed the experiment and the results are displayed in Figure 4.18, showing similar trends as those reported in the VR application case. Overdriving sound is found to hide visual depth differences for **TOY**, **PHONE**, and **SPEAKER**, similarly to **CAT** and **KID** cases in the VR version.

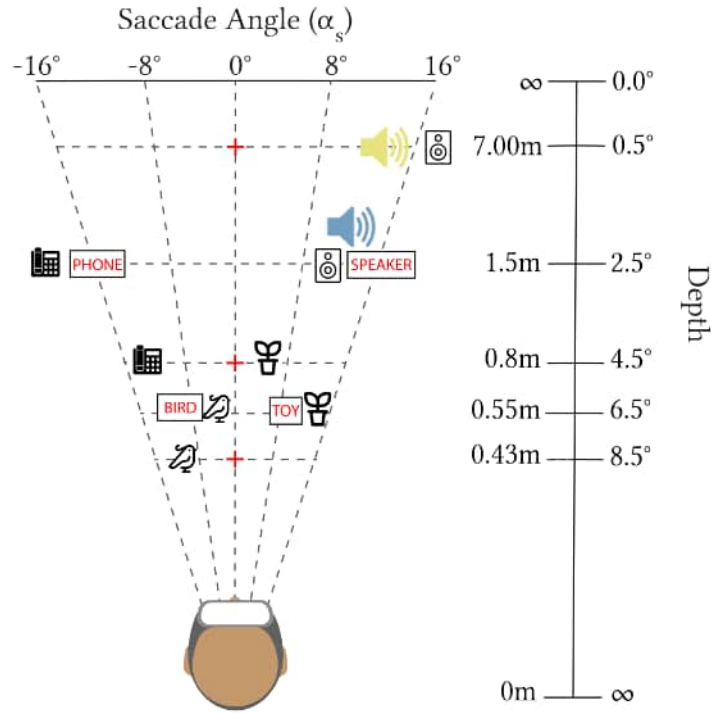


Figure 4.17: Spatial placement of the scene items and their corresponding annotations in the MR experiment. As in the VR case, colocated (blue) and overdriving (yellow) sound conditions are presented while annotations are in a single location for each scene item. Initial fixations are set by the central red crosses.

Different behavior is observed for the **BIRD** case, the closest one in depth, being consistent with the **TOY** case in VR.

Statistical analyses also confirm significant differences between sound conditions, demonstrating the potential of the overdriving effect even in different immersive scenarios such as MR. In this additional experiment, we used the Varjo XR-4 headset, whose specifications are shown in [Section 7.1](#), [Appendix A](#).

Similar depth-overdriving patterns are observed to those reported in the VR application, indicating that the effect generalises to MR environments, where only annotations are rendered content and under a less controlled scenario (luminance changes, head jitter, passthrough latency, etc.). However, it is worth pointing out some differences: while the object occupying the furthest position in the VR experiment (**BIKE**) resulted in significantly different results than the two middle ones (**KID** and **CAT**), this is no longer the case for the MR application. Similarly, the trend seen for the closest object in the MR experiment (**BIRD**) shifts towards the rest, with the accuracy rate rising for both the colocated and overdriven sound in comparison to its VR counterpart (**TOY**). This indicates more consistent behavior across all objects in the MR setup, with overdriven audio still strongly affecting perceived depth. Statistical analyses performed regarding the MR application version are shown in [Tables 7.24](#) and [7.25](#) ([Section 7.3](#), [Appendix C](#)). As can be seen, the sound level is statistically significant in annotation association, while chance level comparisons show significance on **TOY**, **PHONE** and **SPEAKER** colocated cases, even though the overdriven counterparts manage to mask visual depth.

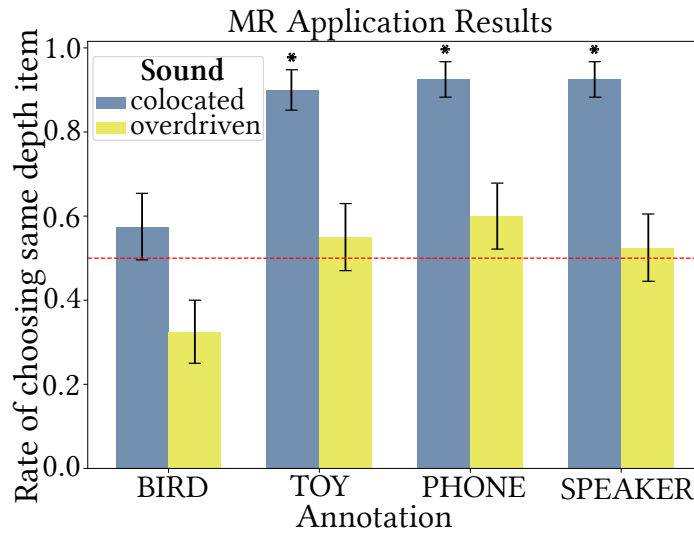


Figure 4.18: Results obtained in the MR version of the application case, displayed in the same format as its VR counterpart in Figure 4.15. Overdriving sound shows similar trends with respect to the colocated condition as in the VR application case, showing effect potential also in MR.

4.8 Limitations and Future Work

While our approach demonstrates the effectiveness of auditory cues in modulating depth perception, limitations and avenues for future exploration remain.

Statistical power and results generalizability The response time experiment (Section 4.5) includes a limited number of participants, as its goal is not to establish a novel effect but to verify that a well-documented relationship (faster responses for smaller vergence changes [Duinkharjav et al. 2023]) is preserved. This confirmatory setup allows us to assess whether vergence range differences relevant to VR, and consequently the presence of the VAC, remain associated with faster visual responses once the overdriving effect is perceptually achieved. While larger participant pools would increase statistical power, sample sizes comparable to those used in this chapter have been shown to be sufficient to reliably detect differences using similar experimental designs [Jiménez-Navarro et al. 2024a; Duinkharjav et al. 2023].

Seamless initial depth recovery In our study, we considered quickly transient targets as well as short-duration sound signals. However, for a visual target shown for a prolonged time after the audio stops, imperceptible disparity manipulations might allow for the recovery of the initial depth during extended user fixation [Kellnhofer et al. 2016a]. This would enable scene coherence restoration for a long-standing effect without any auditory cue. We have observed linear-constant disparity manipulation over the line of sight being particularly promising.

Feedback-driven recalibration and habituation effects Our experiments intentionally avoid feedback or task interaction that could induce perceptual recalibration [Lin et al. 2024], allowing us to isolate the immediate effect of audiovisual decoupling.

Although recalibration may arise in prolonged or interactive scenarios, prior work indicates that distance misperception in VR typically persists due to limited visual cues [Finnegan 2017], while relative audio-loudness differences continue to modulate visual depth perception during exposure [Bruns 2019]. This indicates that users do not automatically adapt to spatial inconsistencies in common VR settings. Our method remains particularly effective in time-critical scenarios, such as alerts or object spawning, where stimuli are brief and users have limited opportunity to recalibrate. In these contexts, the proposed method can enhance spatial interpretation and visual response without long-term perceptual side effects. Future work could explore the longevity of the effect, including cue adaptation or feedback mechanisms.

Extent of vergence changes and range We overdrive vergence angle changes of 2° between T_{ref} and T_{test} , sufficient to elicit measurable visual responses (Section 4.5). Larger vergence shifts ($> 2^\circ$) may be harder to overdrive by sound. Although visual response speedup could increase further, bringing the visual target closer to the fixation depth while the intended location remains farther away makes full concealment by auditory cues unlikely. Also, while near and far vergence regions were studied to explore the effectiveness of the overdriving effect towards the vergence range limits, future research could investigate how it may behave in mid-range regions for completeness.

Occlusions and scene composition Visual occlusions or shading effects as additional visual depth cues are a natural limitation of our approach since they indicate the relative distance positioning of scene elements and overall environment composition. In those contexts, altering visual depth through binocular disparity and eye vergence without considering other existing depth cues may disrupt these relationships, potentially leading to implausible scene interpretations.

Impact on long-term fatigue According to the vergence ranges considered (Table 4.1), our technique inherently reduces vergence magnitude when fixating on newly appearing objects closer to the HMD focal distance (Section 4.6), offering clear potential to mitigate the VAC. This mitigation could be especially beneficial during prolonged VR sessions, where sustained VAC significantly contributes to discomfort. However, further systematic studies are needed to evaluate our technique’s impact on long-term fatigue.

4.9 Conclusion

In this chapter, we investigate the impact of auditory cues on stereoscopic depth perception in VR. Specifically, we demonstrate that stereoscopic depth perception can be overdriven through controlled auditory manipulations based on maximizing sound-based depth cues while carefully ensuring that the integrity of audiovisual cues is preserved. As expected, stronger auditory manipulations induce more consistent depth overdriving, highlighting the potential of multimodal integration as a complementary approach to existing visual-only correction methods. The overdriving effect and its modulation levels are validated in a realistic scenario, showing how the perceived depth range can be effectively extended.

Furthermore, by reducing vergence demands through auditory overdriving, faster depth adjustments and gaze realignment are enabled. These findings provide perceptual guidance for the design of dynamic audiovisual interfaces in AR and VR, where components may be distributed across varying depths. A more detailed exploration of these aspects could be an interesting direction for future research.

Chapter 5

Dynamic Visual Dominance

This chapter addresses how visual dominance influences the underlying process of how visual inputs are fused, giving more weight to one eye’s view over the other. We systematically characterize dynamic visual dominance across the visual field, with a particular focus on peripheral vision, where perceptual asymmetries are most relevant for stereoscopic rendering. First, we show that tolerance to eye-asymmetric blur at the fovea under binocular viewing depends on gaze direction, confirming that eye dominance is not spatially invariant. Then, we show that peripheral dominance is primarily governed by retinal eccentricity, with consistent naso-temporal asymmetries and dominance reversals around the blind spots. We leverage our insights in a dominance-contingent rendering application, where additional blur is selectively applied to the perceptually non-dominant eye regions under binocular viewing.

5.1 Introduction

Human visual perception emerges from the fusion of the two retinal images captured by the eyes, while binocular fusion denotes the neural mechanism that combines these two signals into a single percept. In stereoscopic vision, a clear percept arises when corresponding retinal locations in both eyes receive similar visual information. Small interocular differences, however, can be compensated without conscious awareness, enabling perceptual stability despite local mismatches between the two inputs. Similar to handedness, most individuals exhibit ocular dominance, where one eye contributes more strongly to the integrated percept. Approximately, 70 – 80% of the population is right-eye dominant [Chaurasia and Mathur 1976]. During binocular fusion of corresponding retinal regions, signals from the dominant eye receive greater perceptual weight. As a consequence, attenuating high-frequency content in the non-dominant eye, such as through display-side blurring, can still yield a clear fused percept. Computer graphics has exploited this property by interpreting blur as a reduced sampling density requirement in ray tracing or variable rate shading (VRS).

Classical gaze-contingent rendering relies on the non-uniform distribution of photoreceptors across the visual field and reduces sampling density in peripheral regions to improve computational efficiency [Patney et al. 2016]. Extending this concept, prior

work has investigated eye dominance for enhanced efficiency in stereoscopic displays [Meng et al. 2020; Wang et al. 2023a]. These approaches apply a fixed blur increment to the non-dominant eye and assume that eye dominance is static in magnitude and constant in eye assignment. However, eye dominance is not a fixed characteristic but instead a dynamic feature that adapts with perceptual context [Prummer et al. 2025b].

As shown in Figure 5.1, we investigate and quantify how eye dominance shifts according to gaze position and eccentricity. Our psychophysical experiments show that the dominance is more dynamic and adaptive than commonly assumed in graphics applications, offering key insights for advancing stereoscopic gaze-contingent rendering. First, our results reveal how foveal eye dominance depends on gaze position in pure visual acuity tasks. That is, when focusing on a specific region, the same eye will always be dominant in this region, making the pre-computation of rendering effects possible. As expected, stimuli presented in the right visual field tend to be dominated by the right eye, and vice versa. Second, as stimuli shift towards the periphery, eye dominance and its strength depend mainly on eccentricity. Based on the results, we derive a dominance map that guides peripheral blur levels relatively to the observer’s gaze position.

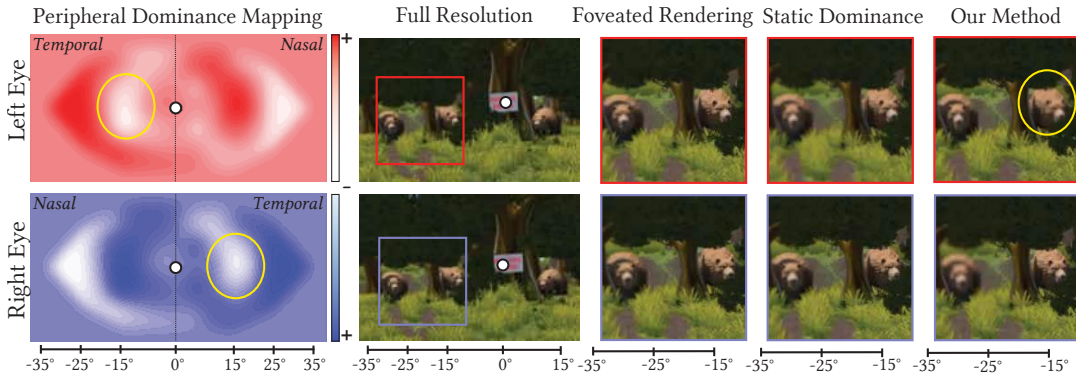


Figure 5.1: We show that visual dominance is not a static, eye-fixed percept, but varies dynamically between the eyes as a function of gaze direction and peripheral offset from the fovea. *Left*: Our retinal dominance maps for the left and right eyes obtained from our psychophysical experiments. The central white markers indicate the fixation point, and the bottom scale denotes retinal eccentricity. Red and blue colors indicate the level of dominance toward the left and right eyes, respectively. As can be observed, the temporal viewing hemifield (corresponding to the nasal retinal region) consistently dominates over the nasal viewing hemifield (corresponding to the temporal retinal region), with the exception of the blind spots (dashed circles). *Right*: We leverage the map to introduce asymmetric blur for foveated rendering. With standard foveated rendering as reference, the first approach applies a static dominance, introducing additional blur in the left eye only, assuming right-eye dominance. Our method, shown in the last column, accounts for dynamic dominance changes by applying additional blur according to the dynamic dominance maps of each eye. While the static approach applies a stronger blur to the nominally non-dominant left eye, the dominance maps reveal a local dominance switch in this region, which can lead to visible artifacts. By leveraging gaze-dependent, dynamically varying visual dominance, our approach enables stronger localized blur with reduced artifacts.

Our findings offer potential implications for real-time stereoscopic rendering. In this direction, we apply our dominance map through rendering shaders to show the potential of enhancing foveated rendering with improved fidelity and reduced display power consumption. Our contributions are summarized as follows:

- We characterize visual dominance shifts in the fovea as a function of gaze direction, and show that tolerance to eye-asymmetric blur under binocular viewing exceeds binocular blur detection thresholds, particularly when applied to the non-dominant eye.
- We show that transitions in peripheral dominance are driven by naso-temporal asymmetry and the eyes' blind spots.
- We apply our insights to a proof-of-concept gaze-contingent rendering application, improving over a static dominance alternative.

To foster future research, our collected anonymized data is publicly available at <https://dynamicvisualdominance.mpi-inf.mpg.de>.

5.2 Related Work

5.2.1 Stereoscopic Perception and Fusion

Stereoscopic perception (stereopsis) arises from the coherent integration of the two retinal images into a single percept [Cang et al. 2023; Jiménez-Navarro et al. 2026b], a process known as binocular fusion [Sperling 1970]. Binocular fusion is possible only when visual signals fall on corresponding retinal locations, a condition that is typically satisfied when the object lies within or close to the focused depth [Sperling 1970]. The overlapping binocular field, within which both retinas receive information from the same region of space, spans approximately 110° horizontally [Stidwill and Fletcher 2017]; outside this region, vision is monocular and cannot support stereopsis.

While the human visual system can tolerate small interocular differences in sharpness, larger blur disparities typically induce interocular blur suppression, whereby the sharper retinal image dominates the percept by suppressing high-frequency content from the blurred eye [Georgeson and Wallis 2014]. When interocular differences become too pronounced, stable fusion may break down and transition into binocular rivalry, characterized by temporal alternations between monocular percepts [Arnold et al. 2007]. At the same time, controlled differential blur can act as an important depth-of-field cue that assists binocular correspondence and reduces rivalry in complex scenes [Hoffman and Banks 2010].

5.2.2 Eye Dominance

Eye dominance describes the preferential weighting of one eye's input during binocular integration of discrepant images and primarily affects overlapping binocular regions. Reported statistics suggesting that approximately 70–80% of the population is right-eye dominant are typically derived from tasks with centrally positioned stimuli, where the notion of eyedness [Chaurasia and Mathur 1976] remains stable in

high-precision frontal positioning tasks such as target aiming, and shows no consistent alignment with hand or foot preference [Carey and Hutchinson 2013; Chaurasia and Mathur 1976].

Subsequent work demonstrates that this static notion of eye dominance is incomplete. Khan and Crawford [2001] showed that *sighting dominance* (SD) switches dynamically between the eyes depending on horizontal gaze angle in a reach-grasp task: For observers with right-eye dominance, dominance reversals occurred when targets were placed in the left visual hemifield (i.e., on the side opposite to the dominant eye) at eccentricities exceeding 15.5° , while observers with left-eye dominance exhibited corresponding switches for targets placed in the right visual hemifield. These gaze-dependent dominance changes are consistent with relative image size effects favoring the eye closer to the target [Banks et al. 2004]. Such gaze-dependent behavior was confirmed across standard clinical tests by Quartley and Firth [2004]. Complementary studies on *sensory eye dominance* (SED), where discrepant content is directly projected to corresponding retinal locations, report a naso-temporal asymmetry favoring the temporal visual hemifield (nasal retina) at eccentricities above 20° [Fahle 1987; Fahle and Schmid 1988], and below 4° [Stanley et al. 2011, 2019]. These effects are commonly attributed to asymmetries in photoreceptor and ganglion cell density and cortical representation, with more recent work reporting deviations at eccentricities below 10° [Paffen 2025], substantial inter-subject variability [Chaumillon et al. 2017], and differences between laboratory and real-world viewing conditions [Candy and Cormack 2022].

Recently, the sighting dominance experiments of Khan and Crawford [2001] were replicated in a VR setting [Prummer et al. 2024], confirming that eye dominance varies dynamically during eye-hand coordination tasks with stimulus alignment. In contrast, our VR experiments target *acuity eye dominance* (AED) in purely visual tasks, focusing on asymmetric blur detection under binocular viewing. We study dominance variations across gaze directions, analogous to SD, and across retinal eccentricities via discrepant stimuli, analogous to SED. This focus on blur detection directly targets a key perceptual factor limiting the visibility of asymmetric distortions in stereoscopic imagery and naturally aligns with exploiting eye dominance in foveated rendering, as well as with related applications such as asymmetric stereoscopic video encoding [Meegan et al. 2001; Seuntiens et al. 2006] for VR/AR streaming and cloud gaming.

5.2.3 Foveation and Eye Dominance in Rendering

Retinal photoreceptor density decreases from the fovea toward the periphery [Curcio and Allen 1990], leading to reduced visual acuity and increased tolerance to image blur [Watson and Ahumada 2011]. This behavior is commonly modeled by extending the contrast sensitivity function to peripheral vision [Peli et al. 1991; Mantiuk et al. 2022], forming the basis of foveated rendering, where sampling density follows peripheral sensitivity falloff [Weier et al. 2017]. Building on this, numerous real-time gaze-driven methods modulate spatial image resolution [Guenter et al. 2012; Patney et al. 2016; Li et al. 2021; Ye et al. 2022], often incorporating local image statistics [Tursun et al. 2019] or visual saliency [Fan et al. 2024; Stengel et al. 2016; Krajancich et al. 2023]. To mitigate artifacts in heavily blurred peripheral regions, prior work synthesizes high-frequency noise [Tariq et al. 2022], reconstructs plausible detail from sparse, temporally accumulated samples using generative models [Kaplanyan et al. 2019], or compensates for losses in motion perception [Tariq and Didyk 2024].

Beyond these monocular considerations, binocular vision offers additional opportunities for perceptual optimization.

Eye dominance and reduced blur sensitivity of the non-dominant eye have therefore been exploited to further accelerate foveated rendering through more aggressive yet perceptually acceptable reductions in sampling density. [Meng et al. \[2020\]](#) introduce kernel foveated rendering (KFR), parameterizing peripheral blur with a single control value tuned to remain imperceptible in the non-dominant eye. [Wang et al. \[2025\]](#) demonstrate that this approach is overly conservative and improves performance by partitioning the field of view into multiple eccentricity-dependent zones for each eye. [Prummer et al. \[2025b\]](#) were the first to postulate exploring dynamic eye dominance for foveated rendering. Motivated by this idea, we derive visual dominance maps that reveal spatially varying dominance as a function of retinal position relative to the fixation point, indicating that strategies assuming fixed dominance may be suboptimal.

5.3 Experiment 1: Dynamic Dominance in Fovea

In this section, we explore sighting dominance ([Section 5.2.2](#)) by studying how dynamic visual dominance influences monocular blur suppression across different fixation locations. Participants performed a blur discrimination task in which controlled amounts of blur were applied either binocularly or monocularly, while fixation was constrained to predefined locations within the headset’s binocular region. Starting from standard binocular detection thresholds for foveal blur as a baseline, we analyze how these thresholds change under monocular blur in different fixation locations. Our results suggest that monocular blur detection thresholds are influenced by both the eye to which the blur is presented and gaze direction.

5.3.1 Experimental Setup

Hardware As in the MR application version presented at the end of [Chapter 4](#), we used a Varjo XR-4 headset with an integrated eye-tracker (200 Hz and 1° accuracy). A chin rest was employed to fix head position and stimuli were generated in screen-space, thereby avoiding head motion or jitter. During the experiment procedure, participants remained seated throughout the session. Rendering resolution of the display was set to 51 PPD. To study visual dominance, we focus on binocularly visible visual field locations, i.e., the overlap region seen by both eyes. To ensure consistency across participants, this overlap region was kept constant by fixing the headset interpupillary distance (IPD) to 64 mm. As a result, the stereoscopic region considered in this experiment spans from -25° to 25° in the visual field, similar to previous works [[Kergaßner et al. 2025](#)]. This overlap region is depicted in [Figure 5.2](#).

Stimuli Following previous work in blur detection and discrimination [[Watson and Ahumada 2011](#)], the visual stimulus employed was a vertical sinusoidal pattern defining an edge between two luminance contrasts [[Mather and Smith 2002](#)]. This pattern provides a combination of vertical and horizontal cues, becoming closer to real scenes than single-dimensional edges. The stimulus pattern was generated at runtime using a GLSL shader following the sine wave equation along the vertical axis:

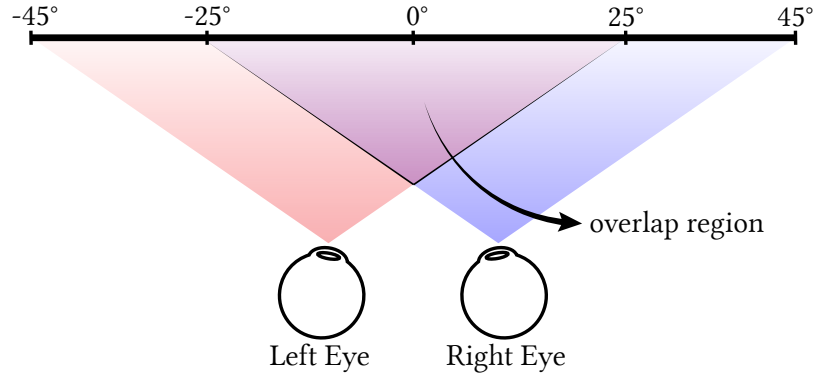


Figure 5.2: Overlap region considered in our study where stereo viewing is available. Each eye individually sees a FOV of 70°, while the common central region is constrained to 50°, 25° from the center of the screen [Kergaßner et al. 2025]. Considering the monocular regions, the effective total FOV of the HMD is 90°.

$$f(y) = 0.5 + A \cdot \sin(2\pi\omega \cdot y + \phi) \quad (5.1)$$

where A is the amplitude of the sinusoid, ω the spatial frequency in cycles per texture UV unit, and ϕ the phase. The values employed are: $A = 0.27$, $\omega = 2$ (cycles displayed), and $\phi = 0$. To achieve the desired stimuli contrast [Mather and Smith 2002], the two regions at each side of the border were assigned grayscale intensity values of 0.8 and 0.2, implemented as linear RGB triplets. The sinusoidal pattern (Figure 5.3, top-right) was dichoptically shown in both eyes, while blur levels applied to each eye were independent. This was achieved by using a second post-processing GLSL shader that applied a stereoscopic Gaussian blur kernel.

The bidirectional (vertical and horizontal) Gaussian kernel was implemented by first computing the spatial weight of each pixel as a function of its offset from the kernel center, and then normalizing the weighted sum to obtain the final blurred pixel appearance, as described by the following equations:

$$w(i, j) = \exp\left(-\frac{i^2 + j^2}{2\sigma^2}\right) \quad (5.2)$$

$$G(x, y) = \frac{1}{\sum_{i=-r}^r \sum_{j=-r}^r w(i, j)} \sum_{i=-r}^r \sum_{j=-r}^r F(x + i, y + j) \cdot w(i, j) \quad (5.3)$$

where σ denotes the standard deviation of the Gaussian kernel, r is the kernel radius, and $F(x, y)$ and $G(x, y)$ represent the input and blurred images, respectively.

The visual stimulus subtended 5° of visual angle, covering the entire foveal region [Kergaßner et al. 2025]. To ensure that blur differences were visible, the stimulus texture resolution was set to 70 PPD. Texture downsampling to display resolution is done via mipmapped bilinear filtering, ensuring no aliasing effects impact the displayed blur level. The scene background was rendered with an intensity value of 0.5. Examples of blurred stimuli are shown in Figure 5.3 (right).

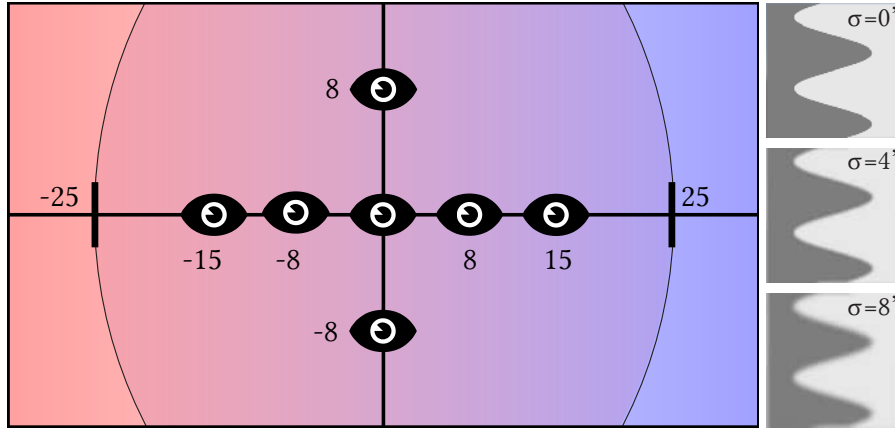


Figure 5.3: *Left*: Illustration of the fixation locations considered in Experiment 1, together with the limits of the binocular region of the headset. *Right*: The visual stimulus employed, with blur levels of $\sigma = \{0', 4', 8'\}$.

Conditions The main factor of this experiment is the *blur mode*, defined as $mode = \{Binocular, Left, Right\}$. In Binocular mode, both eye views receive the same blur level, while in Left and Right modes, blur is applied only to the left or right view, respectively. We consider three *blur levels* for all modes $\sigma = \{0.5', 1', 2'\}$. Since monocular blur is harder to detect, three additional blur levels $\sigma = \{4', 8', 16'\}$ are included for Left and Right modes. To test different gaze directions, we asked participants to fixate on the visual stimulus placed at different positions in the visual field, always within the stereoscopic region of the headset. We define fixation locations along the vertical and horizontal axes of the screen and denote them as (pos_{hor}, pos_{ver}) , expressed in degrees of visual angle relative to the center of the screen, where positive values indicate rightward and upward offsets. We include the following *fixation locations* $= \{(0, 0), (\pm 8, 0), (\pm 15, 0), (0, \pm 8)\}$. These locations follow typical fixation areas after visual saccades, whose amplitudes reach up to 15° without involving head motion [Bahill et al. 1975a]. We employ a fully factorial design with $3 \text{ modes} \times 7 \text{ fixation locations} \times 3 \text{ or } 6 \text{ blur levels } (\sigma) = 105 \text{ conditions}$. Each condition was repeated six times, yielding 630 trials per participant.

Participants Twelve participants took part in the study (eight male, four female, ages 20-26). All reported normal or corrected-to-normal vision. Only two of them were found to be left-dominant after performing a slight variation of the Miles test [Miles 1930], where participants used their own hands to create the hole to look through. Such a right-dominance ratio closely represents the real population distribution. From the participant pool, 83.3% reported having used VR equipment 5 times or less and only one participant had never used it before.

Procedure Participants performed a two-alternative forced-choice (2AFC) task comparing a blurred stimulus against an unblurred *Reference* ($\sigma = 0'$). In each trial, participants viewed two *Test Cases* in sequence and selected the one that appeared more similar to the Reference. One of the two Test Cases always matched the Reference exactly, while the other contained the blurred stimulus under evaluation. In this task, chance-level performance therefore corresponds to blur levels at which participants cannot consistently distinguish between the blurred and unblurred stimuli.

The order of these two Test Cases was randomized on every trial. Each trial followed the same temporal sequence: Reference $\rightarrow$ Test Case 1 $\rightarrow$ Reference $\rightarrow$ Test Case 2, with the user being able to switch between them. A warning message asking the participant to make a decision was displayed after 10 s or if the full sequence was repeated three times. Participants indicated their choice via button press. The system accepted responses only while participants maintained fixation on the stimulus, as verified by the eye-tracking data. We chose this temporal presentation to guarantee precise fixation without introducing spatial offsets that would engage peripheral vision. Each experimental session lasted approximately one and a half hours.

5.3.2 Results and Discussion

We compute success ratios based on the proportion of trials in which participants selected the unblurred Test Case, which is identical to the Reference. Higher success ratios indicate that blur is more reliably detected, reflecting stronger perceptual differences. For each condition, we aggregate responses across trials and participants, and use these ratios to estimate blur detection thresholds. We remove response-time outliers by excluding trials with response times outside the 1st–99th percentile range. We then fit a logistic psychometric function (Equation (5.4)) defined over the logarithm of the blur level $\log_{10}(\sigma)$, parameterized by a location parameter α , slope β , and lapse rate λ . The guess rate is fixed to $\gamma = 0.5$ to match our experimental design. This function is commonly used in perceptual studies to model detection and discrimination performance, as it captures the characteristic sigmoidal relationship between stimulus strength and response probability while allowing for chance-level performance in forced-choice tasks.

$$P(x) = \gamma + (1 - \gamma - \lambda) \frac{1}{1 + \exp\left(-\frac{\log_{10}(x) - \alpha}{\beta}\right)}. \quad (5.4)$$

Since fitting individual curves would be unstable due to limited data, we first pool responses across observers [Patney et al. 2016; Wolski et al. 2022; Bernal-Berdun et al. 2025; Wang and Kristensson 2026]. To verify that pooled fits reflect consistent behaviour, we assessed the monotonic relationship between blur level and detection probability using Spearman’s ρ . In our results, 89% of participant–condition cases exhibited $\rho > 0$, 71% exceeded $\rho > 0.5$, and 11% showed $\rho < 0$, supporting the pooled fits as a stable population-level estimate. The fitted curves for horizontal (avg. RMSE: 0.034 and R^2 : 0.913) and vertical (avg. RMSE: 0.031 and R^2 : 0.932) fixation locations are shown in Figures 5.4 and 5.5, respectively.

As expected, all curves originate at the chance level (50%), consistent with the 2AFC task. We derive detection thresholds at the 75% performance level by obtaining the blur value at which each fitted curve intersects the 75% point (green dotted line). The detection thresholds are shown in Figure 5.6 and a visualization of raw data is provided in Figure 5.7.

As shown in the results, blur detection thresholds in the Binocular condition remain below $1'$, consistent with previous findings in blur detection studies [Watson and Ahumada 2011]. These thresholds do not vary with gaze direction, indicating that foveal acuity remains stable across the stereoscopic region of the field of view considered. This binocular condition, therefore, provides a stable baseline against which monocular effects can be interpreted.

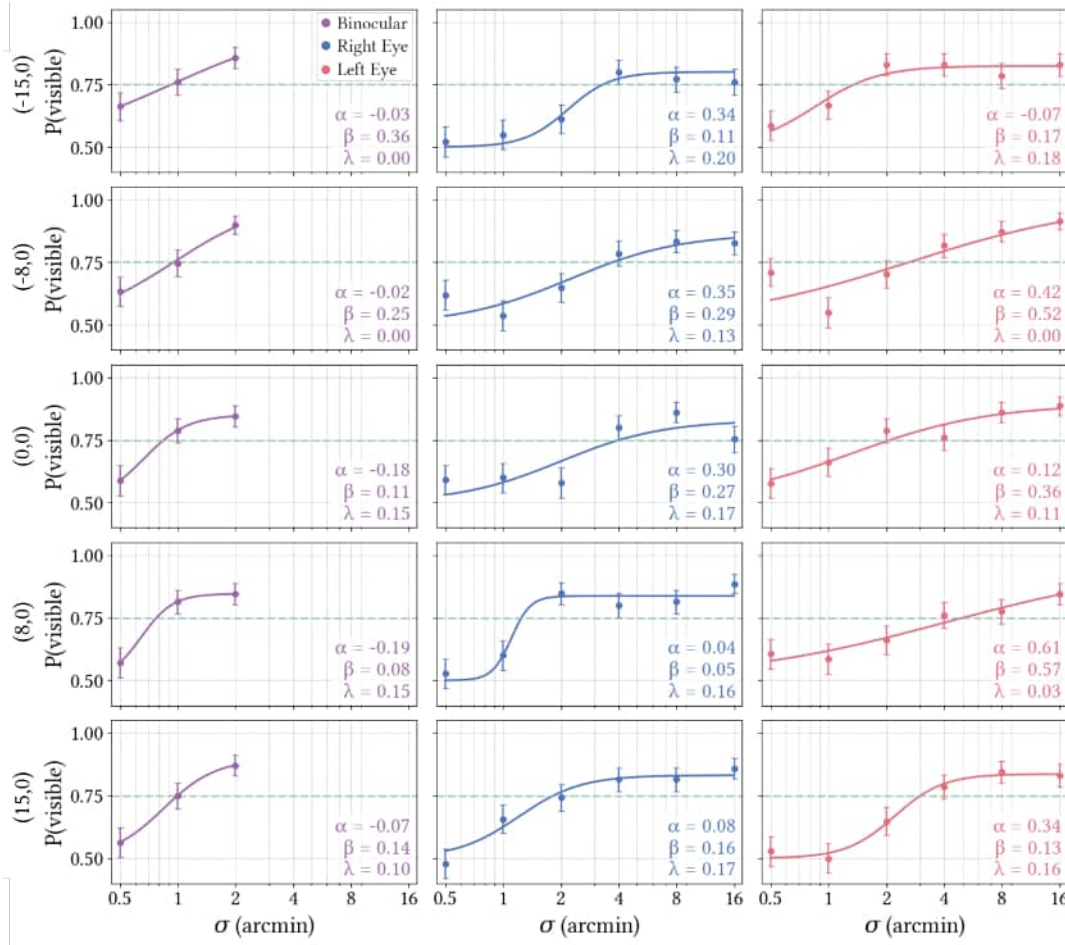


Figure 5.4: Psychometric curves from Experiment 1 (horizontal fixation locations) showing the mean success rate $P(\text{visible})$ as a function of blur strength σ (arcmin), averaged across participants and trials (error bars denote SEM). The horizontal dashed line indicates $P = 0.75$, corresponding to reliable blur detection. The x-axis is shown on a logarithmic scale with ticks at the tested σ levels (0.5, 1, 2, 4, 8, and 16 arcmin). Parameter boxes report the fitted values of α , β , and λ .

When blur is applied to a single eye, detection thresholds increase relative to the binocular baseline. Interestingly, these monocular conditions reveal a dynamic visual dominance with a clear dependence on gaze direction. As shown in Figure 5.6 (top), blur detection thresholds for the left eye increase when participants fixate toward the right side of the screen, and vice versa. This suggests that blur tolerance increases for blurry inputs on the eye that does not fully cover the corresponding side of the field of view (e.g., left eye on the right side). In other words, dynamic visual dominance modulates monocular blur suppression as a function of gaze direction.

In our data, the transition between left and right eye dominance occurs at approximately 3–4° to the right of the screen center. Although most participants exhibit right-eye dominance overall, we observe higher blur tolerance in the Right condition when fixation is centered. This transition point differs slightly from prior reports in hand–eye alignment tasks, which place the dominance switch around 5° to the left [Prummer et al. 2024], suggesting that dominance transitions may depend on task demands and stimulus properties. Despite this shift, our thresholds are consistent

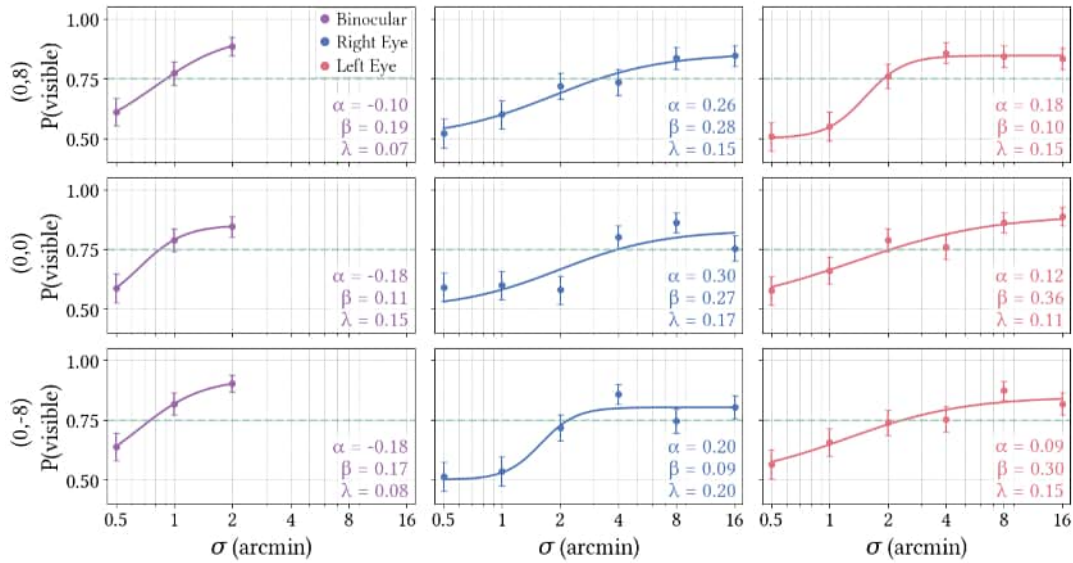


Figure 5.5: Psychometric curves from Experiment 1 (vertical fixation locations) showing the mean success rate $P(\text{visible})$ as a function of blur strength σ (arcmin), averaged across participants and trials (error bars denote SEM). The horizontal dashed line indicates $P = 0.75$, corresponding to reliable blur detection. The x-axis is shown on a logarithmic scale with ticks at the tested σ levels (0.5, 1, 2, 4, 8, and 16 arcmin). Parameter boxes report the fitted values of α , β , and λ .

across conditions, with thresholds below $1'$ for binocular blur [Watson and Ahumada 2011]; thresholds between $1'$ and $3'$ for the dominant eye; and larger thresholds between $3'$ and $5'$ for the non-dominant eye.

For vertical fixation locations, we hypothesized no systematic modulation of visual dominance as both eyes symmetrically cover the vertical extent of the visual field. The results in Figure 5.6 (bottom) qualitatively support this expectation, showing little apparent change in blur detection thresholds. Although the $(0, -8)$ case shows no clear dominance, further exploration is needed to confirm this behavior.

In summary, we show for the first time that monocular blur detection thresholds are not only elevated relative to binocular viewing, but are also systematically modulated by gaze direction and by which eye is blurred.

5.4 Experiment 2: Dynamic Dominance in Periphery

After studying how foveal visual dominance varies with gaze direction, we now extend our analysis to peripheral vision. The goal now is to determine whether similar dominance shifts occur for stimuli presented at different retinal eccentricities relative to the current gaze direction, as in sensory eye dominance experiments (Section 5.2.2). Participants performed a blur discrimination task in which blur was always present but monocularly applied, either to the left or right eye only. While participants fixated at predefined locations, we presented the stimulus at different peripheral positions within the stereoscopic region, and asked them to select the condition that appeared more similar to the unblurred Reference. This design allows us to characterize peripheral visual dominance by comparing preference between left- and right-eye blur as a function of gaze direction and stimulus eccentricity.

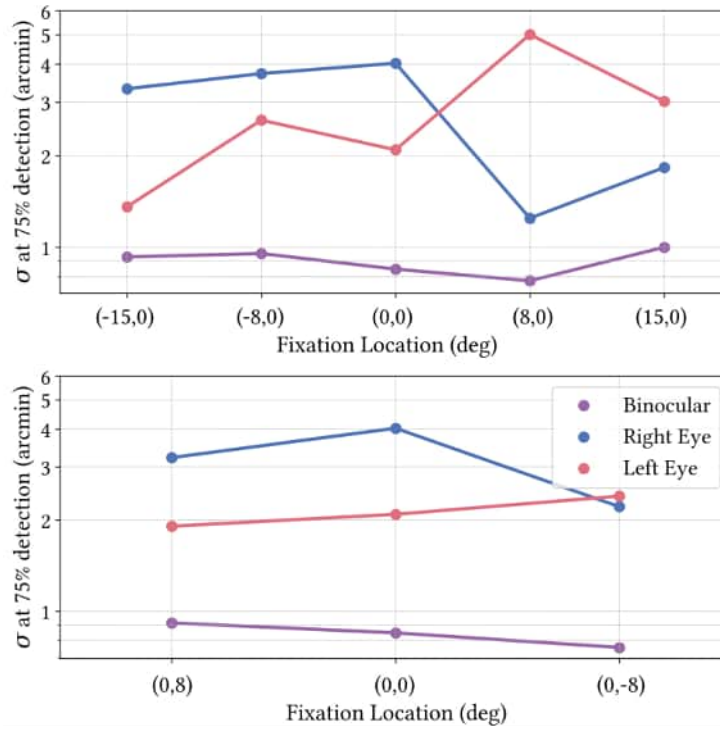


Figure 5.6: Blur Detection Thresholds for horizontal (*top*) and vertical (*bottom*) fixation locations (Experiment 1). The X-axis shows locations from the screen center denoted as (pos_{hor}, pos_{ver}) , expressed in degrees of visual angle. For horizontal locations, a higher threshold is measured for the left eye on the right side of the screen, indicating right-eye dominance. Conversely, left-eye dominance is observed on the left side. Unlike horizontal locations, vertical locations do not flip visual dominance significantly. Instead, eye dominance remains the same as in the central location. Regarding binocular blur, detection thresholds do not strongly depend on the gaze direction.

Conditions Following the same notation (pos_{hor}, pos_{ver}) , we define three fixation locations along the horizontal axis such that $fixation\ location = \{(-15, 0), (0, 0), (15, 0)\}$, expressed in degrees of visual angle relative to the screen center. To study visual dominance in the periphery, stimulus locations are also presented at discrete locations relative to the screen center, ensuring stereoscopic viewing. We define the stimulus locations as $stimulus\ location = \{(\pm 5, 0), (\pm 10, 0), (\pm 15, 0), (\pm 20, 0), (0, \pm 5), (0, \pm 10), (0, \pm 15), (\pm 5, \pm 5), (\pm 10, \pm 10)\}$, where coordinates denote horizontal and vertical offsets in degrees of visual angle. All stimulus locations considered are displayed in [Figure 5.8](#). We employ a fully factorial design with $3\ fixation\ locations \times 22\ stimulus\ locations = 66$ conditions. Each condition is repeated six times, yielding 396 trials per participant.

Procedure The visual stimulus remains the same as described in [Section 5.3.1](#), and we follow a similar sequential presentation. In this experiment, both stimulus presentations (Test Case 1 and Test Case 2) contain blur, but each presentation shows the stimulus blurred in either the *Left* or *Right* view. The participants' task was to select the Test Case that appeared more similar to the unblurred *Reference*. To ensure that monocular blur remains detectable in the periphery, we set the applied blur well above the foveation-dependent detection threshold.

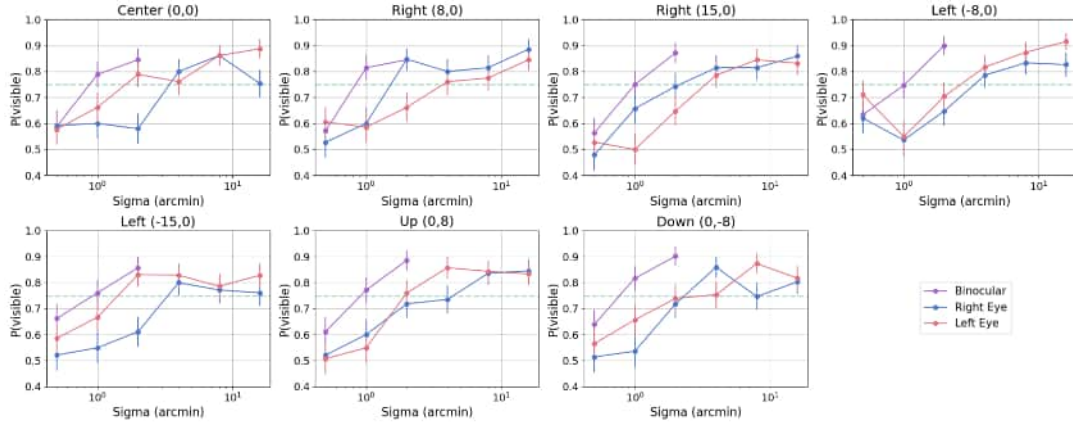


Figure 5.7: Raw data used to fit the psychometric functions used to compute the blur detection thresholds for the fovea (Experiment 1, [Section 5.3](#)).

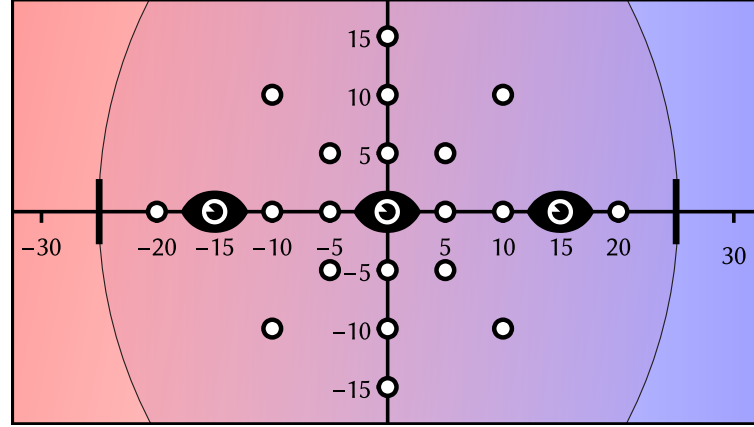


Figure 5.8: Stimuli and fixation locations considered in Experiment 2. Three fixation locations are considered, being at (0,0), (15,0) and (-15,0). For each fixation location, stimulus locations covering the headset overlapping region are represented by dots.

We define the applied blur as: $\sigma_{\text{applied}}(E) = \sigma_{\text{foveation}}(E) \cdot 1.5$; where: $\sigma_{\text{foveation}}(E) = 0.3 \cdot \text{Eccentricity}(\circ)$. Such a blur increase ratio follows established foveated rendering guidelines for peripheral blur [[Mantiuk et al. 2022](#); [Patney et al. 2016](#)]. To guarantee that judgments rely on peripheral perception, participants had to fixate on a cross. Whenever gaze deviated from the intended fixation location, the stimulus disappeared and responses were not accepted.

Participants A total of 7 participants performed the study (4 male, 3 female, ages 21-27) and all of them reported normal or corrected-to-normal vision. In this case, four participants reported having used VR equipment 5 times or less, and only one participant had never used it before.

5.4.1 Results and Discussion

For each condition, we compute the proportion of trials in which participants selected the Right (right eye blurred) or Left (left eye blurred) condition as being more similar to the unblurred Reference. These ratios reflect relative monocular dominance.

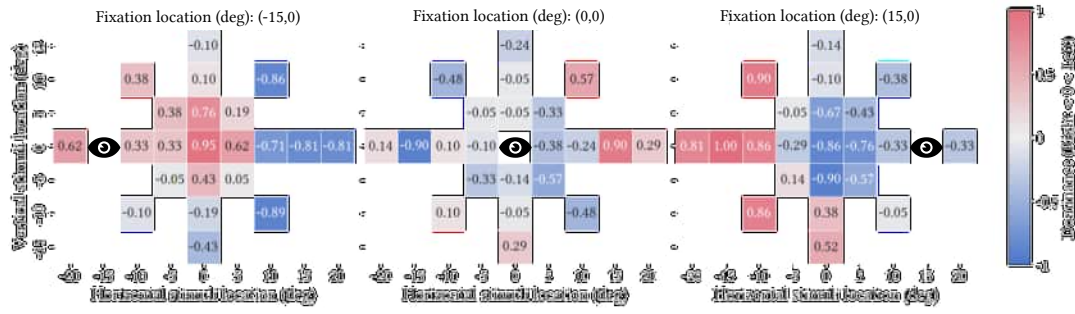


Figure 5.9: Results obtained in Experiment 2. Different fixation locations are represented: *left*, fixation at (0,-15); *center*, fixation at (0,0); *right*, fixation at (0,15). Ratios show monocular dominance by the means of the preferred input being blurred. Values close to 1 show left eye dominance (right eye blur preferred), while values close to -1 show right eye dominance. Axes represent locations in screen space, expressed in degrees of visual angle.

Therefore, positive values indicate a preference for right-eye blur, negative values indicate a preference for left-eye blur, and larger magnitudes correspond to greater blur tolerance in the preferred eye. The resulting dominance ratios are shown in Figure 5.9, grouped by fixation location.

As shown in these results, the dominant view shifts systematically across the peripheral locations within the overlapping region. Consistency across participants is quantified by correlating each participant's response map with the leave-one-out group mean, obtaining strong agreement (mean Pearson $r = 0.71$, $SD = 0.11$; mean Spearman $\rho = 0.69$). This suggests that the spatial dominance pattern is largely shared across observers, despite local variability.

When the fixation location is at (0,0), stronger dominance shifts appear around eccentricities of $\pm 15^\circ$, corresponding to the blind spots of the left and right eyes, respectively. At peripheral locations closer to the fixation, no strong dominance toward either eye is observed, although a slight preference for left-eye blur emerges, consistent with the predominance of right-eye dominance across participants. With fixation location shifted to the left (-15, 0), left-eye dominance appears for small eccentricities around the fovea until the right-eye blind spot at $+15^\circ$ from fixation. Beyond this point, dominance switches to the right eye and persists toward the periphery, up to the maximum eccentricity tested. A symmetric pattern emerges when fixation shifts to the right (15, 0), with right-eye dominance near the fovea and a switch to left-eye dominance beyond the left-eye blind spot. These dominance reversals in peripheral regions starting at $E = 20^\circ$ reflect the naso-temporal asymmetry in binocular vision, where input from the nasal retina dominates over that from the temporal retina at large eccentricities [Fahle 1987; Fahle and Schmid 1988]. The blind spots around $\pm 15^\circ$ act as transition points that trigger local dominance switches, producing a continuous and coherent binocular percept across the visual field.

Retinal Dominance Map To build a unified Retinal Dominance Map, we merged data from all three fixation conditions. As shown in Figure 5.9, visual dominance shifts consistently, with blind-spot transitions occurring at $\pm 15^\circ$ relative to the fovea regardless of gaze direction. This suggests that, within the tested stereoscopic range, peripheral dominance is primarily driven by retinal eccentricity of the stimulus with

respect to the fixation location. Based on this observation, we integrate the data across fixation conditions by transforming stimulus positions to fovea-centered coordinates $(\Delta x, \Delta y) = (s_x - f_x, s_y - f_y)$, pooling samples with identical retinal offsets. For those relative coordinates sampled by multiple gaze locations, agreement is lower near the fovea ($E \leq 10^\circ$: mean $SD = 0.45$) than in the periphery ($10 < E < 20$: mean $SD = 0.15$), as can be observed in Figure 5.9. Therefore, since visual dominance in relative locations close enough to the fovea is affected by gaze direction, such locations are excluded from the peripheral map. For the remaining locations, cross-fovea consistency is tested via a neighbor-based prediction from nearby positions measured under other fixation conditions. Observed and predicted dominance values strongly correlate (Pearson $r = 0.70, p < 0.001$), indicating coherent spatial structure, thus supporting the merging and interpolation. The resulting peripheral dominance map is shown in Figure 5.10 and used in our tested application (Section 5.5.1).

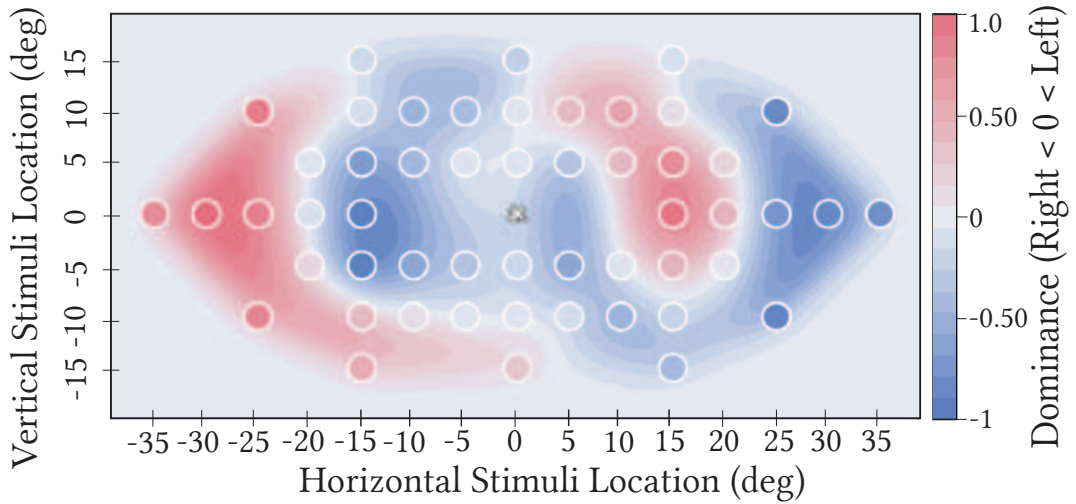


Figure 5.10: Peripheral dominance map obtained by interpolating the data obtained from the eccentricity values considered in Experiment 2 (Figure 5.9). Stimulus locations are relative to the fovea (white cross) and expressed in degrees of visual angle. Small horizontal eccentricities ($E = \pm 5^\circ, \pm 10^\circ$) are not combined, as their proximity to the fovea makes gaze angle influence visual dominance.

5.5 Application Cases

We propose potential applications for the dynamic visual dominance phenomenon. Our findings are used to establish a visual dominance map in the continuous domain to guide gaze-contingent rendering while accounting for dominance shifts. In addition, we estimate computing savings by developing a desktop-based Variable Rate Shading (VRS) prototype and devising a combined map that relates visual dominance to both fixation location and stimulus eccentricity.

5.5.1 Guiding Visual Dominance-Contingent Rendering

Our dominance map from Section 5.4 can be used to add additional blur levels to the peripheral non-dominant regions of each eye. As in gaze-contingent approaches, we focus on peripheral dominance potential while the foveal region is unmodified.

To do so, we first calculate dense peripheral dominance values in a 2D spatial grid using cubic interpolation following a Clough-Tocher scheme [Alfeld 1984; Farin 1986]. Such a 2D spatial grid allows us to estimate ocular dominance, labeled as $d(s_{x,y})$, for any stimulus location within the stereoscopic field (Figure 5.10), where $d \in [-1, 1]$ indicates the visual dominance level. To perform the interpolation, we leverage our experiment-acquired dominance ratios (Section 5.4) as sparse input conditions ($d(s_{x,y}^i)$). The resulting interpolated map allows continuous peripheral dominance transition effects driven by retinal eccentricity. By sampling the map in real time, dynamic dominance can be considered through higher foveation blur in the non-dominant regions of each eye. We adapt the rendering pipeline to apply our dominance map as gaze-contingent foveation blur levels. To perceptually evaluate our method, we use post-processing shaders in Unity. Following the experimental procedure of previous sections, we use standard foveated rendering as baseline, where $\sigma_{left} = \sigma_{right} = \sigma_{foveation}$ (Section 5.4). We implemented static dominance rendering, where blur increments are statically applied in the non-dominant input (left eye) [Wang et al. 2023a, 2025], as an alternative to our method. Specifically, we apply: $\sigma_{left} = 1.5 \cdot \sigma_{foveation}$; $\sigma_{right} = \sigma_{foveation}$. The 1.5 multiplication factor of the non-dominant eye refers to previous work [Meng et al. 2020]. For our method, we derive eye-dependent blur intensity maps from the peripheral dominance map $d(s_{x,y})$ (Figure 5.10). The left-eye blur map $b_l(s_{x,y})$ is obtained by reparameterizing the negative dominance range $[-1, 0]$ to $[1, 0]$, such that higher values indicate increased blur applied to the left eye in regions dominated by the right eye. Conversely, the right-eye blur map $b_r(s_{x,y})$ maintains the positive dominance range $[0, 1]$, assigning higher values to left-eye dominated regions so extra blur is applied in the right view. Thus, $b_{l,r}(s_{x,y}) \in [0, 1]$. Blur kernel $\sigma_{l,r} \in [1, 3]$ is then set as follows:

$$\sigma_{l,r}(s_{x,y}) = 2 \cdot \min(1, \max(2 \cdot b_{l,r}(s_{x,y}) - 1, 0)) + 1. \quad (5.5)$$

The sampling strategy considered to transform dominance blur maps $b_{l,r}(s_{x,y})$ into blur kernel levels $\sigma_{l,r}(s_{x,y})$ is shown in Figure 5.11. This transformation could be optimized in production to increase the non-dominant gains while remaining imperceptible, which would require considering the (dichoptic) peripheral blur detection thresholds in the non-dominant view. Nevertheless, the stereoscopically applied blur is: $\sigma_{applied} = \sigma_{l,r}(s_{x,y}) \cdot \sigma_{foveation}$. Non-dominant regions are dynamically sampled in real-time from the peripheral dominance map based on the eye-tracked gaze. The masks generated by sampling the dominance maps obtained from our interpolated results are shown in Figure 5.12. In the colored areas, additional blur is applied in those regions where the corresponding eye is non-dominant to the final percept, such that imperceptible artifacts are introduced.

As a result, the non-dominance factor ranges in the $[1, 3]$ interval, where 1 corresponds to the dominant regions and 3 to the strongest non-dominant regions. We limit this factor to 3 to test the capabilities of our method by doubling the factor considered in the static dominance approach, and to not apply an excessive blur amount that may cause diplopia or discomfort. Please refer to Section 5.6 for further discussion.

In a perceptual study, we compare our method and the static eye dominance approach (extra foveation applied to the same eye, usually the left one) with standard foveated rendering as *Reference*. Eleven participants (nine male, two female, ages 21–35) took part in the study. The task was to choose which method is perceived closer to the *Reference* in terms of blur and image resolution.

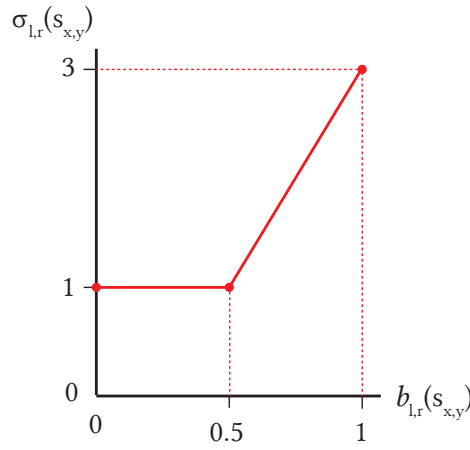


Figure 5.11: Visual representation of the transformation between blur maps ($b_{l,r}(s_{x,y})$) and the blur kernel ($\sigma_{l,r}(s_{x,y})$) performed in the application.

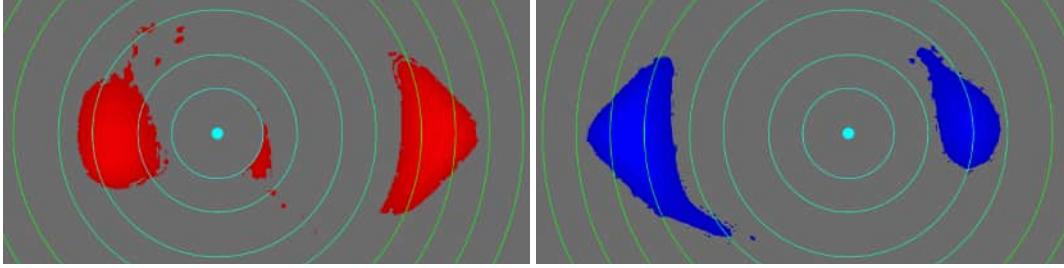


Figure 5.12: Our peripheral dominance mask applied as additional blur in foveation. Coloured regions show non-dominance zones for each eye where additional blur is applied. Red regions show non-dominance regions for the left eye, while blue regions show non-dominance regions for the right eye. Each green line represents an increment of 5° from fixation, while the stereoscopic region always spans 25° from the screen center.

In every tested scene, participants were instructed to fixate on some specific region and focus on: comparing the lion shields (*Sponza*); the bears in the background (*Nature*); and the back of the seated man (*2D Image*). Nevertheless, free viewing was allowed. Regions outside the overlapping region were blacked out to ensure attention control [Kergaßner et al. 2025]. Screenshots of every scene and experiment results are shown in Figure 5.13.

Our method is significantly preferred over the static eye dominance approach, with an 83.8% mean selection rate. A t-test analysis confirmed that all scenes: *Sponza* ($t = 5.043, p < .005$); *Nature* ($t = 6.640, p < .0005$); and *2D Image* ($t = 4.605, p < .005$), show selection rates significantly different from chance level. Additionally, five of those participants repeated the study with full-resolution rendering as *Reference*. Even in such a case, our method remains preferred over the static dominance approach ($rate = 0.844, SEM = 0.055, t = 6.254, p < .001$). Since our aim is to show that visual dominance behaves dynamically, no other foveation methods that consider additional variables (e.g., contrast or saliency) are compared with our approach.

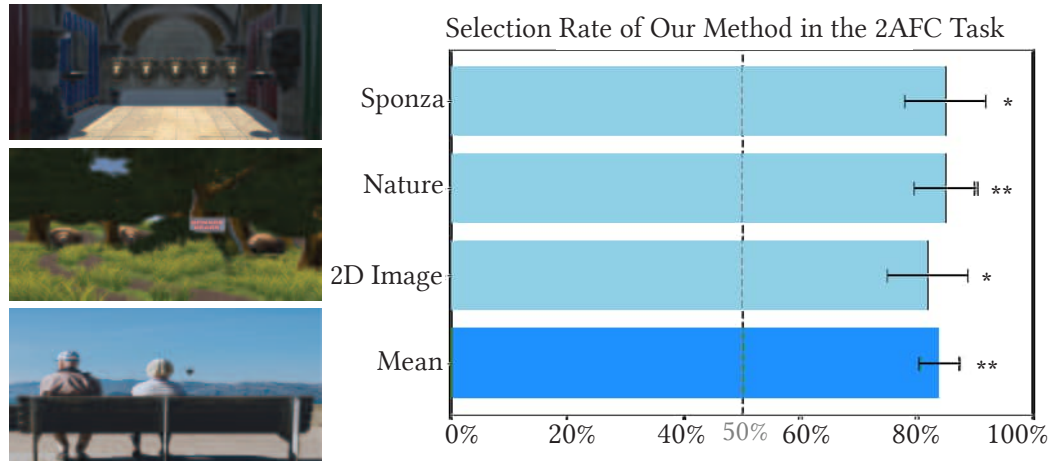


Figure 5.13: Results obtained in our application case. Rates show a preference for choosing our dynamic dominance gaze contingent method over static dominance in the three scenarios considered. Error bars show standard error of the mean (SEM) and stars indicate significance (* < .005, ** < .0005) compared to chance level (50%).

5.5.2 Computational Savings Estimation in a VRS Prototype

A desktop-based variable rate shading (VRS) prototype has been implemented by using the Unity VRS API, which allows us to estimate the computational benefits of using our peripheral dominance map. Display resolution is set at 2550×2550, matching the stereo region of the headset. We encode the dominance map as a shading rate image, assigning corresponding shading rates per peripheral location. By targeting the complete non-dominant view, we compare applying a 4×4 shading rate in the strongest dominance regions with a uniform 1×1 shading rate as a baseline. When more stereo regions are monocularly dominated (e.g., Figure 5.9, right), we obtain around 15% rendering savings. Similar gains are found with 4K or 8K resolution since the mask is constant in screen space. While gains depend on gaze direction and monocularly dominated regions within stereoscopic regions, larger savings are expected in more complex shading scenes, amplifying VRS benefits [Jindal et al. 2021].

The scene employed to evaluate computational savings in the VRS prototype is the one used in the VRS API demo⁷. In this environment, a shading rate image (SRI) is used to encode shading rates in screen space by means of a rendering feature. The peripheral dominance map is provided as SRI and thus converted to dynamic shading rates based on the eccentricity location from the fixation point for the complete non-dominant view. In the tested scene, only a volumetric renderer feature simulating fog takes advantage of VRS. Therefore, increased savings would be expected with more complex shaders implementing additional features or global illumination. The SRI used to estimate the reported savings is displayed in Figure 5.14. Notably, in production, the monocularly dominated region on the left could be extended until the end of the stereoscopic region, maximizing computation savings.

⁷<https://github.com/Unity-Technologies/shading-rate-demo>

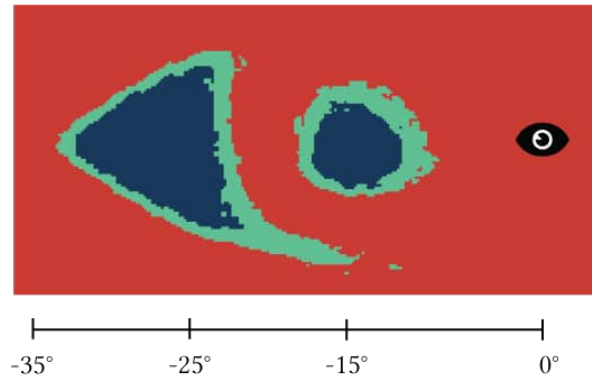


Figure 5.14: Shading Rate Image (SRI) used to estimate computational savings in the prototype desktop-based VRS application. When fixating at (15, 0) expressed in degrees of visual angle relative to the screen center in screen space, the displayed monocularly-dominated areas in the left periphery fall in the stereoscopic region of the headset. Red color corresponds to 1×1 shading rate (SR), while green and blue colors show 2×2 and 4×4 SR, respectively. The horizontal axis shows eccentricities in degrees of visual angle from the fixation location, displayed with the eye icon.

5.5.3 Combined Dominance Map

While our application primarily exploits peripheral dominance effects for visual dominance-contingent rendering (Section 5.5.1), a more complete dominance representation would be beneficial for a broader range of rendering strategies. Applications such as eye-asymmetric rendering require an understanding of how dominance varies across the visual field, from the fovea to the periphery, in order to adapt visual quality continuously as gaze and stimulus position change. Our foveal (Section 5.3) and peripheral (Section 5.4) findings provide complementary evidence toward this goal as both experiments show consistent dominance trends. Thus, Figure 5.15 shows a conceptual dominance representation by relating the fovea-stimuli horizontal offset.

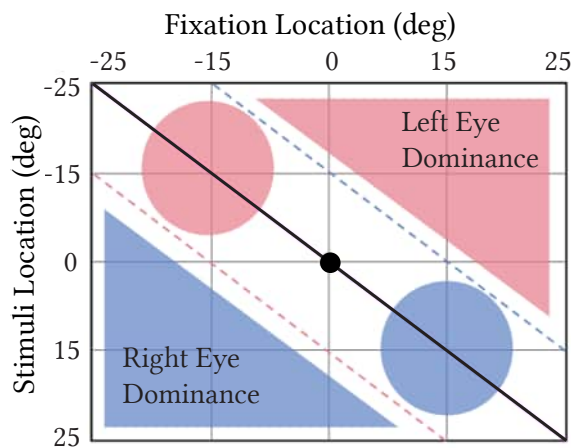


Figure 5.15: Dominance representation along the horizontal dimension for different combinations of fixation and stimuli locations. Red and blue colors show left and right eye dominance, respectively, and geometric shapes are representative. The diagonal line shows fixating at the stimuli and axes are expressed in degrees of visual angle from the screen center.

5.6 Conclusion and Future Work

In this chapter, we characterize dynamic visual dominance in stereoscopic vision, focusing on how asymmetric interocular blur is tolerated across the visual field. Through two complementary psychophysical experiments, we show that blur binocularly applied within the foveal region is detected more readily than blur applied monocularly. We further show that blur detection thresholds vary with gaze direction (fovea) and retinal eccentricity (periphery), revealing dynamic visual dominance even in basic visual acuity tasks. Beyond characterizing this phenomenon, we show how peripheral dominance can be leveraged in an application case for dominance-contingent rendering, where increased blur is selectively applied to the non-dominant eye. Despite the scope of our findings, several limitations remain and motivate future research directions.

To begin with, while non-dominant peripheral regions tolerate substantially higher blur than dominant ones, the upper bound of acceptable monocular blur and its interaction with binocular rivalry onsets (see [Fahle 1982]) were not explored in this study. Characterizing this trade-off could further extend perceptually guided computation savings. Furthermore, in our current application, dominance transition zones are treated conservatively to ensure spatiotemporal stability and avoid visual artifacts. Future work could study the temporal dynamics of these transitions and explore strategies to safely exploit them. In addition, integrating the proposed application into a full foveated rendering or ray-tracing pipeline would allow a precise quantification of performance gains. Finally, visual dominance may extend to other forms of dynamic visual effects. For instance, prior work has explored peripheral color modulations for display power savings [Duinkharjav et al. 2022b], exploited change blindness effects [Groth et al. 2023b], and showed changes in speed perception [Peng et al. 2025; Scholz et al. 2024]. Whether similar dominance-driven tolerance exists for those attributes remains an interesting direction for future research. In this context, our preliminary exploration of peripheral dominance in color space through stereoscopic color variations suggests that dynamic dominance may also be leveraged along these dimensions.

For the first time, our findings show that visual dominance in stereoscopic vision dynamically modulates monocular blur tolerance across both foveal and peripheral vision.

Chapter 6

Conclusions and Future Work

In natural environments, audiovisual cues are semantically related and aligned in both time and space. Immersive virtual environments, which are the focus of this thesis, allow these signals to be deliberately manipulated. Temporal alignment can be precisely adjusted by advancing or delaying auditory signals relative to a visual event, while spatial discrepancies can be introduced with minimal constraints. By bringing sensory manipulation to the visual domain alone, different content can also be presented to each eye through stereoscopic displays on the VR headsets. This differs from real-world conditions, where both eye views are consistent and accurately represent the scene. Throughout this dissertation, manipulations are ensured to remain within the range where multisensory or binocular cues can still be integrated by the brain into a coherent percept.

When multiple misaligned cues are present, the perceptual system must infer the most plausible scene configuration. This thesis investigates how perception can be guided through carefully designed combinations of sensory cues. In [Chapter 3](#) and [Chapter 4](#), we examine specific spatio-temporal manipulations of audiovisual cues aimed at reducing human reaction time. In [Chapter 5](#), similar manipulation principles are applied exclusively to the visual modality, where the impact of dichoptically introducing local blur is investigated as a means of reducing the computational cost. Within this framework, eye dominance is shown to be a substantially more complex phenomenon than the commonly assumed notion of a single, fixed dominant eye.

The following concluding sections synthesize and discuss the findings of the presented chapters, alongside potential directions for future research.

6.1 Visual Saccades Latency Acceleration

In [Chapter 3](#), we demonstrate that auditory modality can enhance visual response performance, evidencing facilitatory effects of multimodal interaction. In this chapter, one sensory input (vision) benefits from another (hearing), extending previous findings on audiovisual integration. Specifically, we quantify how this facilitation modulates saccadic latency as a function of the spatiotemporal shifts between the audiovisual cues and the stimulus eccentricity.

Saccadic latency decreases monotonically with increasing audiovisual temporal shifts in which the auditory signal precedes the visual cue, reaching a plateau at approximately 400 ms. Notably, this latency reduction remains achievable even when the auditory cue is semantically unrelated to the visual stimulus. Although audiovisual cues are naturally aligned, some real-world cases may exhibit such spatiotemporal misalignments originating from different sources. Such relationships could unintentionally reduce reaction times if they are neglected. For example, a referee blowing a whistle near a basketball player may inadvertently accelerate the visual reaction of the player who will receive the pass, which is a plausible real-world occurrence.

The primary contribution of this chapter is the predictive saccade latency model within the eccentricity range over which purely visual responses are elicited without head motion, relating saccadic latency to audiovisual spatiotemporal offsets. Validation results show that saccadic latency prediction for eccentricities within our range differs from eye-tracker measurements by less than 10 ms.

Future work could extend this approach by combining sound-based acceleration with the previously reported influence of pictorial cues on saccadic latency [Duinkharjav et al. 2022a], or by exploring whether vertical saccades similarly benefit from auditory-based latency reduction. Although deliberately designing temporally misaligned auditory cues requires additional effort, application domains such as e-sports or time-critical visual systems (e.g., flight control or driving assistance) may benefit from these findings.

6.2 Sound-(over)driven Depth Perception

We extend audiovisual interaction further in [Chapter 4](#), where auditory modality not only enhances visual performance but can also replace visual information in depth perception tasks. In our studies, we demonstrate that auditory depth cues can overdrive or compensate for conflicting visual depth information, effectively restoring the original depth percept even in the presence of contradictory visual signals. This suggests that our brain can integrate and prioritize sensory inputs according to their relative reliability. In this direction, relative loudness differences can facilitate the depth discrimination task compared to degraded visual depth cues. Consequently, original visual depth variations can be substituted by auditory information, restoring the initial depth perception and enabling faster visual responses.

In contrast to [Chapter 3](#), where auditory signals were employed to reduce visual saccadic reaction time in 2D displays without depth variation, [Chapter 4](#) focuses on speeding up visual responses to depth variations during combined eye movements. Specifically, we aim to perceptually control the final depth perception through auditory cues that overdrive visual depth cues. To this end, the manipulation performed focuses on two components of the auditory signal: the spatial location and the frequency spectrum. For convergent eye movements, the spatial location of the audio source is adjusted not only considering audiovisual fusion limits constraints but also maximizing interaural differences (IAD). On the other hand, regarding divergent eye movements, the sound's frequency spectrum is adjusted accordingly. The results obtained demonstrate how auditory cues can bias depth perception to the extent that they overdrive contradictory visual depth information arising from vergence movements and binocular disparity. Consequently, by reducing the visual depth variation between stimuli and compensating it with the corresponding auditory manipulation,

the visual combined response is accelerated, as shown in the second experiment reported in [Chapter 4](#), while the vergence variation remains perceptually unnoticed.

Further research could explore a wider range of vergence amplitudes susceptible to auditory overdriving, as well as alternative sound manipulation techniques. Rather than maximizing IADs under convergence conditions, future approaches could instead modulate the spatial location of the audio source according to the target's visual eccentricity. This strategy would imply not varying IADs but keeping the lateralization of the audio source constant, thus varying only audio loudness. It would help address pure vergence visual movements without any change in eccentricity, disentangling auditory depth cues for a more precise characterization of auditory influence on depth perception.

6.3 Dynamic Visual Dominance

Lastly, we investigate how visual information is processed when a single sensory modality is involved. Focusing on the HVS and purely visual cues, stereovision provides two distinct inputs, with each eye contributing independently to the final visual representation. In [Chapter 5](#), we examine the relative contribution of each eye to the final percept across fixation locations in the central visual field and at varying eccentricities. This analysis allows us to map visual dominance at different regions in the periphery, revealing local dominance shifts and well-defined monocularly dominated areas. Based on these findings, we build a retinal dominance map that can be used to guide gaze-contingent rendering techniques, applying stronger foveation levels at those peripheral regions where the corresponding eye is non-dominant. This reduces computational requirements through lower-resolution rendering in those non-dominant regions without any perceptible artifacts, improving user experience compared to static eye dominance approaches.

As there is no hearing modality involved, dichoptic visual manipulations are exclusively employed through the perceptual studies performed in this chapter. While still focusing on the central region of the screen to fixate on, thus without involving head motions, the complete stereoscopic region of the headset is systematically sampled to construct the retinal dominance map. Within this space, visual stimuli are rendered with asymmetric blur applied across the two eyes, allowing us to quantify the relative contribution of each eye to the final percept and characterize local dominance variations in the periphery.

Next steps could consider how mismatches between the headset IPD and the observer's real IPD may affect dominance estimation in the overlapping region. Another direction could focus on characterizing binocular rivalry thresholds within the monocularly dominated peripheral regions, aiming to reduce potential discomfort. Moreover, it remains to be determined whether the retinal dominance map generalizes to other visual attributes, such as color processing. In monocularly dominated areas, accurately rendering the desired color only in the dominant view while introducing controlled color manipulations in the non-dominant view could reduce power consumption without perceptible degradation. Such color foveation driven by dynamic visual dominance could extend the practical impact of these findings in real-time rendering systems.

6.4 Concluding Remarks

Overall, this dissertation investigates how the brain selects and integrates sensory information to form a coherent percept, leading to measurable improvements in both visual performance and rendering efficiency. Across three complementary projects, we model how one sensory modality facilitates another ([Chapter 3](#)), demonstrate how one sensory modality can overdrive conflicting information conveyed by another ([Chapter 4](#)), and how competing inputs from the same sensory modality are weighted to determine dominance and the final percept ([Chapter 5](#)). These findings contribute to a deeper understanding of (1) the effects of multimodal interaction on audiovisual perception and (2) visual input processing mechanisms under stereovision, guiding the design of immersive environments toward improved performance and user experience.

Bibliography - Own Work

- Daniel Jiménez-Navarro, Colin Groth, Kenneth Chen, Qi Sun, Karol Myszkowski, Hans-Peter Seidel, and Ana Serrano. 2026a. Dynamic Visual Dominance in Stereoscopic Foveation. In *ACM SIGGRAPH 2026 Conference Papers*. 1–11.
- Daniel Jiménez-Navarro, Colin Groth, Xi Peng, Jorge Pina, Qi Sun, Praneeth Chakravarthula, Karol Myszkowski, Hans-Peter Seidel, and Ana Serrano. 2026b. Overdriving Visual Depth Perception via Sound Modulation in VR. *IEEE Transactions on Visualization and Computer Graphics* (2026).
- Daniel Jiménez-Navarro, Xi Peng, Yunxiang Zhang, Karol Myszkowski, Hans-Peter Seidel, Qi Sun, and Ana Serrano. 2024a. Accelerating saccadic response through spatial and temporal cross-modal misalignments. In *ACM SIGGRAPH 2024 Conference Papers*. 1–12.
- Daniel Jiménez-Navarro, Ana Serrano, and Sandra Malpica. 2024b. Minimally disruptive auditory cues: their impact on visual performance in virtual reality. *The Visual Computer* (2024).
- Xi Peng, Colin Groth, Daniel Jiménez-Navarro, Zihao Zou, Yan Zhu, Ana Serrano, Karol Myszkowski, Qi Sun, and Praneeth Chakravarthula. 2025. Why Slow Feels Fast and Fast Feels Slow: Evaluating and Predicting Speed Misperception. In *2025 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 1075–1084.
- Xi Peng, Yunxiang Zhang, Daniel Jiménez-Navarro, Ana Serrano, Karol Myszkowski, and Qi Sun. 2024. Measuring and predicting multisensory reaction latency: A probabilistic model for visual-auditory integration. *IEEE Transactions on Visualization and Computer Graphics* (2024).

Bibliography

- Rachel Albert, Anjul Patney, David Luebke, and Joohwan Kim. 2017. Latency requirements for foveated rendering in virtual reality. *ACM Transactions on Applied Perception (TAP)* 14, 4 (2017), 1–13.
- Peter Alfeld. 1984. A trivariate clough—tocher scheme for tetrahedral data. *Computer Aided Geometric Design* 1, 2 (1984), 169–181.
- V. Ralph Algazi, Richard O. Duda, Dennis M. Thompson, and Carlos Avendano. 2001. The cipc hrtf database. In *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No. 01TH8575)*. IEEE, 99–102.
- James Anliker. 1976. Eye movements-On-line measurement, analysis, and control. In *Eye Movements and Psychological Processes*.
- Elena Arabadzhiyska, Okan Tarhan Tursun, Karol Myszkowski, Hans-Peter Seidel, and Piotr Didyk. 2017. Saccade landing position prediction for gaze-contingent rendering. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 1–12.
- Derek H. Arnold, Philip M. Grove, and Thomas SA. Wallis. 2007. Staying focused: A functional account of perceptual suppression during binocular rivalry. *Journal of Vision* 7, 7 (2007), 7–7.
- Derek H. Arnold, Alan Johnston, and Shinya Nishida. 2005. Timing sight and sound. *Vision Research* 45, 10 (2005), 1275–1284.
- A. Terry Bahill, Deborah Adler, and Lawrence Stark. 1975a. Most naturally occurring human saccades have magnitudes of 15 degrees or less. *Investigative Ophthalmology & Visual Science* 14, 6 (1975), 468–469.
- A. Terry Bahill, Michael R. Clark, and Lawrence Stark. 1975b. The Main Sequence, a Tool for Studying Human Eye Movements. *Mathematical Biosciences* 24, 3-4 (1975), 191–204.
- Robert W. Baloh, Andrew W. Sills, Warren E. Kumley, and Vicente Honrubia. 1975. Quantitative measurement of saccade amplitude, duration, and velocity. *Neurology* 25, 11 (1975), 1065–1065.
- Martin S. Banks, Tandra Ghose, and James M. Hillis. 2004. Relative image size, not eye position, determines eye dominance switches. *Vision Research* 44, 3 (2004), 229–234.
- Peter GJ. Barten. 1999. *Contrast sensitivity of the human eye and its effects on image quality*. SPIE press.
- Peter Battaglia, Robert Jacobs, and Richard Aslin. 2003. Bayesian integration of visual and auditory signals for spatial localization. *Journal of the Optical Society of America A* 20 (2003), 1391–1397.

- Wolfgang Becker and Albert F. Fuchs. 1969. Further properties of the human saccadic system: eye movements and correction saccades with and without visual fixation points. *Vision Research* 9, 10 (1969), 1247–1258.
- AH. Bell, MA. Meredith, AJ. Van Opstal, and DP. Munoz. 2006. Stimulus intensity modifies saccadic reaction time and visual response latency in the superior colliculus. *Experimental Brain Research* 174 (2006), 53–59.
- Christopher C. Berger, Mar Gonzalez-Franco, Ana Tajadura-Jiménez, Dinei Florencio, and Zhengyou Zhang. 2018. Generic HRTFs may be good enough in virtual reality. Improving source localization through cross-modal plasticity. *Frontiers in Neuroscience* 12 (2018), 21.
- Edurne Bernal-Berdun, Mateo Vallejo, Qi Sun, Ana Serrano, and Diego Gutierrez. 2025. Audiovisual Disparities in VR: Impact on Spatial Perception. In *2025 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 1–8.
- Matthias Bernhard, Camillo Dell’mour, Michael Hecher, Efstathios Stavrakakis, and Michael Wimmer. 2014. The effects of fast disparity adjustment in gaze-controlled stereoscopic applications. In *Proceedings of the Symposium on Eye Tracking Research and Applications*. 111–118.
- Jonathan S. Berry, David AT. Roberts, and Nicolas S. Holliman. 2014. 3D sound and 3D image interactions: a review of audio-visual depth perception. *Human Vision and Electronic Imaging XIX* 9014 (2014), 33–48.
- Stefan Berti and Erich Schröger. 2001. A comparison of auditory and visual distraction effects: behavioral and event-related indices. *Cognitive Brain Research* 10, 3 (2001), 265–273.
- Piotr Bilinski, Jens Ahrens, Mark RP. Thomas, Ivan J. Tashev, and John C. Platt. 2014. HRTF magnitude synthesis via sparse representation of anthropometric features. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4468–4472.
- M. Bischoff, B. Walter, CR. Blecker, K. Morgen, D. Vaitl, and G. Sammer. 2007. Utilizing the ventriloquism-effect to investigate audio-visual binding. *Neuropsychologia* 45, 3 (2007), 578–586.
- D. Boghen, BT. Troost, RB. Daroff, LF. Dell’Osso, and JE. Birkett. 1974. Velocity characteristics of normal human saccades. *Investigative Ophthalmology & Visual Science* 13, 8 (1974), 619–623.
- R. Bolia, W. D’Angelo, and R. McKinley. 1999. Aurally aided visual search in three-dimensional space. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 41, 4 (1999), 664–669.
- Benjamin M. Bolker, Mollie E. Brooks, Connie J. Clark, Shane W. Geange, John R. Poulsen, M. Henry H. Stevens, and Jada-Simone S. White. 2009. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution* 24, 3 (2009), 127–135.
- Kris Boyen, Deniz Baskent, and Pim van Dijk. 2015. The gap detection test: can it be used to diagnose tinnitus? *Ear and Hearing* 36, 4 (2015), e138–e145.

- AW. Bronkhorst and T. Houtgast. 1999. Auditory distance perception in rooms. *Nature* 397, 6719 (1999), 517–520.
- Douglas S. Brungart and William M. Rabinowitz. 1999. Auditory localization of nearby sources. Head-related transfer functions. *The Journal of the Acoustical Society of America* 106, 3 (1999), 1465–1479.
- Patrick Bruns. 2019. The ventriloquist illusion as a tool to study multisensory processing: An update. *Frontiers in Integrative Neuroscience* 13 (2019), 51.
- T. Rowan Candy and Lawrence K. Cormack. 2022. Recent understanding of binocular vision in the natural environment with clinical implications. *Progress in Retinal and Eye Research* 88 (2022), 101014.
- Jianhua Cang, Jieming Fu, and Seiji Tanabe. 2023. Neural circuits for binocular vision: Ocular dominance, interocular matching, and disparity selectivity. *Frontiers in Neural Circuits* 17 (2023), 1084027.
- David P. Carey and Claire V. Hutchinson. 2013. Looking at eye dominance from a different angle: is sighting strength related to hand preference? *Cortex* 49, 9 (2013), 2542–2552.
- R.H.S. Carpenter. 2004. Contrast, Probability, and Saccadic Latency: Evidence for Independence of Detection and Decision. *Current Biology* 14, 17 (2004), 1576–1580.
- Roberto Cecere, Joachim Gross, and Gregor Thut. 2016. Behavioural evidence for separate mechanisms of audiovisual temporal binding as a function of leading sensory modality. *European Journal of Neuroscience* 43, 12 (2016), 1561–1568.
- Chenliang Chang, Kiseung Bang, Gordon Wetzstein, Byoungcho Lee, and Liang Gao. 2020. Toward the next-generation VR/AR optics: a review of holographic near-eye displays from a human-centric perspective. *Optica* 7, 11 (2020), 1563–1578.
- Geneviève Charbonneau, Marie Véronneau, Colin Boudrias-Fournier, Franco Lepore, and Olivier Collignon. 2013. The ventriloquist in periphery: impact of eccentricity-related reliability on audio-visual localization. *Journal of Vision* 13, 12 (2013), 20–20.
- Romain Chaumillon, Nadia Alahyane, Patrice Senot, Judith Vergne, Christelle Lemoine-Lardennois, Jean Blouin, Karine Doré-Mazars, Alain Guillaume, and Dorine Vergilino-Perez. 2017. Asymmetry in visual information processing depends on the strength of eye dominance. *Neuropsychologia* 96 (2017), 129–136.
- BD. Chaurasia and BBL. Mathur. 1976. Eyedness. *Cells Tissues Organs* 96, 2 (1976), 301–305.
- Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. 2020. Soundspaces: Audio-visual navigation in 3d environments. In *European Conference on Computer Vision*. Springer, 17–36.
- Kenneth Chen, Thomas Wan, Nathan Matsuda, Ajit Ninan, Alexandre Chapiro, and Qi Sun. 2024. Pea-pods: Perceptual evaluation of algorithms for power optimization in XR displays. *ACM Transactions on Graphics (TOG)* 43, 4 (2024), 1–17.

- Shaoyu Chen, Budmonde Duinkharjav, Xin Sun, Li-Yi Wei, Stefano Petrangeli, Jose Echevarria, Claudio Silva, and Qi Sun. 2022. Instant reality: Gaze-contingent perceptual optimization for 3d virtual reality streaming. *IEEE Transactions on Visualization and Computer Graphics* 28, 5 (2022), 2157–2167.
- Yi-Chuan Chen and Charles Spence. 2017. Assessing the role of the ‘unity assumption’ on multisensory integration: A review. *Frontiers in Psychology* 8 (2017), 445.
- James J. Clark and Alan L. Yuille. 1990. *Data Fusion for Sensory Information Processing Systems*. Kluwer Academic, Norwell, Mass.
- Han Collewyn, Casper J. Erkelens, and Robert M. Steinman. 1995. Voluntary Binocular Gaze-Shifts in the Plane of Regard. *Vision Research* 35, 23 (1995), 3335–3358.
- Hans Colonius and Petra Arndt. 2001. A two-stage model for visual-auditory interaction in saccadic latencies. *Perception & Psychophysics* 63, 1 (2001), 126–147.
- BD. Corneil, M. Van Wanrooij, DP. Munoz, and AJ. Van Opstal. 2002. Auditory-visual interactions subserving goal-directed saccades in a complex scene. *Journal of Neurophysiology* 88, 1 (2002), 438–454.
- RA. Crone and OMA. Leuridan. 1973. Tolerance for aniseikonia: I. Diplopia thresholds in the vertical and horizontal meridians of the visual field. *Albrecht von Graefes Archiv für klinische und experimentelle Ophthalmologie* 188, 1 (1973), 1–16.
- Richard S. Crutchfield, Donald G. Woodworth, and Ruth E. Albrecht. 1958. *Perceptual performance and the effective person*. Technical Report.
- Bruce G. Cumming and Gregory C. DeAngelis. 2001. The physiology of stereopsis. *Annual Review of Neuroscience* 24, 1 (2001), 203–238.
- Christine A. Curcio and Kimberly A. Allen. 1990. Topography of ganglion cells in human retina. *Journal of Comparative Neurology* 300, 1 (1990), 5–25.
- J. Cutting and P. Vishton. 1995. Perceiving layout and knowing distances: The integration, relative potency, and contextual use of different information about depth. In *Perception of Space and Motion (Handbook of Perception and Cognition)*, W. Epstein and S. Rogers (Eds.). Academic Press, 69–117.
- Joan M. Dafoe, Irene T. Armstrong, and Doug P. Munoz. 2007. The influence of stimulus direction and eccentricity on pro-and anti-saccades in humans. *Experimental Brain Research* 179 (2007), 563–570.
- Nianchen Deng, Zhenyi He, Jiannan Ye, Budmonde Duinkharjav, Praneeth Chakravarthula, Xubo Yang, and Qi Sun. 2022. Fov-nerf: Foveated neural radiance fields for virtual reality. *IEEE Transactions on Visualization and Computer Graphics* 28, 11 (2022), 3854–3864.
- Piotr Didyk, Tobias Ritschel, Elmar Eisemann, Karol Myszkowski, and Hans-Peter Seidel. 2011. A perceptual model for disparity. *ACM Transactions on Graphics (TOG)* 30, 4 (2011), 1–10.
- Adele Diederich, Annette Schomburg, and Hans Colonius. 2012. Saccadic reaction times to audiovisual stimuli show effects of oscillatory phase reset. *Plos One* (2012).

- Neil A. Dodgson. 2004. Variation and Extrema of Human Interpupillary Distance. In *Electronic Imaging 2004*. International Society for Optics and Photonics, 36–46.
- Oliver Doehrmann and Marcus J. Naumer. 2008. Semantics and the multisensory brain: how meaning modulates processes of audio-visual integration. *Brain Research* 1242 (2008), 136–150.
- Budmonde Duinkharjav, Praneeth Chakravarthula, Rachel Brown, Anjul Patney, and Qi Sun. 2022a. Image features influence reaction time: A learned probabilistic perceptual model for saccade latency. *ACM Transactions on Graphics (TOG)* 41, 4 (2022), 1–15.
- Budmonde Duinkharjav, Kenneth Chen, Abhishek Tyagi, Jiayi He, Yuhao Zhu, and Qi Sun. 2022b. Color-perception-guided display power reduction for virtual reality. *ACM Transactions on Graphics (TOG)* 41, 6 (2022), 1–16.
- Budmonde Duinkharjav, Benjamin Liang, Anjul Patney, Rachel Brown, and Qi Sun. 2023. The shortest route is not always the fastest: Probability-modeled stereoscopic eye movement completion time in vr. *ACM Transactions on Graphics (TOG)* 42, 6 (2023), 1–14.
- David Dunn, Okan Tursun, Hyeonseung Yu, Piotr Didyk, Karol Myszkowski, and Henry Fuchs. 2020. Stimulating the human visual system beyond real world performance in future augmented reality displays. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 90–100.
- Inc. Encyclopædia Britannica. [n. d.]. *Edgar Bergen*. <https://www.britannica.com/biography/Edgar-Bergen> Article.
- Isaac Engel, Rapolas Daugintis, Thibault Vicente, Aidan OT. Hogg, Johan Pauwels, Arnaud J. Tournier, and Lorenzo Picinali. 2023. The sonicom HRTF dataset. *Journal of the Audio Engineering Society* 71, 5 (2023), 241–253.
- JT. Enright. 1986. Facilitation of Vergence Changes by Saccades: Influences of Mis-focused Images and of Disparity Stimuli in Man. *The Journal of Physiology* 371, 1 (1986), 69–87.
- James T. Enright. 1984. Changes in vergence mediated by saccades. *The Journal of Physiology* 350, 1 (1984), 9–31.
- Charles W. Eriksen. 1963. *Perception And Personality*. (1963).
- Marc Ernst and Heinrich Bülthoff. 2004. Merging the Senses into a Robust Percept. *Trends in Cognitive Sciences* 8 (2004), 162–9.
- M. Fahle. 1982. Binocular rivalry: Suppression depends on orientation and spatial frequency. *Vision research* 22, 7 (1982), 787–800.
- Manfred Fahle. 1987. Naso-temporal asymmetry of binocular inhibition. *Investigative Ophthalmology & Visual Science* 28, 6 (1987), 1016–1017.
- M. Fahle and M. Schmid. 1988. Naso-temporal asymmetry of visual perception and of the visual cortex. *Vision Research* 28, 2 (1988), 293–300.

- Runze Fan, Xuehuai Shi, Kangyu Wang, Qixiang Ma, and Lili Wang. 2024. Scene-aware foveated rendering. *IEEE Transactions on Visualization and Computer Graphics* (2024).
- Gerald Farin. 1986. Triangular bernstein-bézier patches. *Computer Aided Geometric Design* 3, 2 (1986), 83–127.
- Derek Fender and Bela Julesz. 1967. Extension of Panum’s fusional area in binocularly stabilized vision. *Journal of the Optical Society of America* 57, 6 (1967), 819–830.
- Daniel J. Finnegan. 2017. *Compensating for distance compression in virtual audiovisual environments*. Ph. D. Dissertation. University of Bath.
- B. Fischer, S. Gezeck, and K. Hartnegg. 1997. The analysis of saccadic eye movements from gap and overlap paradigms. *Brain Research Protocols* 2, 1 (1997), 47–52.
- Burkhart Fischer and Heike Weber. 1993. Express saccades and visual attention. *Behavioral and Brain Sciences* 16, 3 (1993), 553–567.
- Philippe Fournier and Sylvie Hébert. 2013. Gap detection deficits in humans with tinnitus as assessed with the acoustic startle paradigm: does tinnitus fill in the gap? *Hearing Research* 295 (2013), 16–23.
- Linus Franke, Laura Fink, Jana Martschinke, Kai Selgrad, and Marc Stamminger. 2021. Time-Warped Foveated Rendering for Virtual Reality Headsets. In *Computer Graphics Forum*, Vol. 40. Wiley Online Library, 110–123.
- Maarten A. Frens, A. John Van Opstal, and Robert F. Van der Willigen. 1995. Spatial and temporal factors determine auditory-visual interactions in human saccadic eye movements. *Perception & Psychophysics* 57 (1995), 802–816.
- Werner Gaik. 1993. Combined evaluation of interaural time and intensity differences: Psychoacoustic results and computer modeling. *The Journal of the Acoustical Society of America* 94, 1 (1993), 98–110.
- Miguel A. García-Pérez and Rocío Alcalá-Quintana. 2015. Converging evidence that common timing processes underlie temporal-order and simultaneity judgments: A model-based analysis. *Attention, Perception, & Psychophysics* 77 (2015), 1750–1766.
- Mark A. Georgeson and Stuart A. Wallis. 2014. Binocular fusion, suppression and diplopia for blurred edges. *Ophthalmic and Physiological Optics* 34, 2 (2014), 163–185.
- James J. Gibson. 1950. The perception of visual surfaces. *The American Journal of Psychology* 63, 3 (1950), 367–384.
- Martine Godfroy, Corinne Roumes, and Pierre Dauchy. 2003. Spatial variations of visual—auditory fusion areas. *Perception* 32, 10 (2003), 1233–1245.
- Guilherme Gonçalves, Miguel Melo, José Vasconcelos-Raposo, and Maximino Bessa. 2019. Impact of different sensory stimuli on presence in credible virtual environments. *IEEE Transactions on Visualization and Computer Graphics* 26, 11 (2019), 3231–3240.
- Google DeepMind. 2025. Introducing Nano Banana Pro. <https://blog.google/technology/ai/nano-banana-pro/>. Accessed: 2026-05-29.

- Marcin Gorzel, David Corrigan, Gavin Kearney, John Squires, and Frank Boland. 2012. Distance perception in virtual audio-visual environments. In *25th UK Conference of the Audio Engineering Society: Spatial Audio In Today's 3D World (2012)*. 1–8.
- D. Wesley Grantham. 1984. Discrimination of dynamic interaural intensity differences. *The Journal of the Acoustical Society of America* 76, 1 (1984), 71–76.
- M. Grohn, T. Lokki, and T. Takala. 2005. Comparison of auditory, visual and audio-visual navigation in a 3D space. *ACM Transactions on Applied Perception* 2, 4 (2005), 564–570.
- Colin Groth, Sascha Fricke, Susana Castillo, and Marcus Magnor. 2023a. Wavelet-based fast decoding of 360 videos. *IEEE Transactions on Visualization and Computer Graphics* 29, 5 (2023), 2508–2516.
- Colin Groth, Marcus Magnor, Steve Grogorkick, Martin Eisemann, and Piotr Didyk. 2024. Cybersickness reduction via gaze-contingent image deformation. *ACM Transactions on Graphics (TOG)* 43, 4 (2024), 1–14.
- Colin Groth, Timon Scholz, Susana Castillo, Jan-Philipp Tauscher, and Marcus Magnor. 2023b. Instant Hand Redirection in Virtual Reality Through Electrical Muscle Stimulation-Triggered Eye Blinks. In *Proceedings of the 29th ACM Symposium on Virtual Reality Software and Technology*. 1–11.
- Quan Gu, Qikai Zhang, Yueming Han, Pengfei Li, Zaifeng Gao, and Mowei Shen. 2024. Microsaccades reflect attention shifts: a mini review of 20 years of microsaccade research. *Frontiers in Psychology* 15 (2024), 1364939.
- Brian Guenter, Mark Finch, Steven Drucker, Desney Tan, and John Snyder. 2012. Foveated 3D graphics. *ACM Transactions on Graphics (TOG)* 31, 6 (2012), 1–10.
- William M. Hartmann. 2021. Localization and lateralization of sound. *Binaural Hearing: With 93 Illustrations* (2021), 9–45.
- David Hecht, Miriam Reiner, and Gad Halevy. 2006. Multimodal Virtual Environments: Response Times, Attention, and Presence. *Presence* 15 (10 2006), 515–523.
- Selig Hecht. 1928. On the binocular fusion of colors and its relation to theories of color vision. *Proceedings of the National Academy of Sciences* 14, 3 (1928), 237–241.
- Selig Hecht. 1937. Rods, cones, and the chemical basis of vision. *Physiological Reviews* 17, 2 (1937), 239–290.
- Anita E. Hendrickson and Cristine Yuodelis. 1984. The morphological development of the human fovea. *Ophthalmology* 91, 6 (1984), 603–612.
- Souta Hidaka and Masakazu Ide. 2015. Sound can suppress visual perception. *Scientific reports* 5, 1 (2015), 10483.
- Rebecca J. Hirst, David P. McGovern, Annalisa Setti, Ladan Shams, and Fiona N. Newell. 2020. What you see is what you hear: Twenty years of research using the Sound-Induced Flash Illusion. *Neuroscience & Biobehavioral Reviews* 118 (2020), 759–774.
- David M. Hoffman and Martin S. Banks. 2010. Focus information is used to interpret binocular images. *Journal of Vision* 10, 5 (2010), 13–13.

- David M. Hoffman, Ahna R. Girshick, Kurt Akeley, and Martin S. Banks. 2008. Vergence–accommodation conflicts hinder visual performance and cause visual fatigue. *Journal of Vision* 8, 3 (2008), 33–33.
- Torin Hopkins, Suibi Che-Chuan Weng, Rishi Vanukuru, Emma Wenzel, Amy Banic, Mark D. Gross, and Ellen Yi-Luen Do. 2022. Studying the Effects of Network Latency on Audio-Visual Perception During an AR Musical Task. In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 26–34.
- Ian P. Howard and Brian J. Rogers. 2012. *Perceiving in Depth, Volume 2: Stereoscopic Vision*. Oxford University Press.
- Shinobu Ishihara. 1951. Tests for colour-blindness.
- William H. Ittelson. 1951. Size as a cue to distance: Static localization. *The American Journal of Psychology* 64, 1 (1951), 54–67.
- Charles E. Jack and Willard R. Thurlow. 1973. Effects of Degree of Visual Association and Angle of Displacement on the “Ventriloquism” Effect. *Perceptual and Motor Skills* 37, 3 (1973), 967–979.
- Akshay Jindal, Krzysztof Wolski, Karol Myszkowski, and Rafał K. Mantiuk. 2021. Perceptual model for adaptive local shading and refresh rate. *ACM Transactions on Graphics (TOG)* 40, 6 (2021), 1–18.
- Graham R. Jones, Delman Lee, Nicolas S. Holliman, and David Ezra. 2001. Controlling perceived depth in stereoscopic images. In *Stereoscopic Displays and Virtual Reality Systems VIII*, Vol. 4297. SPIE, 42–53.
- RP. Kalesnykas and PE. Hallett. 1987. The differentiation of visually guided and anticipatory saccades in gap and overlap paradigms. *Experimental Brain Research* 68 (1987), 115–121.
- RP. Kalesnykas and PE. Hallett. 1994. Retinal eccentricity and the latency of eye saccades. *Vision Research* 34, 4 (1994), 517–531.
- Anton S. Kaplanyan, Anton Sochenov, Thomas Leimkühler, Mikhail Okunev, Todd Goodall, and Gizem Rufo. 2019. DeepFovea: Neural reconstruction for foveated rendering and video compression using learned statistics of natural videos. *ACM Transactions on Graphics (TOG)* 38, 6 (2019), 1–13.
- Z. Kapoula and DA. Robinson. 1986. Saccadic undershoot is not inevitable: Saccades can be accurate. *Vision Research* 26, 5 (1986), 735–743.
- Emine Merve Kaya and Mounya Elhilali. 2017. Modelling auditory attention. *Philosophical Transactions of the Royal Society B: Biological Sciences* 372, 1714 (2017), 20160101.
- Petr Kellnhofer, Piotr Didyk, Karol Myszkowski, Mohamed M. Hefeeda, Hans-Peter Seidel, and Wojciech Matusik. 2016a. GazeStereo3D: Seamless disparity manipulations. *ACM Transactions on Graphics (TOG)* 35, 4 (2016), 1–13.
- Petr Kellnhofer, Piotr Didyk, Tobias Ritschel, Belen Masia, Karol Myszkowski, and Hans-Peter Seidel. 2016b. Motion parallax in stereo 3D: Model and applications. *ACM Transactions on Graphics (TOG)* 35, 6 (2016), 1–12.

- Sophie Kergaßner, Taimoor Tariq, and Piotr Didyk. 2025. Towards Understanding Depth Perception in Foveated Rendering. In *ACM SIGGRAPH 2025 Conference Papers*. 1–9.
- Daniel Kersten, David C. Knill, Pascal Mamassian, and Isabella Bühlhoff. 1996. Illusory motion from shadows. *Nature* 379, 6560 (1996), 31–31.
- Aarlenne Z. Khan and J. Douglas Crawford. 2001. Ocular dominance reverses as a function of horizontal gaze angle. *Vision Research* 41, 14 (2001), 1743–1748.
- Hayeon Kim and In-Kwon Lee. 2022. Studying the effects of congruence of auditory and visual stimuli on virtual reality experiences. *IEEE Transactions on Visualization and Computer Graphics* 28, 5 (2022), 2080–2090.
- Joohwan Kim, Josef Spjut, Morgan McGuire, Alexander Majercik, Ben Boudaoud, Rachel Albert, and David Luebke. 2019. Esports Arms Race: Latency and Refresh Rate for Competitive Gaming Tasks. *Journal of Vision* 19, 10 (2019), 2–2.
- Alan Kingstone and Raymond M. Klein. 1993. Visual offsets facilitate saccadic latency: does predisengagement of visuospatial attention mediate this gap effect? *Journal of Experimental Psychology: Human Perception and Performance* 19, 6 (1993), 1251.
- Akiyoshi Kitaoka and Hiroshi Ashida. 2003. Phenomenal characteristics of the peripheral drift illusion. *Vision* 15, 4 (2003), 261–262.
- D. C. Knill and W. Richards. 1996. *Perception as Bayesian Inference*. Cambridge University Press.
- Andrew J. Kolarik, Brian CJ. Moore, Pavel Zahorik, Silvia Cirstea, and Shahina Pardhan. 2016. Auditory distance perception in humans: a review of cues, development, neuronal bases, and effects of sensory loss. *Attention, Perception, & Psychophysics* 78 (2016), 373–395.
- Helga Kolb. 2012. Simple anatomy of the retina. (2012).
- Robert Konrad, Anastasios Angelopoulos, and Gordon Wetzstein. 2020. Gaze-contingent ocular parallax rendering for virtual reality. *ACM Transactions on Graphics (TOG)* 39, 2 (2020), 1–12.
- Denis Koposov, Maria Semenova, Andrey Somov, Andrey Lange, Anton Stepanov, and Evgeny Burnaev. 2020. Analysis of the reaction time of esports players through the gaze tracking and personality trait. In *2020 IEEE 29th International Symposium on Industrial Electronics (ISIE)*. IEEE, 1560–1565.
- Eileen Kowler. 2011. Eye Movements: The Past 25 Years. *Vision Research* 51, 13 (2011), 1457–1483.
- Brooke Krajancich, Petr Kellnhofer, and Gordon Wetzstein. 2023. Towards attention-aware foveated rendering. *ACM Transactions on Graphics (TOG)* 42, 4 (2023), 1–10.
- Alexander Kramer, Brigitte Röder, and Patrick Bruns. 2020. Feedback modulates audio-visual spatial recalibration. *Frontiers in Integrative Neuroscience* 13 (2020), 74.
- Gregory Kramida. 2015. Resolving the vergence-accommodation conflict in head-mounted displays. *IEEE Transactions on Visualization and Computer Graphics* 22, 7 (2015), 1912–1931.

- Jérôme Kudnick, Martin Weier, Colin Groth, Biying Fu, and Robin Horst. 2025. WarpVision: Using Spatial Curvature to Guide Attention in Virtual Reality. *IEEE Transactions on Visualization and Computer Graphics* (2025).
- Daniel P. Kumpik, Connor Campbell, Jan WH. Schnupp, and Andrew J. King. 2019. Re-weighting of sound localization cues by audiovisual training. *Frontiers in Neuroscience* 13 (2019), 1164.
- Yuna Kwak, Eric Penner, Xuan Wang, Mohammad R. Saeedpour-Parizi, Olivier Mercier, Xiuyun Wu, Scott Murdison, and Phillip Guan. 2024. Saccade-contingent rendering. In *ACM SIGGRAPH 2024 Conference Papers*. 1–9.
- M. S. Landy, L. T. Maloney, E. B. Johnston, and M. Young. 1995. Measurement and modeling of depth cue combination: In defense of weak fusion. *Vision Research* 35, 3 (1995), 389–412.
- Manuel Lang, Alexander Hornung, Oliver Wang, Steven Poulakos, Aljoscha Smolic, and Markus Gross. 2010. Nonlinear disparity mapping for stereoscopic 3D. *ACM Transactions on Graphics (TOG)* 29, 4 (2010), 1–10.
- Shahid Latif, Hagen Tarner, and Fabian Beck. 2021. Talking realities: audio guides in virtual reality visualizations. *IEEE Computer Graphics and Applications* 42, 1 (2021), 73–83.
- R. John Leigh and David S. Zee. 2015. *The neurology of eye movements*. Oxford University Press.
- Willem JM Levelt. 1965a. Binocular brightness averaging and contour information. *British Journal of Psychology* 56, 1 (1965), 1–13.
- Willem JM. Levelt. 1965b. *On binocular rivalry*. Ph. D. Dissertation. Van Gorcum Assen.
- David Li, Ruofei Du, Adharsh Babu, Camelia D. Brumar, and Amitabh Varshney. 2021. A log-rectilinear transformation for foveated 360-degree video streaming. *IEEE Transactions on Visualization and Computer Graphics* 27, 5 (2021), 2638–2647.
- Wan-Yi Lin, Rohith Venkatakrishnan, Roshan Venkatakrishnan, Sabarish V. Babu, Christopher Pagano, and Wen-Chieh Lin. 2024. An empirical evaluation of the calibration of auditory distance perception under different levels of virtual environment visibilities. In *2024 IEEE conference virtual reality and 3D user interfaces (VR)*. IEEE, 690–700.
- Matteo Lisi, Joshua A. Solomon, and Michael J. Morgan. 2019. Gain control of saccadic eye movements is probabilistic. *Proceedings of the National Academy of Sciences* 116, 32 (2019), 16137–16142.
- Alex D. Little, Donald H. Mershon, and Patrick H. Cox. 1992. Spectral Content as a Cue to Perceived Auditory Distance. *Perception* 21, 3 (1992), 405–416.
- Wenxuan Liu, Budmonde Duinkharjav, Qi Sun, and Sai Qian Zhang. 2025. Fovealnet: Advancing ai-driven gaze tracking solutions for efficient foveated rendering in virtual reality. *IEEE Transactions on Visualization and Computer Graphics* (2025).
- Margaret S. Livingstone. 2014. *Vision and Art: The Biology of Seeing* (updated and expanded edition ed.). Abrams, New York.

- Andrew Maimone, Gordon Wetzstein, Matthew Hirsch, Douglas Lanman, Ramesh Raskar, and Henry Fuchs. 2013. Focus 3D: Compressive accommodation display. *ACM Transactions on Graphics (TOG)* 32, 5 (2013), 153–1.
- Sandra Malpica, Ana Serrano, Julia Guerrero-Viu, Daniel Martin, Edurne Bernal, Diego Gutierrez, and Belen Masia. 2022. Auditory stimuli degrade visual performance in virtual reality. In *ACM SIGGRAPH 2022 Posters*. 1–2.
- George R. Mangun. 1995. Neural mechanisms of visual selective attention. *Psychophysiology* 32, 1 (1995), 4–18.
- Rafał K. Mantiuk, Maliha Ashraf, and Alexandre Chapiro. 2022. stelaCSF: a unified model of contrast sensitivity as the function of spatio-temporal frequency, eccentricity, luminance and area. *ACM Transactions on Graphics (TOG)* 41, 4 (2022), 1–16.
- Lawrence E. Marks. 1979. A theory of loudness and loudness judgments. *Psychological Review* 86, 3 (1979), 256.
- David Marr. 1980. Visual information processing: The structure and creation of visual representations. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences* 290, 1038 (1980), 199–218.
- Daniel Martin, Sandra Malpica, Diego Gutierrez, Belen Masia, and Ana Serrano. 2022. Multimodality in VR: A Survey. *ACM Comput. Surv.* 54, 10s, Article 216 (2022).
- Susana Martinez-Conde, Stephen L. Macknik, Xoana G. Troncoso, and Thomas A. Dyar. 2006. Microsaccades counteract visual fading during fixation. *Neuron* 49, 2 (2006), 297–305.
- George Mather and David RR. Smith. 2002. Blur discrimination and its relation to blur-mediated depth perception. *Perception* 31, 10 (2002), 1211–1219.
- Ethel Martin. 1974. Saccadic suppression: a review and an analysis. *Psychological Bulletin* 81, 12 (1974), 899.
- Dennis McFadden and Edward G. Pasanen. 1976. Lateralization at high frequencies based on interaural time differences. *The Journal of the Acoustical Society of America* 59, 3 (1976), 634–639.
- Daniel V. Meegan, Lew B. Stelmach, and W. James Tam. 2001. Unequal weighting of monocular inputs in binocular combination: Implications for the compression of stereoscopic imagery. *Journal of Experimental Psychology: Applied* 7, 2 (2001), 143.
- Xiaoxu Meng, Ruofei Du, and Amitabh Varshney. 2020. Eye-dominance-guided foveated rendering. *IEEE Transactions on Visualization and Computer Graphics* 26, 5 (2020), 1972–1980.
- Xiaoxu Meng, Ruofei Du, Matthias Zwicker, and Amitabh Varshney. 2018. Kernel foveated rendering. *Proceedings of the ACM on Computer Graphics and Interactive Techniques* 1, 1 (2018), 1–20.
- Yuri Mikawa and Taiki Fukiage. 2024. Low-latency ocular parallax rendering and investigation of its effect on depth perception in virtual reality. *IEEE Transactions on Visualization and Computer Graphics* (2024).

- Walter R Miles. 1930. Ocular dominance in human adults. *The Journal of General Psychology* 3, 3 (1930), 412–430.
- Nancy J. Minshew and Jessica A. Hobson. 2008. Sensory sensitivities and performance on sensory perceptual tasks in high-functioning individuals with autism. *Journal of Autism and Developmental Disorders* 38, 8 (2008), 1485–1498.
- Donald E. Mitchell. 1966. Retinal disparity and diplopia. *Vision Research* 6, 7-8 (1966), 441–451.
- Mark Mon-Williams, James R. Tresilian, and Andrew Roberts. 2000. Vergence provides veridical depth perception from horizontal retinal image disparities. *Experimental Brain Research* 133 (2000), 407–413.
- Theo Moons, Luc Van Gool, and Maarten Vergauwen. 2010. 3D reconstruction from multiple images part 1: Principles. *Foundations and Trends in Computer Graphics and Vision* 4, 4 (2010), 287–404.
- Aileen C. Naef, Marie-Madlen Jeitziner, Samuel EJ. Knobel, Matthias Thomas Exl, René M. Müri, Stephan M. Jakob, Tobias Nef, and Stephan M. Gerber. 2022. Investigating the role of auditory and visual sensory inputs for inducing relaxation during virtual reality stimulation. *Scientific Reports* 12, 1 (2022), 17073.
- Marc Naguib and R. Haven Wiley. 2001. Estimating the distance to a source of sound: mechanisms and adaptations for long-range communication. *Animal Behaviour* 62, 5 (2001), 825–837.
- Kumpei Ogawa, Kazuyuki Fujita, Shuichi Sakamoto, Kazuki Takashima, and Yoshifumi Kitamura. 2023. Exploring Visual-Auditory Redirected Walking using Auditory Cues in Reality. *IEEE Transactions on Visualization and Computer Graphics* (2023).
- Robert P. O’Shea, Shane G. Blackburn, and Hiroshi Ono. 1994. Contrast as a depth cue. *Vision Research* 34, 12 (1994), 1595–1604.
- Thomas Oskam, Alexander Hornung, Huw Bowles, Kenny Mitchell, and Markus H. Gross. 2011. OSCAM-optimized stereoscopic camera control for interactive 3D. *ACM Trans. Graph.* 30, 6 (2011), 189.
- Jorge Otero-Millan, Stephen L. Macknik, and Susana Martinez-Conde. 2012. Microsaccades and blinks trigger illusory rotation in the “rotating snakes” illusion. *Journal of Neuroscience* 32, 17 (2012), 6043–6051.
- Matthias Ott. 2006. Visual accommodation in vertebrates: mechanisms, physiological response and stimuli. *Journal of Comparative Physiology A* 192, 2 (2006), 97–111.
- Chris LE. Paffen. 2025. Variations in sensory eye dominance along the horizontal meridian. *Journal of Vision* 25, 8 (2025), 1–1.
- Dawid Pająk, Robert Herzog, Radosław Mantiuk, Piotr Didyk, Elmar Eisemann, Karol Myszkowski, and Kari Pulli. 2014. Perceptual depth compression for stereo applications. In *Computer Graphics Forum*, Vol. 33. Wiley Online Library, 195–204.
- Stephen Palmer. 1999. *Vision Science: Photons to Phenomenology*. MIT Press.

- Anjul Patney, Marco Salvi, Joohwan Kim, Anton Kaplanyan, Chris Wyman, Nir Benty, David Luebke, and Aaron Lefohn. 2016. Towards foveated rendering for gaze-tracked virtual reality. *ACM Transactions On Graphics (TOG)* 35, 6 (2016), 1–12.
- Eli Peli, Jian Yang, and Robert B. Goldstein. 1991. Image invariance with changes in size: The role of peripheral contrast thresholds. *Journal of the Optical Society of America A* 8, 11 (1991), 1762–1774.
- Xi Peng, Kenneth Chen, Iran Roman, Juan Pablo Bello, Qi Sun, and Praneeth Chakravarthula. 2025. Perceptually-Guided Acoustic" Foveation". In *2025 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*. IEEE, 450–460.
- Valeria I. Petkova and H. Henrik Ehrsson. 2008. If I were you: perceptual illusion of body swapping. *Plos One* 3, 12 (2008), e3832.
- Charles Pierrot-Deseilligny, Sophie Rivaud, Bertrand Gaymard, René Müri, and Anne-Isabelle Vermersch. 1995. Cortical control of saccades. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society* 37, 5 (1995), 557–567.
- Gian F. Poggio and Tomaso Poggio. 1984. The analysis of stereopsis. *Annual Review of Neuroscience* 7, 1 (1984), 379–412.
- Clare Porac and Stanley Coren. 1976. The dominant eye. *Psychological Bulletin* 83, 5 (1976), 880.
- Ivan Poupyrev, Tadao Ichikawa, Suzanne Weghorst, and Mark Billinghurst. 1998. Egocentric object manipulation in virtual environments: empirical evaluation of interaction techniques. In *Computer Graphics Forum*, Vol. 17. Wiley Online Library, 41–52.
- Albert R. Powers, Andrea R. Hillock, and Mark T. Wallace. 2009. Perceptual training narrows the temporal window of multisensory binding. *Journal of Neuroscience* 29, 39 (2009), 12265–12274.
- Jan M. Provis, Adam M. Dubis, Ted Maddess, and Joseph Carroll. 2013. Adaptation of the central retina for high acuity vision: cones, the fovea and the avascular zone. *Progress in retinal and eye research* 35 (2013), 63–81.
- Franziska Prummer, Mohamed Shereef Abdelwahab, Florian Weidner, Yasmeen Abdrabou, and Hans Gellersen. 2025a. It's Not Always the Same Eye That Dominates: Effects of Viewing Angle, Handedness and Eye Movement in 3D. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–10.
- Franziska Prummer, Ludwig Sidenmark, and Hans Gellersen. 2024. Dynamics of eye dominance behavior in virtual reality. *Journal of Eye Movement Research* 17, 3 (2024), 14.
- Franziska Prummer, Florian Weidner, and Hans Gellersen. 2025b. All But Static: Exploring Dynamic Eye Dominance for Foveated Rendering. In *2025 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, 376–380.
- Ning Qian. 1997. Binocular disparity and the perception of depth. *Neuron* 18, 3 (1997), 359–368.

- Jo Quartley and Alison Y. Firth. 2004. Binocular sighting ocular dominance changes with different angles of horizontal gaze. *Binocular Vision & Strabismus Quarterly* 19, 1 (2004), 25–30.
- Vilayanur S. Ramachandran. 1992. Blind spots. *Scientific American* 266, 5 (1992), 86–91.
- Patricia A. Reuter-Lorenz, Howard C. Hughes, and Robert Fendrich. 1991. The reduction of saccadic latency by prior offset of the fixation point: an analysis of the gap effect. *Perception & Psychophysics* 49, 2 (1991), 167–175.
- Flavio Ribeiro, Cha Zhang, Dinei A. Florêncio, and Demba Elimane Ba. 2010. Using reverberation to improve range and elevation discrimination for small array sound source localization. *IEEE Transactions on Audio, Speech, and Language Processing* 18, 7 (2010), 1781–1792.
- DA. Robinson. 1965. The mechanics of human smooth pursuit eye movement. *The Journal of Physiology* 180, 3 (1965), 569.
- DA. Robinson. 1966. The mechanics of human vergence eye movement. 31–37 pages.
- Brian Rogers and Maureen Graham. 1979. Motion parallax as an independent cue for depth perception. *Perception* 8, 2 (1979), 125–134.
- Hermann Rorschach. 1921. *Psychodiagnostik: Methodik und ergebnisse eines wahrnehmungsdiagnostischen Experiments (deutenlassen von zufallsformen)*. Vol. 2. E. Bircher.
- John Ross, M. Concetta Morrone, and David C. Burr. 1997. Compression of visual space before saccades. *Nature* 386, 6625 (1997), 598–601.
- John Ross, M. Concetta Morrone, Michael E. Goldberg, and David C. Burr. 2001. Changes in visual perception at the time of saccades. *Trends in Neurosciences* 24, 2 (2001), 113–121.
- Susan M. Ross and Leonard E. Ross. 1981. Saccade latency and warning signals: effects of auditory and visual stimulus onset and offset. *Perception & Psychophysics* 29 (1981), 429–437.
- Olli Rummukainen, Thomas Robotham, Axel Plinge, Frank Wefers, Jürgen Herre, Emanuel Habets, et al. 2019. Listening tests with individual versus generic head-related transfer functions in six-degrees-of-freedom virtual reality. In *Audio for Virtual, Augmented and Mixed Realities: Proceedings of ICASA 2019; 5th International Conference on Spatial Audio; September 26th to 28th, 2019, Ilmenau, Germany*. 55–62.
- Carl Schissler, Aaron Nicholls, and Ravish Mehra. 2016. Efficient HRTF-based spatial audio for area and volumetric sources. *IEEE Transactions on Visualization and Computer Graphics* 22, 4 (2016), 1356–1366.
- Timon Scholz, Colin Groth, Susana Castillo, Martin Eisemann, and Marcus Magnor. 2024. Measuring Velocity Perception Regarding Stimulus Eccentricity. In *ACM Symposium on Applied Perception (SAP)*. ACM, 1–9.
- Clifton Schor, Ivan Wood, and Jane Ogawa. 1984. Binocular sensory fusion is limited by spatial resolution. *Vision Research* 24, 7 (1984), 661–665.

- Clifton M. Schor, Lori A. Lott, David Pope, and Andrew D. Graham. 1999. Saccades reduce latency and increase velocity of ocular accommodation. *Vision Research* 39, 22 (1999), 3769–3795.
- Clifton M. Schor and Christopher W. Tyler. 1981. Spatio-temporal properties of Panum’s fusional area. *Vision Research* 21, 5 (1981), 683–692.
- John Semmlow and Paul Wetzel. 1979. Dynamic contributions of the components of binocular vergence. *Journal of the Optical Society of America* 69, 5 (1979), 639–645.
- Stefania Serafin, Michele Geronazzo, Cumhur Erkut, Niels C. Nilsson, and Rolf Nordahl. 2018. Sonic interactions in virtual reality: State of the art, current challenges, and future directions. *IEEE Computer Graphics and Applications* 38, 2 (2018), 31–43.
- Ana Serrano, Incheol Kim, Zhili Chen, Stephen DiVerdi, Diego Gutierrez, Aaron Hertzmann, and Belen Masia. 2019. Motion parallax for 360 RGBD video. *IEEE Transactions on Visualization and Computer Graphics* 25, 5 (2019), 1817–1827.
- Pieter Seuntjens, Lydia Meesters, and Wijnand Ijsselstein. 2006. Perceived quality of compressed stereoscopic images: Effects of symmetric and asymmetric JPEG coding and camera separation. *ACM Transactions on Applied Perception (TAP)* 3, 2 (2006), 95–109.
- Ladan Shams, Yukiyasu Kamitani, and Shinsuke Shimojo. 2000. What you see is what you hear. *Nature* 408, 6814 (2000), 788–788.
- Ladan Shams, Wei Ji Ma, and Ulrik Beierholm. 2005. Sound-induced flash illusion as an optimal percept. *Neuroreport* 16, 17 (2005), 1923–1927.
- James E. Sheedy, Ian L. Bailey, Markus Buri, and Eric Bass. 1986. Binocular vs. monocular task performance. *American Journal of Optometry and Physiological Optics* 63, 10 (1986), 839–846.
- Xuehuai Shi, Yucheng Li, Jiaheng Li, Jian Wu, Jieming Yin, Xiaobai Chen, and Lili Wang. 2025. Audio-visual aware Foveated Rendering. *IEEE Transactions on Visualization and Computer Graphics* (2025).
- Takashi Shibata, Joohwan Kim, David M. Hoffman, and Martin S. Banks. 2011. The zone of comfort: Predicting visual discomfort with stereo displays. *Journal of Vision* 11, 8 (2011), 11–11.
- Einat Shneur and Shaul Hochstein. 2005. Effects of eye dominance in visual perception. In *International Congress Series*, Vol. 1282. Elsevier, 719–723.
- John Siderov and Ronald S. Harwerth. 1995. Stereopsis, spatial frequency and retinal eccentricity. *Vision Research* 35, 16 (1995), 2329–2337.
- Daniel A. Slutsky and Gregg H. Recanzone. 2001. Temporal and spatial dependency of the ventriloquism effect. *Neuroreport* 12, 1 (2001), 7–10.
- Charles Spence. 2011. Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics* 73 (2011), 971–995.
- Charles Spence and Jon Driver. 2004. *Crossmodal space and crossmodal attention*. Oxford University Press.

- George Sperling. 1970. Binocular vision: A physical and a neural theory. *The American Journal of Psychology* (1970), 461–534.
- Daniel P. Spiegel and Ian M. Erkelens. 2024. Vergence-accommodation conflict increases time to focus in augmented reality. *Journal of the Society for Information Display* 32, 5 (2024), 194–205.
- John S. Stahl. 2001. Eye-head coordination and the variation of eye-movement accuracy with orbital eccentricity. *Experimental Brain Research* 136 (2001), 200–210.
- Jody Stanley, Olivia Carter, and Jason Forte. 2011. Color and luminance influence, but can not explain, binocular rivalry onset bias. *Plos One* 6, 5 (2011), e18978.
- Jody Stanley, Jason D. Forte, and Olivia Carter. 2019. Rivalry onset in and around the fovea: The role of visual field location and eye dominance on perceptual dominance bias. *Vision* 3, 4 (2019), 51.
- Barry E. Stein and Terrence R. Stanford. 2008. Multisensory integration: current issues from the perspective of the single neuron. *Nature Reviews Neuroscience* 9, 4 (2008), 255–266.
- Michael Stengel, Steve Grogorick, Martin Eisemann, and Marcus Magnor. 2016. Adaptive image-space sampling for gaze-contingent real-time rendering. In *Computer Graphics Forum*, Vol. 35. Wiley Online Library, 129–139.
- Ryan A. Stevenson and Mark T. Wallace. 2013. Multisensory temporal integration: task and stimulus dependencies. *Experimental Brain Research* 227 (2013), 249–261.
- Ryan A. Stevenson, Raquel K. Zemtsov, and Mark T. Wallace. 2012. Individual differences in the multisensory temporal binding window predict susceptibility to audiovisual illusions. *Journal of Experimental Psychology: Human Perception and Performance* 38, 6 (2012), 1517.
- David Stidwill and Robert Fletcher. 2017. *Normal binocular vision: Theory, investigation and practical aspects*. John Wiley & Sons.
- Thomas Z. Strybel and David R. Perrott. 1984. Discrimination of relative distance in the auditory modality: the success and failure of the loudness discrimination hypothesis. *The Journal of the Acoustical Society of America* 76 1 (1984), 318–20.
- Taimoor Tariq and Piotr Didyk. 2024. Towards motion metamers for foveated rendering. *ACM Transactions on Graphics (TOG)* 43, 4 (2024), 1–10.
- Taimoor Tariq, Cara Tursun, and Piotr Didyk. 2022. Noise-based enhancement for foveated rendering. *ACM Transactions on Graphics (TOG)* 41, 4 (2022), 1–14.
- Wolfgang A. Teder-Sälejärvi, Francesco Di Russo, John J. McDonald, and Steven A. Hillyard. 2005. Effects of spatial congruity on audio-visual multimodal integration. *Journal of Cognitive Neuroscience* 17, 9 (2005), 1396–1409.
- Donald J. Tellinghuisen and Erin J. Nowak. 2003. The inability to ignore auditory distractors as a function of visual task perceptual load. *Perception & Psychophysics* 65 (2003), 817–828.

- Krzysztof Templin, Piotr Didyk, Karol Myszkowski, Mohamed M. Hefeeda, Hans-Peter Seidel, and Wojciech Matusik. 2014. Modeling and optimizing eye vergence response to stereoscopic cuts. *ACM Transactions on Graphics (TOG)* 33, 4 (2014), 1–8.
- Frank Tong, Ming Meng, and Randolph Blake. 2006. Neural bases of binocular rivalry. *Trends in Cognitive Sciences* 10, 11 (2006), 502–511.
- Amy Turner, Jonathan Berry, and Nick Holliman. 2011. Can the perception of depth in stereoscopic images be influenced by 3D sound?. In *Stereoscopic Displays and Applications XXII*, Vol. 7863. SPIE, 71–80.
- Okan Tarhan Tursun, Elena Arabadzhiyska-Koleva, Marek Wernikowski, Radosław Mantiuk, Hans-Peter Seidel, Karol Myszkowski, and Piotr Didyk. 2019. Luminance-contrast-aware foveated rendering. *ACM Transactions on Graphics (TOG)* 38, 4 (2019), 1–14.
- Robert J. van Beers. 2008. Saccadic eye movements minimize the consequences of motor noise. *Plos One* 3, 4 (2008), e2070.
- Erik Van der Burg, David Alais, and John Cass. 2013. Rapid recalibration to audiovisual asynchrony. *Journal of neuroscience* 33, 37 (2013), 14633–14637.
- N. Van der Stoep, Charles Spence, TCW. Nijboer, and S. Van der Stigchel. 2015. On the relative contributions of multisensory integration and crossmodal exogenous spatial attention to multisensory response enhancement. *Acta Psychologica* 162 (2015), 20–28.
- Marc M. Van Wanrooij, Andrew H. Bell, Douglas P. Munoz, and A. John Van Opstal. 2009. The effect of spatial–temporal audiovisual disparities on saccades in a complex scene. *Experimental Brain Research* 198 (09 2009), 425–437.
- Argiro Vatakis and Charles Spence. 2006. Audiovisual synchrony perception for music, speech, and object actions. *Brain Research* 1111, 1 (2006), 134–142.
- Argiro Vatakis and Charles Spence. 2007. Crossmodal binding: Evaluating the “unity assumption” using audiovisual speech stimuli. *Perception & Psychophysics* 69 (2007), 744–756.
- Manuel Vidal, Andrea Desantis, and Laurent Madelain. 2020. Irrelevant auditory and tactile signals, but not visual signals, interact with the target onset and modulate saccade latencies. *Plos One* 15, 2 (2020), e0221192.
- Manuel Vidal and Françoise Vitu. 2022. Multisensory temporal binding induces an illusory gap/overlap that reduces the expected audiovisual interactions on saccades but not manual responses. *Plos One* 17, 4 (2022), e0266468.
- Jean Vroomen and Mirjam Keetels. 2010. Perception of intersensory synchrony: a tutorial review. *Attention, Perception, & Psychophysics* 72, 4 (2010), 871–884.
- Mark T. Wallace and Ryan A. Stevenson. 2014. The construct of the multisensory temporal binding window and its dysregulation in developmental disabilities. *Neuropsychologia* 64 (2014), 105–123.
- Brian A. Wandell. 1995. *Foundations of vision*. Sinauer Associates.

- Chin-An Wang, Gunnar Blohm, Jeff Huang, Susan E. Boehnke, and Douglas P. Munoz. 2017. Multisensory integration in orienting behavior: Pupil size, microsaccades, and saccades. *Biological Psychology* 129 (2017), 36–44.
- Jialin Wang, Rongkai Shi, Wenxuan Zheng, Weijie Xie, Dominic Kao, and Hai-Ning Liang. 2023c. Effect of frame rate on user experience, performance, and simulator sickness in virtual reality. *IEEE Transactions on Visualization and Computer Graphics* 29, 5 (2023), 2478–2488.
- Lili Wang, Xuehuai Shi, and Yi Liu. 2023b. Foveated rendering: A state-of-the-art survey. *Computational Visual Media* 9, 2 (2023), 195–228.
- Qisong Wang and Per Ola Kristensson. 2026. Stiffness Discrimination Thresholds of Force-Feedback Gloves for Volumetric Data Exploration in Virtual Reality. *IEEE Transactions on Visualization and Computer Graphics* (2026).
- Zhimin Wang, Xiangyuan Gu, and Feng Lu. 2023a. Deamp: Dominant-eye-aware foveated rendering with multi-parameter optimization. In *2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 632–641.
- Zhimin Wang, Xiangyuan Gu, and Feng Lu. 2025. Dominant-Eye-Aware Asymmetric Foveated Rendering for Virtual Reality. *IEEE Transactions on Visualization and Computer Graphics* (2025).
- John P. Wann, Simon Rushton, and Mark Mon-Williams. 1995. Natural problems for stereoscopic depth perception in virtual environments. *Vision Research* 35, 19 (1995), 2731–2736.
- Richard M. Warren, Eugene A. Sersen, and Edwin B. Pores. 1958. A basis for loudness-judgments. *The American Journal of Psychology* 71, 4 (1958), 700–709.
- Andrew B. Watson and Albert J. Ahumada. 2011. Blur clarified: A review and synthesis of blur discrimination. *Journal of Vision* 11, 5 (2011), 10–10.
- Martin Weier, Thorsten Roth, Ernst Kruijff, André Hinkenjann, Arsène Pérard-Gayot, Philipp Slusallek, and Yongmin Li. 2016. Foveated real-time ray tracing for head-mounted displays. In *Computer Graphics Forum*, Vol. 35. Wiley Online Library, 289–298.
- Martin Weier, Michael Stengel, Thorsten Roth, Piotr Didyk, Elmar Eisemann, Martin Eisemann, Steve Grogork, André Hinkenjann, Ernst Kruijff, Marcus Magnor, et al. 2017. Perception-driven accelerated rendering. In *Computer Graphics Forum*, Vol. 36. Wiley Online Library, 611–643.
- Elizabeth M. Wenzel, Marianne Arruda, Doris J. Kistler, and Frederic L. Wightman. 1993. Localization using nonindividualized head-related transfer functions. *The Journal of the Acoustical Society of America* 94, 1 (1993), 111–123.
- Frederic L. Wightman and Doris J. Kistler. 1992. The dominant role of low-frequency interaural time differences in sound localization. *The Journal of the Acoustical Society of America* 91, 3 (1992), 1648–1661.
- Jeremy M. Wolfe. 1983. Influence of spatial frequency, luminance, and duration on binocular rivalry and abnormal fusion of briefly presented dichoptic stimuli. *Perception* 12, 4 (1983), 447–456.

- Krzysztof Wolski, Laura Trutoiu, Zhao Dong, Zhengyang Shen, Kevin Mackenzie, and Alexandre Chapiro. 2022. Geo-metric: A perceptual dataset of distortions on faces. *ACM Transactions on Graphics (TOG)* 41, 6 (2022), 1–13.
- Muzhe Wu, Yi Fei Cheng, and David Lindlbauer. 2024. New Ears: An Exploratory Study of Audio Interaction Techniques for Performing Search in a Virtual Reality Environment. In *2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 386–395.
- Shimpei Yamagishi and Shigeto Furukawa. 2020. Factors influencing saccadic reaction time: Effect of task modality, stimulus saliency, spatial congruency of stimuli, and pupil size. *Frontiers in Human Neuroscience* 14 (2020), 571893.
- Qing Yang, Maria Pia Bucci, and Zoï Kapoula. 2002. The latency of saccades, vergence, and combined eye movements in children and in adults. *Investigative Ophthalmology & Visual Science* 43, 9 (2002), 2939–2949.
- Qing Yang and Zoï Kapoula. 2004. Saccade–vergence dynamics and interaction in children and in adults. *Experimental Brain Research* 156, 2 (2004), 212–223.
- Jiannan Ye, Anqi Xie, Susmija Jabbireddy, Yunchuan Li, Xubo Yang, and Xiaoxu Meng. 2022. Rectangular mapping-based foveated rendering. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 756–764.
- Pavel Zahorik. 2003. Auditory and visual distance perception: The proximity-image effect revisited. *Journal of the Acoustical Society of America* 113 (2003), 2270–2270.
- Pavel Zahorik, Douglas Brungart, and Adelbert Bronkhorst. 2005. Auditory distance perception in humans: A summary of past and present research. *Acta Acustica united with Acustica* 91 (2005), 409–420.
- Daniela Zambarbieri. 2002. The latency of saccades toward auditory targets in humans. *Progress in Brain Research* 140 (2002), 51–59.
- Daniela Zambarbieri, Giorgio Beltrami, and Maurizio Versino. 1995. Saccade latency toward auditory targets depends on the relative position of the sound source with respect to the eyes. *Vision Research* 35, 23-24 (1995), 3305–3312.
- David S. Zee, Edmond J. Fitzgibbon, and Lance M. Optican. 1992. Saccade-vergence interactions in humans. *Journal of Neurophysiology* 68, 5 (1992), 1624–1641.
- ZhiYing Zhou, Adrian David Cheok, Yan Qiu, and Xubo Yang. 2007. The role of 3-D sound in human reaction and performance in augmented reality environments. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 37, 2 (2007), 262–272.
- DYN. Zotkin, Jane Hwang, R. Duraiswaini, and Larry S. Davis. 2003. HRTF personalization using anthropometric measurements. In *2003 IEEE workshop on applications of signal processing to audio and acoustics (IEEE Cat. No. 03TH8684)*. Ieee, 157–160.
- Heng Zou, Hermann J Müller, and Zhuanghua Shi. 2012. Non-spatial sounds regulate eye movements and enhance visual search. *Journal of Vision* 12, 5 (2012), 2–2.

Chapter 7

Appendices

7.1 Appendix A

In this appendix, hardware details are provided regarding the devices employed throughout this thesis. In [Tables 7.1](#) and [7.2](#), we specify the hardware details of the Varjo head-mounted displays used during the user studies, both the Varjo Aero and the Varjo XR-4, respectively. In addition, [Table 7.3](#) contains the details of the audio headphones.

Table 7.1: Specifications of the Varjo Aero headset

Display Resolution	2880 × 2720 pixels per eye
Display Refresh Rate	90 Hz
Eye Tracking Frequency	200 Hz
Eye Tracking Accuracy	Sub-Degree

Table 7.2: Specifications of the Varjo XR-4 headset

Display Resolution	3840 × 3744 pixels per eye
Display Refresh Rate	90 Hz
Passthrough Cameras	2 × 20 Mpx
Eye Tracking Frequency	200 Hz
Eye Tracking Accuracy	Sub-Degree

Table 7.3: Specifications of the Beyerdynamic DT 770M headphones

Frequency Response	5 Hz - 30000 Hz
Noise Cancellation Technology	Passive
Nominal Impedance	80 Ohm
Sensitivity	105 dB/mW

7.2 Appendix B

In this appendix, we include the tables corresponding to all pairwise comparisons in the statistical analysis for both the main experiment and the basketball application covered in [Chapter 3](#). In [Tables 7.4](#) and [7.5](#), we show the statistical results of the main experiment regarding saccade latency and landing accuracy, respectively.

Table 7.4: Statistical analysis of the results of the main experiment. Effect of the analyzed factors on saccade latency.

Factor	t-value	p-value
ΔT	-4.032	5.64e-05
E	13.721	<2e-16
ΔS	-1.252	0.210479
E: ΔS	0.766	0.443796
ΔT :E	-3.891	0.000102
ΔT : ΔS	0.891	0.372795
ΔT :E: ΔS	-0.908	0.363684

Table 7.5: Statistical analysis of the results of the main experiment. Effect of the analyzed factors on saccade landing accuracy.

Factor	t-value	p-value
ΔT	-0.513	0.6082
E	8.198	3.27e-16
ΔS	-0.655	0.51224
E: ΔS	0.675	0.49957
ΔT :E	1.917	0.05525
ΔT : ΔS	2.150	0.05163
ΔT :E: ΔS	-2.121	0.05402

Following the results in these tables, we run post-hocs analysis for statistically significant factors. In [Table 7.6](#), we show pairwise comparisons (post-hocs) for each temporal shift considered.

Table 7.6: Statistical analysis of the results of the main experiment. Post-hocs for the ΔT factor on saccade latency.

Comparison	z-ratio	p-value
$\Delta T_0 - \Delta T_{100}$	8.826	<.0001
$\Delta T_0 - \Delta T_{200}$	15.725	<.0001
$\Delta T_0 - \Delta T_{300}$	19.665	<.0001
$\Delta T_0 - \Delta T_{400}$	21.027	<.0001
$\Delta T_{100} - \Delta T_{200}$	6.654	<.0001
$\Delta T_{100} - \Delta T_{300}$	10.507	<.0001
$\Delta T_{100} - \Delta T_{400}$	12.007	<.0001
$\Delta T_{200} - \Delta T_{300}$	3.897	0.0900
$\Delta T_{200} - \Delta T_{400}$	5.524	0.0601
$\Delta T_{300} - \Delta T_{400}$	1.692	0.4388

In [Table 7.7](#), we show pairwise comparisons (post-hocs) for the interaction of ΔT and eccentricity. We do not run post-hocs for the eccentricity, since it only has two levels.

Table 7.7: Statistical analysis of the results of the main experiment. Post-hocs for the ΔT *Eccentricity interaction on saccade latency.

Comparison	z-ratio	p-value
$\Delta T_0 E_8 - \Delta T_{100} E_8$	5.658	<.0001
$\Delta T_0 E_8 - \Delta T_{200} E_8$	9.569	<.0001
$\Delta T_0 E_8 - \Delta T_{300} E_8$	12.014	<.0001
$\Delta T_0 E_8 - \Delta T_{400} E_8$	12.689	<.0001
$\Delta T_{100} E_8 - \Delta T_{200} E_8$	3.850	0.046
$\Delta T_{100} E_8 - \Delta T_{300} E_8$	6.211	<.0001
$\Delta T_{100} E_8 - \Delta T_{400} E_8$	7.027	<.0001
$\Delta T_{200} E_8 - \Delta T_{300} E_8$	2.333	0.3675
$\Delta T_{200} E_8 - \Delta T_{400} E_8$	3.244	0.0691
$\Delta T_{300} E_8 - \Delta T_{400} E_8$	0.974	0.9937
$\Delta T_0 E_{16} - \Delta T_{100} E_{16}$	6.709	<.0001
$\Delta T_0 E_{16} - \Delta T_{200} E_{16}$	12.659	<.0001
$\Delta T_0 E_{16} - \Delta T_{300} E_{16}$	15.751	<.0001
$\Delta T_0 E_{16} - \Delta T_{400} E_{16}$	17.035	<.0001
$\Delta T_{100} E_{16} - \Delta T_{200} E_{16}$	5.541	0.046
$\Delta T_{100} E_{16} - \Delta T_{300} E_{16}$	8.605	<.0001
$\Delta T_{100} E_{16} - \Delta T_{400} E_{16}$	9.924	<.0001
$\Delta T_{200} E_{16} - \Delta T_{300} E_{16}$	3.169	0.0492
$\Delta T_{200} E_{16} - \Delta T_{400} E_{16}$	4.565	0.0002
$\Delta T_{300} E_{16} - \Delta T_{400} E_{16}$	1.415	0.9230

Finally, in [Tables 7.8](#) and [7.9](#), we show the statistical results of the basketball application case regarding saccade latency and landing accuracy, respectively. [Table 7.10](#) shows the post-hoc analysis for the effect of ΔT on saccade latency for the basketball application case.

Table 7.8: Statistical analysis of the results of the basketball application. Effect of the analyzed factors on saccade latency.

Factor	t-value	p-value
ΔT	-12.694	<2e-16
ΔS	-1.745	0.0812
$\Delta T:\Delta S$	0.182	0.8558

Table 7.9: Statistical analysis of the results of the basketball application. Effect of the analyzed factors on saccade landing accuracy.

Factor	t-value	p-value
ΔT	-0.107	0.9149
ΔS	-1.081	0.2799
$\Delta T:\Delta S$	0.761	0.4469

Table 7.10: Statistical analysis of the results of the basketball application. Post-hocs for the ΔT factor on saccade latency.

Factor	t-value	p-value
$\Delta T_0 - \Delta T_{100}$	5.680	<.0001
$\Delta T_0 - \Delta T_{200}$	10.181	<.0001
$\Delta T_0 - \Delta T_{300}$	13.104	<.0001
$\Delta T_0 - \Delta T_{400}$	13.737	<.0001
$\Delta T_{100} - \Delta T_{200}$	4.112	0.0004
$\Delta T_{100} - \Delta T_{300}$	7.102	<.0001
$\Delta T_{100} - \Delta T_{400}$	8.027	<.0001
$\Delta T_{200} - \Delta T_{300}$	3.177	0.0132
$\Delta T_{200} - \Delta T_{400}$	4.300	0.0602
$\Delta T_{300} - \Delta T_{400}$	1.241	0.7272

7.3 Appendix C

In this appendix, we include the tables corresponding to all pairwise comparisons in the statistical analysis for the main experiment, the visual response experiment, and the final exploring application covered in [Chapter 4](#). In [Table 7.11](#), we show the significance of each factor on the success ratio in the main analysis.

Table 7.11: Statistical analysis of the results of the main experiment. Effects of the analyzed factors on the success ratio.

Factor	z-ratio	p-value
Matching Depth	-2.047	0.0406
Overdriving level 1	-5.046	< .0001
Overdriving level 2	-6.080	< .0001
Overdriving level 3	-7.473	< .0001
Vergence Direction	6.102	< .0001
Eccentricity	-3.288	0.001
Vergence Range	3.447	< .0001
Side	-0.390	0.6962

Following the results of that table, we run post-hocs analysis for statistically significant differences between *loudness shift* (ΔL) levels. Pairwise comparisons for each sound *loudness shift* considered are shown in [Table 7.12](#). We do not run post-hocs for *vergence direction* (V), *vergence range* (R) and *eccentricity* (E) as they have two levels.

Table 7.12: Statistical analysis of the results of the main experiment. Post-hocs for the *loudness shift* factor (ΔL) on success ratios.

Comparison	z-ratio	p-value
Colocated - Matching Depth	2.047	0.243
Colocated - Overdriving Level 1	5.046	< .0001
Colocated - Overdriving Level 2	6.080	< .0001
Colocated - Overdriving Level 3	7.473	< .0001
Matching Depth - Overdriving Level 1	3.034	0.020
Matching Depth - Overdriving Level 2	4.089	< .001
Matching Depth - Overdriving Level 3	5.516	< .0001
Overdriving Level 1 - Overdriving Level 2	1.069	0.822
Overdriving Level 1 - Overdriving Level 3	2.523	0.085
Overdriving Level 2 - Overdriving Level 3	1.457	0.590

Binomial test comparisons are performed between all sound modulation levels and the chance level (0.5) for each variable considered in the experiment, in order to find out if the proportion of success ratios for each condition is significantly different from the chance level. These tests are shown in [Tables 7.13](#) to [7.18](#).

Table 7.13: Statistical analysis of the results of the main experiment. Binomial test between *loudness shift* (ΔL) cases and Chance Rate for $E = 8$ condition.

Sound Level	p-value
Colocated	< .0001
Matching Depth	< .0001
Overdriving Level 1	< .0001
Overdriving Level 2	< .0001
Overdriving Level 3	0.0018

Table 7.14: Statistical analysis of the results of the main experiment. Binomial test between *loudness shift* (ΔL) cases and Chance Rate for $E=16$ condition.

Sound Level	p-value
Colocated	< .0001
Matching Depth	< .0001
Overdriving Level 1	< .001
Overdriving Level 2	0.0332
Overdriving Level 3	0.113

Table 7.15: Statistical analysis of the results of the main experiment. Binomial test between *loudness shift* (ΔL) cases and Chance Rate for Convergence condition.

Sound Level	p-value
Colocated	< .0001
Matching Depth	< .0001
Overdriving Level 1	0.00125
Overdriving Level 2	0.0433
Overdriving Level 3	0.702

Table 7.16: Statistical analysis of the results of the main experiment. Binomial test between *loudness shift* (ΔL) cases and Chance Rate for Divergence condition.

Sound Level	p-value
Colocated	< .0001
Matching Depth	< .0001
Overdriving Level 1	< .0001
Overdriving Level 2	< .0001
Overdriving Level 3	< .0001

Regarding the visual response experiment, the results of the analyses are shown in [Tables 7.19 to 7.21](#), showing the effect of each variable in the response times recorded for the first target (T_1), second target (T_2) and the combination of both ($T_1 + T_2$), respectively.

Table 7.17: Statistical analysis of the results of the main experiment. Binomial test between *loudness shift* (ΔL) cases and Chance Rate for Near range condition.

Sound Level	p-value
Colocated	< .0001
Matching Depth	< .0001
Overdriving Level 1	< .0001
Overdriving Level 2	0.0118
Overdriving Level 3	0.0139

Table 7.18: Statistical analysis of the results of the main experiment. Binomial test between *loudness shift* (ΔL) cases and Chance Rate for Far range condition.

Sound Level	p-value
Colocated	< .0001
Matching Depth	< .0001
Overdriving Level 1	< .0001
Overdriving Level 2	< .0001
Overdriving Level 3	0.0251

Table 7.19: Statistical analysis of the results of the validation test. Effects of the analysed factors on response times for T_1 , first target appearing.

Factor	z-ratio	p-value
$\Delta 2^\circ$	2.732	0.006
$\Delta 4^\circ$	6.196	< .0001
Eccentricity	5.895	< .0001
Side	1.504	0.133
First Vergence Direction	0.644	0.519

Table 7.20: Statistical analysis of the results of the validation test. Effects of the analysed factors on response times for T_2 , second target appearing.

Factor	z-ratio	p-value
$\Delta 6^\circ$	4.675	< .0001
$\Delta 8^\circ$	4.979	< .0001
Eccentricity	1.796	0.072
Side	-0.660	0.509
First Vergence Direction	-0.766	0.444

Finally, in [Table 7.22](#), we show the statistical results of the application case in evaluating factor levels significance on the success ratio of selecting the scene item colocated with the annotation. Binomial test comparisons between each condition's success ratio and chance level (0.5) are also shown in [Table 7.23](#). The same analysis is performed regarding the application case in MR, whose results are displayed in [Tables 7.24](#) and [7.25](#), respectively.

Table 7.21: Statistical analysis of the results of the validation test. Effects of the analysed factors on response times for $T_1 + T_2$, considering total trial response time.

Factor	z-ratio	p-value
$\Delta 8^\circ$	4.649	< .0001
$\Delta 12^\circ$	7.004	< .0001
Eccentricity	4.871	< .0001
Side	0.547	0.0585
First Vergence Direction	-0.066	0.948

Table 7.22: Statistical analysis of the results of the application case. Effects of the analysed factors on the success ratio of selecting the scene item at the same visual depth as the annotation.

Factor	z-ratio	p-value
Sound Level	-10.375	< .0001
Vergence Direction (D)	0.945	0.3449
Eccentricity (16)	0.601	0.5475

Table 7.23: Statistical analysis of the results of the application case. Binomial test between the association ratios for each sound case and the chance rate (0.5) for each scene item condition.

Annotation	Sound Level	t_stat	p-value
Bike	Colocated	0.2223	0.8247
Bike	Compensated	-18.3518	< .0001
Cat	Colocated	7.6024	< .0001
Cat	Compensated	0.6685	0.5058
Kid	Colocated	10.9000	< .0001
Kid	Compensated	-0.8933	0.3744
Toy	Colocated	-2.0525	0.0434
Toy	Compensated	-16.0645	< .0001

Table 7.24: Statistical analysis of the results of the application case in MR. Effects of the analysed factors on the success ratio of selecting the scene item at the same visual depth as the annotation.

Factor	z-ratio	p-value
Sound Level	-6.749	< .0001
Vergence Direction (D)	2.280	0.023
Eccentricity (16)	3.257	0.001

Table 7.25: Statistical analysis of the results of the exploring application in MR. Binomial test between the association ratios for each sound case and the chance rate (0.5) for each scene item condition.

Annotation	Sound Level	t_stat	$p - value$
Bird	Colocated	0.9475	0.3492
Bird	Compensated	-2.3333	0.0249
Phone	Colocated	10.0767	< .0001
Phone	Compensated	1.2748	0.2099
Speaker	Colocated	10.0767	< .0001
Speaker	Compensated	0.3126	0.7562
Toy	Colocated	8.3267	< .0001
Toy	Compensated	0.6276	0.5339