# PSYCHOPHYSICAL METHODS FOR ENHANCING IMMERSIVE GRAPHICS SYSTEMS

**DISSERTATION**

Submitted in Partial Fulfillment of

the Requirements for

the Degree of

DOCTOR OF PHILOSOPHY (Computer Science)

at the

NEW YORK UNIVERSITY
TANDON SCHOOL OF ENGINEERING

by

Budmonde Duinkharjav

May 2025

# PSYCHOPHYSICAL METHODS FOR ENHANCING IMMERSIVE GRAPHICS SYSTEMS

DISSERTATION

Submitted in Partial Fulfillment of

the Requirements for

the Degree of

DOCTOR OF PHILOSOPHY (Computer Science)

at the

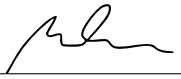
NEW YORK UNIVERSITY
TANDON SCHOOL OF ENGINEERING

by

Budmonde Duinkharjav

May 2025

Approved:



Department Chair Signature

May 1, 2025

Date

Approved by the Guidance Committee:

Major: Computer Science

---

**Qi Sun**
Assistant Professor
NYU Tandon School of Engineering

Qi Sun

Date
April 30, 2025

---

**Kenneth Perlin**
Professor
NYU Courant Institute of Mathematical Sciences

Kenneth Perlin

Date
April 28, 2025

---

**Daniele Panozzo**
Associate Professor
NYU Courant Institute of Mathematical Sciences

Date
April 27, 2025

---

**Claudio Silva**
Professor
NYU Tandon School of Engineering

Date
April 30, 2025

---

**Gordon Wetzstein**
Associate Professor
Stanford University

Date
April 30, 2025

# Vita

Budmonde Duinkharjav was born in Erdenet, Mongolia. He earned his B.S. and M.Eng. degrees in Computer Science and Engineering from the Massachusetts Institute of Technology (MIT) in 2018 and 2019, respectively. In January 2021, he enrolled in the Ph.D. program in the Department of Computer Science and Engineering at NYU Tandon School of Engineering, where he conducted all of the research presented in this dissertation.

His Ph.D. research was partially supported by National Science Foundation grants #2225861, #2232817, and #2044963, as well as by DARPA PTG and ICS programs. During his doctoral studies, he also collaborated with researchers at NVIDIA during the summer terms of 2022 and 2024, and at Adobe in the summer of 2023.

## List of First-Authored Publications

*SIGGRAPH Asia 2024* Evaluating Visual Perception of Object Motion in Dynamic Environments
***Duinkharjav**, Kang, Miller, Xiao, Sun*

*SIGGRAPH Asia 2023* The Shortest Route Is Not Always the Fastest: Probability-Modeled Stereoscopic Eye Movement Completion Time in VR
***Duinkharjav**, Liang, Patney, Brown, Sun*

*SIGGRAPH Asia 2022* Color-Perception-Guided Display Power Reduction for Virtual Reality
***Duinkharjav***, Chen*, Tyagi, He, Zhu, Sun (* co-authors)*

*SIGGRAPH 2022* Image Features Influence Reaction Time: A Learned Probabilistic Perceptual Model for Saccade Latency
***Duinkharjav**, Brown, Chakravarthula, Patney, Sun*
**Best Paper**

## List of Contributing Publications

*IEEE VR 2025* FovealNet: Advancing AI-Driven Gaze Tracking Solutions for Efficient Foveated Rendering in Virtual Reality
*Liu, **Duinkharjav**, Sun, Zhang*

*ASPLOS 2024* Exploiting Human Color Discrimination for Memory-and Energy-Efficient Image Encoding in Virtual Reality
*Ujjainkar, Shahan, Chen, **Duinkharjav**, Sun, Zhu*

*SIGGRAPH 2023* Imperceptible Color Modulation for Power Saving in VR/AR
*Chen, **Duinkharjav**, Ujjainkar, Shahan, Tyagi, He, Zhu, Sun*

*SID Display Week 2022* Modeling And Optimizing Human-In-The-Loop Visual Perception Using Immersive Displays: A Review
*Sun, **Duinkharjav**, Patney*

*JID 2022* Reconstructing Room Scales With a Single Sound for Augmented Reality Displays
*Liang, Liang, Roman, Weiss, **Duinkharjav**, Bello, Sun*

*ISMAR 2022* FoV-NeRF: Foveated Neural Radiance Fields for Virtual Reality
*Deng, He, Ye, **Duinkharjav**, Chakravarthula, Yang, Sun*
**Best Journal Paper**

*IEEE VR 2022* Instant Reality: Gaze-Contingent Perceptual Optimization for 3D Virtual Reality Streaming
*Chen, **Duinkharjav**, Sun, Wei, Petrangeli, Echevarria, Silva, Sun*

# Acknowledgements

First and foremost, I would like to thank my advisor, Qi Sun. We began this journey together—new professor and new student—and your mentorship has been instrumental in shaping my research path. You've taught me to approach research holistically, encouraging me to focus on the broader impact of my work rather than getting lost in the minutiae.

Thank you, Anjul and Rachel, for introducing me to human perception research and guiding me during my early days. Praneeth, Yuhao and Ruth, your critical questions and challenges have significantly affected my trajectory towards becoming a better researcher. I'm grateful to my teams at NVIDIA and Adobe for providing such enriching research environments during my internships. Joohwan, you've been the best manager I could've asked for. Loudon and Chang, your mentorship has been invaluable. To Ken, Daniele, Claudio, and Gordon—thank you for serving on my PhD committee. Your feedback has played a key role in shaping this dissertation, and I'm hopeful it reflects the best of our collective efforts. I'd also like to thank the CUSP and CSE communities for fostering such a welcoming environment, and the Immersive Computing Lab cohort for the great conversations and boba breaks. To the Carrasco lab members—thank you for showing me what rigorous vision science research looks like.

To my friends and family— Nancy, James, Cynthia, and Charlotte: thank you for helping me move to NYC during the tail end of COVID. Mom, Dad, Enkhlen, and Sunder—your unwavering support has meant everything. Thank you for being my foundation through the ups and downs of this journey. And lastly, to Sophia—thank you for standing by me through it all, in both the best and hardest moments. I honestly don't think I could have finished this in one piece without you.

To my friends and family, without whom I would not have reached this milestone.

## ABSTRACT

**PSYCHOPHYSICAL METHODS FOR ENHANCING IMMERSIVE GRAPHICS SYSTEMS**

**by**

**Budmonde Duinkharjav**

**Advisor: Prof. Qi Sun, Ph.D.**

**Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy (Computer Science)**

**May 2025**

Insights into how humans perceive and react to their visual surroundings have driven advancements in computer graphics, improving the efficiency and fidelity of display and rendering technologies. Computational models that capture the capabilities, limitations, and nuances of human vision have revealed numerous optimization strategies that enhance system performance without perceptible degradation in user experience. The emergence of applications with complex computational capabilities and human interaction-aware technologies—such as XR, assisted driving, video games, and

esports—has not only introduced new opportunities for optimizing graphics but also for enhancing human performance, productivity, and safety beyond conventional limits.

In this PhD dissertation, we investigate various aspects of the human visual system and develop computational models and algorithms that complement perceptual and behavioral constraints to enhance user experience. We explore topics such as leveraging color encoding limitations to optimize display output, identifying and correcting inaccuracies in motion perception, and measuring human decision-making and motor control latency to assess the temporal effects of displayed imagery.

Through this work, we demonstrate how psychophysical methodologies, originally designed to study human perception and behavior, can be applied to understanding human-computer joint systems. By addressing inefficiencies, bottlenecks, and inaccuracies within this system, we show how computers can be improved to reduce power consumption, computation, and bandwidth, while human users can be enhanced in speed and accuracy.

# Table of Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

Immersive graphics systems rely on an accurate understanding of human perception and behavior to ensure a high-fidelity experience, and serve as an effective intermediary between computer systems and users, facilitating seamless information exchange. Therefore, new technological advancements in display and rendering pipelines often trigger research into the boundaries of human interaction with the imagery and downstream applications enabled by these advancements. Moreover, research into how humans perceive and react within these new paradigms has also led to new understandings of human vision and cognition. Additionally, it prompts consideration of whether we can move past merely optimizing graphics systems for their *own* performance toward designing systems that also enhance *human* performance beyond typical capabilities.

The field of perceptual graphics encompasses research aimed at integrating the latest findings in human perception and vision science into state-of-the-art display and rendering technologies, and thus, plays a pivotal role in addressing these inquiries. As a crucial linchpin, this field also drives further research in both areas by addressing technological gaps on the computer systems side as well as scientific gaps in human

perception. The emergence of newly available display technologies such as *virtual/augmented reality* (VR/AR), advanced rendering techniques like *neural radiance fields* (NeRFs) and 3D Gaussian Splatting, and proliferation of new application domains—from video games and esports to remote and assisted vehicle driving—underscores the need for detailed analysis through the lens of human perception and cognition. Hence, there is a pressing need for research aimed at understanding the perceptual factors relevant within state-of-the-art immersive graphics pipelines.

Our research aims is to utilize methods for measuring and understanding aspects of human perception and behavior relevant to these emerging technologies. We seek to apply these insights to enhance the performance of underlying computer systems, and conversely, improve human performance in interactive tasks beyond typical capabilities outside of the immersive graphics systems. To this end, this dissertation presents methods that leverage psychophysics, an essential tool in psychology and neuroscience, to design and implement computer systems that advance both fields. Our approach relies on identifying and measuring features and limitations of human perception, and utilize our findings to develop computational frameworks integrated into downstream applications. These frameworks aim to minimize the impact of system limitations experienced by users and enhance users' ability to navigate and interact effectively within immersive virtual environments.

Throughout this dissertation, we focus on how the brain processes visual signals, explore the underlying neural circuitry and scientific understanding of its end-to-end functions in human perception and behavior, and ultimately propose methods for holistically improving the efficiency of human-computer systems. Just as optimizing computer system pipelines requires tracing critical pathways of information transmission, we take a similar approach in this work. In the following chapters, we provide an overview

of the human visual pathway, examining how neural signals are transmitted throughout the brain. We identify areas where the accuracy and latency of these signals can be enhanced, as well as where inherent bandwidth limitations can be leveraged for system-side optimizations.

Specifically, in Chapter 3, we first examine how the conversion of light signals within the retina and their subsequent transmission to the brain exhibit bandwidth limitations—and how these limitations can be exploited to optimize computer systems. By accounting for constraints in neural signal transmission, we minimize the generation of visual information that would otherwise be discarded along the visual pathway, leading to more efficient system design.

In Chapter 4, we continue exploring the visual information transmission pathway, investigating how the visual system integrates multiple sources of low-level information to construct a comprehensive understanding of 3D environments. We also examine how different display conditions degrade perception and discuss potential strategies to mitigate these effects. In Chapter 5, we analyze the mechanics of decision-making, exploring how our perception of displayed imagery affects reaction times and how optimizing content appearance can enhance human performance. Finally, in Chapter 6, we examine how eye movement decisions are executed via control signals transmitted back to the eyes. The accuracy and efficiency of eye movement control influence how different scene layouts affect our ability to quickly scan our visual surroundings, as well as the costs associated with visually interacting with virtual environments. Across these chapters, we illustrate the application of computational frameworks for improving performance in various contexts, including VR/AR headsets and 2D display environments, demonstrating their practical utility.

# Chapter 2

# Background Literature

In this chapter, we review the relevant literature and establish the mathematical and computational frameworks that underpin this work. We begin by surveying prior research in computer graphics that incorporates human perceptual and behavioral factors. Next, we introduce the psychophysical framework for studying human perception and behavior, which informs the modeling approaches discussed throughout this manuscript. Finally, we provide an overview of the human visual system, covering existing psychophysical models aimed at understanding various aspects of visual information processing, along with key neural and physiological evidence that support these models.

## 2.1 Immersive Graphics Systems

This section reviews existing literature on how computer graphics systems incorporate human perception and behavior, how the constraints of both humans and computers influence hardware and graphics system design, and how human behaviors create opportunities for optimization and enhancement in human-computer interaction.

### 2.1.1 Human Vision-Aware Display Systems

Display systems leverage the limitations of human vision to bridge the gap between natural and displayed visual content in terms of fidelity and realism. For instance, as display resolution surpasses the perceptual resolution limit, we lose the ability to discern whether a visual target is a cohesive object or a collection of individual pixels depicting it [Campbell and Robson, 1968]. Since the primary goal of display systems is to present visual content to human users, their specifications (e.g., resolution, luminance, color gamut) are determined by both the constraints of display hardware and the limitations of human vision.

The spatial resolution at the center of human vision is approximately 120 *pixels-per-degree* (ppd) [Campbell and Robson, 1968], meaning that the display resolution required to match human limits varies depending on the viewing angle of the display, a.k.a., *field-of-view* (fov). While high-end TVs provide sufficient resolution for eye-display distances (which affect the fov) of $\geq$ 1.5 m[1], state-of-the-art AR/VR displays, with resolutions of up to 51 ppd[2], have yet to reach the limits of human vision.

Other aspects of display system capabilities are still short of human vision limits. Human eyes can adapt to luminance levels ranging from $10^{-6}$ to $10^{6}$ nits, spanning approximately 12 log units [Wang and Zhao, 2022], while preserving a static contrast sensitivity of around $1:100$ across the whole luminance range [Barten, 1999a]. However, state-of-the-art high dynamic range display systems only reach up to 4000 nits [3] while maintaining minimum contrast sensitivity across all adaptation levels. Similarly, human color vision, as specified by the CIE 1931 colorimetric standards [Smith and Guild, 1931]

[1] Estimated based on the 117 pixels-per-inch resolution reported by Samsung (`https://www.samsung.com/africa_en/tvs/tv-buying-guide/what-is-8k-tv/`)

[2] `https://varjo.com/products/xr-4/`

[3] `https://www.lgcorp.com/media/release/28575`

(further details in Section 2.3.2), exceeds the color gamuts of any state-of-the-art display systems available today [Chen et al., 2017]. Moreover, emerging display technologies, particularly in AR displays, introduce new challenges in color generation, such as color blending and color alignment [Hassani and Murdoch, 2016; Murdoch et al., 2015; Zhang et al., 2021a].

Beyond the physical constraints of display hardware, additional computational requirements further limit the visual fidelity of displayed content. The graphics rendering pipeline demands substantial computation to operate in real time. In fact, research on user preferences in low-latency demanding applications, such as competitive video games, has shown that task-crucial visual factors like latency often dictate the required graphics settings [Claypool et al., 2006]. Maintaining such performance requirements consistently necessitates high computational power and significant energy consumption. These constraints pose challenges for mobile displays, as achieving high computational performance and power storage becomes increasingly difficult within the form factors required for such devices. As a result, energy- and compute-aware methods for displaying content have gained interest in the literature, focusing on reducing the rendering power of graphics algorithms [Debattista et al., 2018; Wang et al., 2016; Zhang et al., 2018, 2021b], exploring the relationship between display power, luminance and color in mobile computing [Dash and Hu, 2021; Dong et al., 2009; Dong and Zhong, 2011; Shye et al., 2009], analyzing the effects of display luminance on human vision [Shye et al., 2009; Yan et al., 2018], and investigating the impacts of hardware design optimizations on display power consumption [Boroson et al., 2009; Miller et al., 2007, 2008, 2006; Shin et al., 2013].

A common theme across research on display power reduction is the quantification of power savings per unit change in display luminance [Shye et al., 2009], chrominance,

and hue [Dash and Hu, 2021; Dong et al., 2009; Dong and Zhong, 2011]. Studies have proposed measuring changes in display characteristics in terms of human physiological [Shye et al., 2009] and perceptual responses [Mantiuk et al., 2021, 2024] (further details in Section 2.2.1). Perceptually driven power-saving strategies have influenced display hardware design, leading to the integration of color-transformation functions into hardware [Shin et al., 2013] and even the introduction of four-color OLED structures [Miller et al., 2007, 2008].

Ultimately, the design objectives, constraints, and methodologies that shape display systems are driven by human visual perception—what the observer can and cannot see—how displayed content influences perception, and how users interact with digital content efficiently and effectively.

### 2.1.2 Gaze-Contingent Computer Graphics

In Section 2.1.1, we discussed how display systems are designed to accommodate limitations of the human visual system. However, the perceptual requirements we outlined do not fully capture the complexities of human vision. Vision is most sensitive at the center of the visual field, a.k.a., the fovea, and degrades toward the periphery (further details in Section 2.3.1). Examples of limitations include reduced spatial resolution in peripheral vision [Watson, 2014], diminished color perception [Cohen et al., 2020], and reduced sensitivity to flicker [Tyler, 1987]. Unfortunately, traditional display systems are unable to leverage these characteristics and must instead optimize the entire display as if every region were subject to foveal scrutiny.

Gaze-contingent computer graphics methods seek to exploit this overlooked aspect of human vision by incorporating high-speed eye tracking to determine the user's gaze location and using this information to optimize graphics algorithms and display systems.

One of the most significant optimizations enabled by eye tracking is gaze-contingent rendering. These methods dynamically adjust rendering algorithms to enhance the perceived realism of displayed content, improving effects such as parallax [Konrad et al., 2020], ocular and stereo depth perception [Krajancich et al., 2020; Sun et al., 2020], depth of field [Duchowski et al., 2014; Hillaire et al., 2008; Mauderer et al., 2014]. By leveraging the perceptual differences between foveal and peripheral vision, these approaches also enhance interactive computer graphics techniques, forming the foundation for gaze-contingent foveated rendering and display systems.

Early implementations of foveated methods relied on reducing the sampling rate of pixels in the peripheral visual field to decrease computational bandwidth and rendering latency [Guenter et al., 2012; Meng et al., 2018; Patney et al., 2016]. Further research has demonstrated that image statistics [Kaplanyan et al., 2019; Tursun et al., 2019; Walton et al., 2021], temporal frame re-use [Franke et al., 2021], dynamic level-of-detail [Chen et al., 2022], and variable shading rates [Denes et al., 2020; Jindal et al., 2021] can further refine foveated rendering. The same fundamental concept has also been applied to develop foveated path tracing [Koskela et al., 2019, 2016; Polychronakis et al., 2021; Weier et al., 2016], light field displays [Sun et al., 2017], neural rendering [Deng et al., 2022], and gaussian splatting [Lin et al., 2025].

In summary, the integration of eye-tracking technology into immersive display systems and graphics methods has unlocked a wealth of technological optimizations and enhancements to the user experience.

### 2.1.3 Behavior-Aware Computer Graphics

The use of eye tracking in immersive computer graphics applications represents only the beginning of how graphics systems can be adapted to user needs, with significant

untapped potential for enhancing applications to complement the limitations of human perception and behavior. Recent research has shown that a deeper understanding of these limitations can lead to more user-friendly and perceptually optimized graphics applications, significantly enriching this field.

For example, insights into human eye movement behaviors have enabled the development of predictive foveation techniques that anticipate gaze shifts to optimize rendering performance [Arabadzhiyska et al., 2017; Kwak et al., 2024]. Incorporating research on multimodal stimulus integration into graphics applications has led to novel approaches, such as manipulating users' spatial perception of their surroundings [Bernal-Berdun et al., 2024] and accelerating human reaction times [Jiménez Navarro et al., 2024; Peng et al., 2024]. Studies on behavioral correlates of cyber-sickness [Tovar et al., 2024] and its mitigation [Hu et al., 2019; Park et al., 2022], along with research into the perceptual requirements for gaze-contingent distortion correction [Guan et al., 2022] and world-locked rendering [Guan et al., 2023; Lutwak et al., 2023], further highlight the challenges of bridging the gap between immersive display systems such as AR/VR and real-world perception.

Additionally, research on the effects of different display systems on human perception and behavior has helped refine their applications for specific tasks. In particular, scene layout, depth perception, and motion understanding within immersive virtual environments are influenced by a combination of visual [Didyk et al., 2011; Lutwak et al., 2022; Murray, 1994] and non-visual cues, such as vestibular input [DeAngelis and Angelaki, 2012]. These findings suggest that for tasks requiring accurate depth and motion estimation, immersive display systems such as AR/VR can enhance human perceptual performance [Xie et al., 2020a; Xing and Saunders, 2022]. Studies on how gaze behavior changes in AR/VR environments have also provided valuable insights for

improving the design of these hardware systems [Aizenman et al., 2022; Shi et al., 2022].

Ultimately, a major research goal in advancing computer graphics algorithms and display systems is to align them more closely with human perception and behavior, ensuring they better serve the users for whom they are designed.

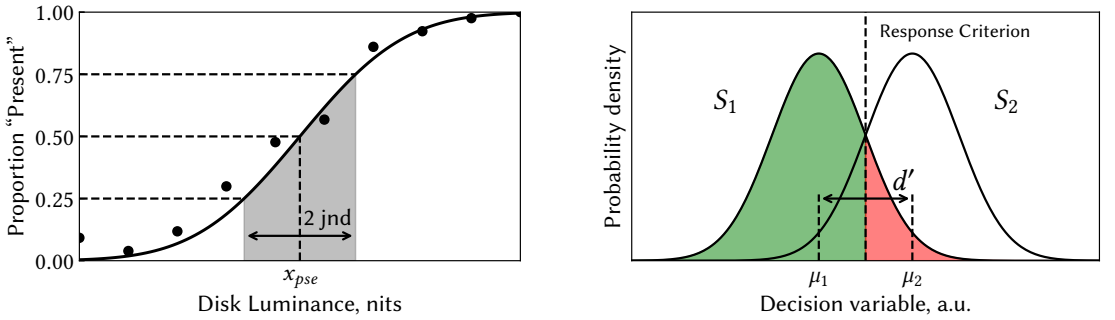
## 2.2 Psychophysical Methods

Psychophysics is the scientific discipline concerned with quantitatively examining the relationship between physical stimuli and the sensations and perceptions they elicit. In the context of vision, when we observe a visual stimulus, the physical signal is first detected by photoreceptors in the retina and subsequently transmitted through various neural pathways to the brain, where it gives rise to sensory experiences and perceptual judgments. As we will discuss in Section 2.3, while physiological experiments provide insight into the structure and function of these neural pathways, psychophysics is primarily concerned with drawing end-to-end inferences about how physical stimuli influence perception and behavior [Bruce et al., 2014].

Throughout this dissertation, we employ a range of psychophysical techniques in experimental design, numerical analysis, and computational modeling. This section provides essential background on these methods, as well as a review of relevant literature that informs the research presented in this work.

### 2.2.1 Perceptual Discrimination

**The Psychometric Function.** When investigating how humans perceive physical stimuli and what internal representations underlie perceptual judgments, simple perceptual discrimination tasks provide valuable insights [Hautus et al., 2021, Chapter 1].




(a) the psychometric function

(b) signal detection theory

Figure 2.1: *Psychophysics measures.* (a) Hypothetical proportion "present" responses are plotted as scatter points in a solid disk detection task. A cumulative Gaussian psychometric function is fitted to the data, with the *point of subjective equivalence* (PSE) and *just-noticeable difference* (jnd) values annotated. (b) The Signal Detection Theory (SDT) decision variable and its relationship to psychophysical responses are illustrated. The decision variable associated with stimulus, $S_1$, is randomly sampled from its probability distribution. Depending on the location of the response criterion, the sampled value results in either a correct (shaded green), or incorrect (shaded red) response. Stimulus $S_2$ is sampled in the same manner (results not visualized). The distance between the means of the decision variable distributions for the stimuli represents the sensitivity, $d'$, in discriminating between them.

For example, consider an experiment designed to measure the visibility of a solid gray disk presented against a solid gray background of lower luminance (measured in nits). By repeatedly asking subjects whether they perceived the disk on each trial, and systematically varying the disk luminance across different levels, we can record the proportion of "present" responses at each stimulus luminance level, $x$. Plotting these response proportions as a function of stimulus luminance yields the *psychometric function*, as illustrated in Figure 2.1a [Hautus et al., 2021, Chapter 4].

The psychometric function aims to quantitatively describe how perception transitions between two perceptual states—specifically, from being unable to detect a stimulus to reliably detecting it. For most perceptual tasks, the psychometric function exhibits a characteristic sigmoidal shape [Woodworth and Schlosberg, 1954]. Accordingly, analyses of psychometric functions typically focus on two key attributes: the intercept and

the slope [Luce et al., 1963].

The intercept, known as the *point of subjective equivalence* (PSE), corresponds to the stimulus level $x_{pse}$ at which the subject reports detecting the stimulus 50% of the time, that is $p(x_{pse}) = 0.5$. The slope is typically quantified as half the difference in stimulus levels between the 25th and 75th percentile response rates, and is referred to as the *just-noticeable difference* (jnd) [Hautus et al., 2021; Luce et al., 1963].

When quantifying perceptual thresholds, the PSE is often used because it is reflects the stimulus level at which the subject has an equal probability of reporting the stimulus as present or absent. In the case of the gray disk detection task, the PSE represents the minimum disk luminance at which the observer has a 50% chance of detecting the disk. The jnd, on the other hand, reflects the increment in luminance required to increase the observer's detection probability from 50% to 75%.

Empirically measured psychometric functions are typically modeled using analytic fits based on well-known cumulative distribution functions [Hautus et al., 2021; Wichmann and Hill, 2001]. One common model is the cumulative normal distribution:

$$p(x) = \Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] dx, \tag{2.1}$$

which is used throughout this work. Alternatively, psychometric functions are also frequently modeled using the logistic cumulative distribution or the Weibull distribution (formulations omitted here for brevity; see Hautus et al. [2021, Chapter 11] for details). In the case of the cumulative normal model, the PSE corresponds directly to the distribution mean, $\mu$, while the jnd is proportional to the standard deviation, specifically $0.675 \times \sigma$. The PSE for normal distribution models is equivalent to the distribution mean, $\mu$, while the jnd equals a fraction of the standard deviation, $0.675 \times \sigma$.

Additionally, depending on the experimental protocol, psychometric functions are often adjusted to account for a non-zero guess rate, denoted as $\gamma$ [Wichmann and Hill, 2001]. This adjustment reflects the probability of a correct response due to chance, independent of the perceptual signal. The resulting observed psychometric function is defined as:

$$p_{obs}(x) = \gamma + (1 - \gamma)\, p(x). \tag{2.2}$$

For example, in *two-alternative forced-choice* (2AFC) experiments, the task is to determine the order of two stimuli, such as $\langle S_1, S_2 \rangle$ or $\langle S_2, S_1 \rangle$. In this case, the chance-level performance corresponds to $\gamma = 0.5$, reflecting the 50% probability of a correct response when guessing [McKee et al., 1985].

**Adaptive Threshold Measurement.** In practice, however, determining the entire psychometric function through repeated trials across numerous stimulus levels is often impractical and time-consuming. For many applications, as we will see throughout this dissertation, it is sufficient to estimate the stimulus level corresponding to a specific performance threshold. To this end, adaptive threshold methods are commonly employed to reduce both the number of stimulus levels and the number of trial repetitions required for threshold estimation. Broadly, adaptive threshold measurement methods refer to a family of protocols in which the stimulus level of each trial is determined based on the responses of previous trials [Hautus et al., 2021, Chapter 11].

One of the simplest and most widely used of these protocols is the *up-down transformed-response* (UDTR), a.k.a., the adaptive staircase method [Wetherill and Levitt, 1965]. In this method, the sequence of prior responses is compared to a predefined rule to adjust the stimulus level. For instance, a 1-up-2-down protocol increases the

stimulus level after every incorrect response and decreases is after two consecutive correct responses. At steady state, the probability of increasing and decreasing the stimulus level equilibrates to 0.5. This behavior corresponds to a target task performance of $p = \sqrt{0.5} \approx 0.71$, since two consecutive correct responses occur with probability $p^2 = 0.5$. See Figure 3.1b for an example of such an adaptive staircase sequence. Similar protocols, such as 1-up-3-down and 1-up-4-down, target higher performance thresholds of $p \approx 0.79$ and $p \approx 0.84$, respectively. Beyond UDTR methods, more sophisticated adaptive procedures leverage maximum likelihood estimation to refine threshold estimates, such as PEST [Taylor et al., 1967] and QUEST [Watson and Pelli, 1983].

Parameters such as step size schedule, termination criteria, and threshold determination strategies differ across methods. In the experiments presented throughout this dissertation, step sizes were held constant within each experiment, termination criteria were based on a fixed number of reversals, and thresholds were determined by averaging the stimulus levels at reversal points. For a comprehensive overview of adaptive threshold measurement methods, see Hautus et al. [2021, Chapter 11].

**Signal Detection Theory.** Beyond the characterization of perceptual performance through psychometric functions, modern psychophysics builds heavily on foundational work on Signal Detection Theory (SDT) [Hautus et al., 2021]. The SDT framework provides a principled approach for analyzing not only the sensitivity of perceptual systems but also the decision-making processes underlying perceptual judgments. In particular, SDT offers insights into the nature of the *abstract* internal representations and cognitive strategies that mediate the relationship between physical stimuli and behavioral responses [Hautus et al., 2021, Chapter 1].

For example, consider the gray disk detection task described earlier. Within the SDT framework, it is assumed that, on each trial, the visual system generates a sample of an abstract *decision variable*, drawn from a normal distribution (illustrated in Figure 2.1b). A binary present/absent judgment is then made by comparing this sample against a fixed *decision criterion*; if the sample exceeds the criterion, the observer reports the disk as visible, otherwise it is reported as invisible [Blackwell, 1946]. This simple probabilistic model has been shown to accurately account for empirical patterns observed in perceptual detection experiments [Hautus et al., 2021].

In experimental designs that examine multiple target luminance levels, increasing the disk luminance results in a corresponding increase in the mean of the decision variable distribution, as illustrated in Figure 2.1b, while the variance of the distribution remains constant. This constant variance reflects an additive noise component, typically interpreted as internal noise accumulated along the visual processing pathway [Bruce et al., 2014]. Notably, applying a single, fixed decision criterion across all luminance conditions can successfully predict the proportion of correct judgments in each condition.

In this example, the decision variable represents an abstract measure of *visibility*, and together with the decision criterion, accounts for the trial-to-trial variability observed in subjects' judgments. Crucially, while the decision criterion may shift depending on experimental protocols or task instructions, the underlying decision variable remains determined by the physical properties of the visual stimulus [Hautus et al., 2021, Chapter 2]. This distinction reflects the fact that the criterion is influenced by the subject's response bias, whereas the decision variable captures stimulus-driven sensory information. In the SDT framework, the separation between decision variable distributions for different stimuli is quantified by the sensitivity index, $d'$, which measures the distance

between distributions in units of the system's internal noise standard deviation [Hautus et al., 2021, Chapter 1]. For example, two stimuli with $d' = 1$ are separated by one standard deviation of internal noise, while higher values of $d'$ indicate that the internal representation of the stimuli are more distinct, and exhibit less overlap. In that sense, in the context of detection tasks, given an experiment with unbiased responses, the jnd and $d'$ measures are linearly related as both indirectly measure the width of the decision variable [Hautus et al., 2021, Chapter 4].

### 2.2.2 Speeded Decision-Making

In the psychology literature, efforts to understand the timing characteristics of decision-making processes are commonly referred to as speeded decision-making. As discussed in the previous section, psychophysical modeling provides a framework for constructing internal representations of visual signals that underpin perceptual judgments and decision-making. Several modeling approaches have been proposed to describe speeded decision-making processes, in which an internal representation of *decision evidence* is accumulated stochastically over time. In these frameworks, evidence continues to accumulates until it reaches a pre-defined decision criterion, at which point an action or judgment is triggered [Mazurek et al., 2003].

The parallels between evidence accumulation models and SDT models are notable: the rate of evidence accumulation in these models, much like the decision variable distributions from SDT, have been shown to be modulated by stimulus characteristics [Bell et al., 2006; Carpenter, 2004; Mahadevan et al., 2018]. Similarly, the decision criterion serves as a measure of response bias [Reddi et al., 2003; Yamagishi and Furukawa, 2020] (cf. decision criterion in SDT).

Among evidence accumulation models, two of the most prominent are the Drift-

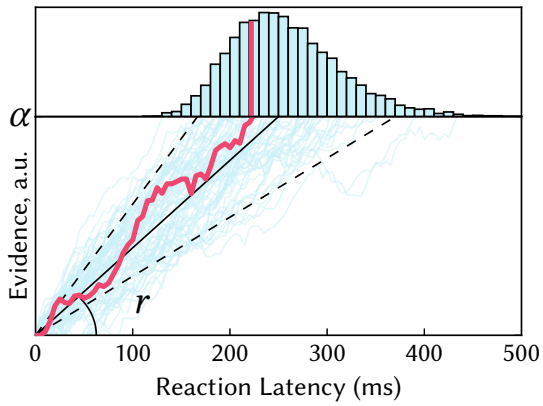



Figure 2.2: *The Drift Diffusion Model (DDM).* The $x$-axis represents time, and the $y$-axis represents accumulated evidence levels. Each decision follows a random walk process (e.g., the red trajectory), where the decision criterion $\alpha$ reflects response bias, and the drift rate $r$ determines the speed of evidence accumulation. A decision is triggered when the accumulated evidence reaches the pre-determined criterion $\alpha$. Due to cognitive noise, individual decisions vary (e.g., the light blue trajectories), making action timing a probabilistic event.

Diffusion Model (DDM) [Palmer et al., 2005; Ratcliff, 1978] which models the decision process as a stochastic random walk analogous to Brownian Motion, and the LATER model, which assumes that the rate of evidence accumulation is randomly sampled at the onset of a evidence accumulation and remains constant throughout the accumulation process [Carpenter and Williams, 1995; Reddi et al., 2003]. Throughout this work, we employ the DDM framework and therefore examine it in further detail below.

**The Drift Diffusion Model.** The DDM has been shown to be an effective and accurate model across a wide range of decision-making contexts in psychology and neuroscience, enabling the quantification of reaction latencies and choice behavior in binary discrimination tasks [Fudenberg et al., 2020; Myers et al., 2022]. In this model, the process of evidence accumulation is treated as a stochastic random process, reflecting the noisy neural signal transmission and integration that occur during decision-making processes [Gupta et al., 2022]. As the name suggests, the DDM models the observed

evidence as a diffusion process with non-zero drift, commonly referred to as Brownian motion with drift.

Formally, the accumulated evidence at time $t$ is represented as a stochastic process $\{A(t;r)\}_{t\geq 0}$. At the onset of a decision-making process, no evidence has been accumulated, meaning that the initial state is $A(0;r) = 0$. The evolution of the diffusion process is then defined by the rule,

$$A(t;r) = rt + W(t), \tag{2.3}$$

where $W(t)$ denotes a Wiener process, which captures the accumulation of noise over time. The increments of the Wiener process, $W(t+\Delta t) - W(t)$, are normally distributed with mean zero and variance $\Delta t$ for any $\Delta t > 0$, that is, $\sim \mathcal{G}(0, \Delta t)$ [Dobrow, 2016]. This implies that the variability introduced by noise in the evidence variable increases proportionally with time, $t$.

In essence, as time progresses, the accumulated evidence grows linearly on average at a rate $r$, while also exhibiting stochastic variability due to internal noise. A decision is triggered once the accumulated evidence reaches a pre-defined decision threshold, as illustrated in Figure 2.2.

If we are only interested in characterizing the *distribution* of accumulated evidence at a fixed time $t$, the stochastic process can be solved for that time slice, yielding a Gaussian distribution:

$$A(t;r) \sim \mathcal{G}(rt, t). \tag{2.4}$$

However, our primary goal is to characterize the distribution of the *time* it takes for the accumulated evidence to reach a specific criterion—that is, the time required to trigger

a decision. Formally, we seek the distribution of the *first-passage time*, $T(\alpha; r)$, defined as the earliest time at which the evidence reaches a pre-defined threshold $\alpha$:

$$T(\alpha; r) \coloneqq \inf_t \{A(t; r) = \alpha\}. \tag{2.5}$$

Solving for $T(\alpha; r)$ using Equations (2.3) and (2.5) (see Section 2.A for derivation), we find that first-passsage time follows an Inverse Gaussian (IG), also known as the Wald distribution [Folks and Chhikara, 1978]:

$$T(\alpha; r) \sim \mathcal{IG}(\alpha, r), \tag{2.6}$$

with the probability density function:

$$f(t; \alpha, r) = \frac{\alpha}{\sqrt{2\pi t^3}} \exp \frac{-(\alpha - rt)^2}{2t}. \tag{2.7}$$

Intuitively, the random variables describing accumulated evidence, $A(t; r)$ and decision latency, $T(\alpha; r)$, can be understood as inverses of the same stochastic process: the former describes the distribution of evidence at a given time, while the latter describes the distribution of time at a given evidence criterion. Accordingly, these variables follow Gaussian and Inverse Gaussian distributions, respectively. For derivation details of Equation (2.7) refer to Section 2.A.

## 2.3 The Human Visual System

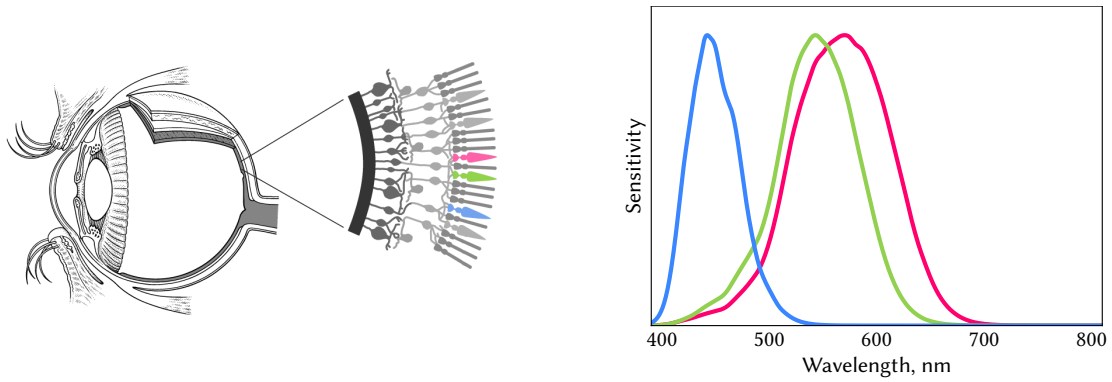
The Human Visual System is a complex system that perceives the visual world using a series of optical components, retinal photoreceptors, and neural structures. In this

section, we review literature from psychology and neuroscience on topics relevant to the work presented in this dissertation. Specifically, we provide an overview of visual signal processing, beginning with the detection of light by photoreceptors in the retina and tracing the transmission of neural signals to the visual cortex. We also discuss prior research on the neural correlates and perceptual mechanisms underlying color and motion perception, as well as studies of eye movement control, which governs how we actively sample information from the visual field.

### 2.3.1 Visual Signal Processing

**Retinal Photoreceptors.** Light entering the human eye through the pupil is detected at the retina by photoreceptor rod and cone cells as illustrated in Figure 2.3a [Tovée, 2008]. Rod cells are exclusively responsible for visual signaling in low luminance environments (below approximately 3 nits), supporting what is known as scotopic and mesopic vision, whereas cone cells are primarily responsible for detecting light in higher luminance conditions beyond the sensitivity range of rod cells [Roufs, 1978]. In addition to their role in photopic (daylight) vision, cone cells enable color perception, as there are distinct families of cone cells, each sensitive to different ranges of light wavelengths, as shown in Figure 2.3b [Williamson and Cummins, 1983]. The colorimetric implications of these cone cell families will be discussed in more detail in Section 2.3.2.

**Cortical Signal Processing.** The light signals detected by the photoreceptors are converted into neural electrical signals and transmitted through the optical nerve to the back of the brain, where the primary visual cortex is located, as illustrated in Figure 2.4 [Tovée, 2008]. Upon reaching the visual cortex, these neural signals propagate forward through multiple streams of neural pathways and are processed to extract various




(a) retinal cone cells (b) cone cell sensitivities

Figure 2.3: *Retinal Photoreceptors.* (a) The organizations of cone photoreceptors (among other structures) in the human retina are shown. Illustration credits: J. Hirshfeld (`https://www.sciencenews.org/article/how-rewire-eye`). (b) The spectral sensitivities of the three cone cell types are visualized in red/green/blue for $L/M/S$ respectively.

statistics features of the visual input [Goodale and Milner, 1992; Henderson et al., 2023]. For instance, neural correlates selectively tuned to spatial patterns [Schwartz et al., 2002], color [Kim et al., 2020], motion [Braddick et al., 2001], and depth [Von Der Heydt et al., 2000] have been identified through a combination of psychophysical and physiological studies. Crucially, visual signals are progressively summarized into higher-level statistical representations as they traverse the visual processing hierarchy [Groen et al., 2017]. This process involves a degree of downsampling and abstraction, suggesting that the brain allocates its computational resources toward extracting and representing task-relevant high-level information at the expense of raw signal fidelity [Freeman and Simoncelli, 2011]. One striking consequence of this summarization is the human inability to reliably distinguish between natural images and modified "mongrel" images—visualizations that preserve certain statistical properties of natural scenes but lack coherent structure—due to information loss in the ventral visual stream [Freeman and Simoncelli, 2011]. Ultimately, these higher-level visual representations are integrated with other cognitive processes in brain regions such as the prefrontal

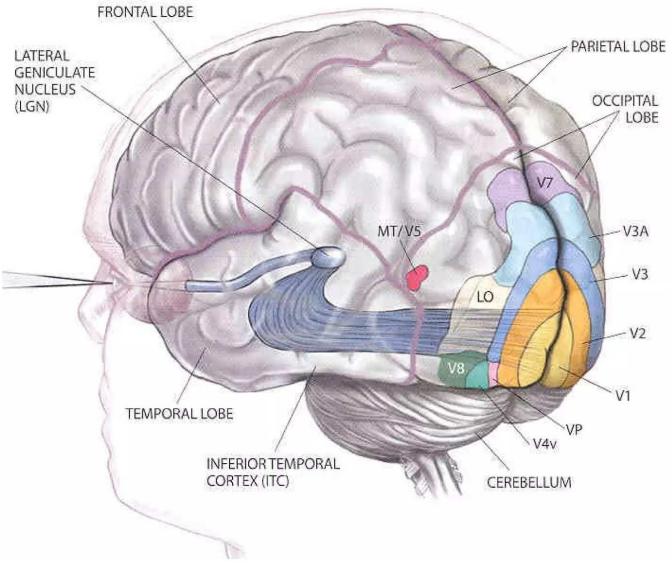



Figure 2.4: *The visual pathway.* Visual signal is transmitted from the retina through the optical nerve, passing through the lateral geniculate nucleus (LGN) before reaching the visual cortex in the occipital lobe. Illustration credits: Brain from top to bottom (`https://thebrain.mcgill.ca/`).

cortex, where perceptual information contributes to decision-making [Skirzewski et al., 2022].

**Foveal vs Peripheral Vision.** As illustrated in Figure 2.5, the distribution of cone cells in the human retina decreases rapidly toward the periphery. This non-uniformity implies that the majority of visual information sampled by the retina is concentrated at the center of our visual field [Roorda and Williams, 1999]. Importantly, this uneven allocation of sensory resources is not limited to the photoreceptor layer; it persists throughout subsequent stages of neural processing [Freeman and Simoncelli, 2011].

For example, measurement of receptive field sizes—the regions of visual space that influence the activity of individual neurons—show a similar non-uniform trend in the visual cortex. Specifically, receptive fields associated with peripheral visual inputs are considerably larger and less densely packed than those representing foveal input

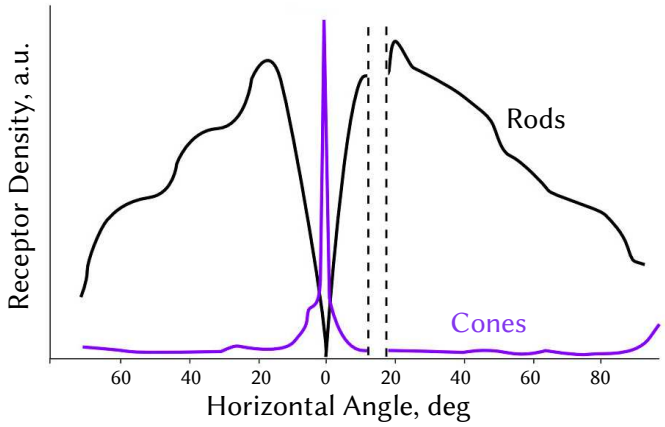



Figure 2.5: *Retinal Photoreceptor Densities.* Retinal cone and rod photoreceptor densities are plotted as a function of eccentricity (i.e., angular distance from the fovea). The discontinuity denoted with dotted lines corresponds to the location of the optic nerve where no photoreceptors are present.

[Tovée, 2008]. This pattern, known as cortical magnification, reflects the fact that a disproportionately large portion of neural computational resources is dedicated to processing visual signals from the central (foveal) region of the visual field [Daniel and Whitteridge, 1961; Tovée, 2008]. Furthermore, neural signals from foveal and peripheral vision are predominantly processed through distinct pathways—namely the magnocellular and parvocellular pathways, respectively. These pathways not only differ in bandwidth and receptive field allocation but also exhibit distinct temporal response characteristics [Hermann et al., 2021; Solomon, 2021].

Consequently, human peripheral vision exhibits several well-documented perceptual limitations and peculiarities. Beyond its significantly lower spatial acuity compared to foveal vision, peripheral vision shows asymmetries between detection and resolution tasks [Thibos et al., 1987a,b], heightened sensitivity to motion [McKee and Nakayama, 1984], including elevated critical flicker-fusion frequencies [Hartmann et al., 1979], reduced capacity for accurate color perception [Cohen et al., 2020; Noorlander et al., 1983].

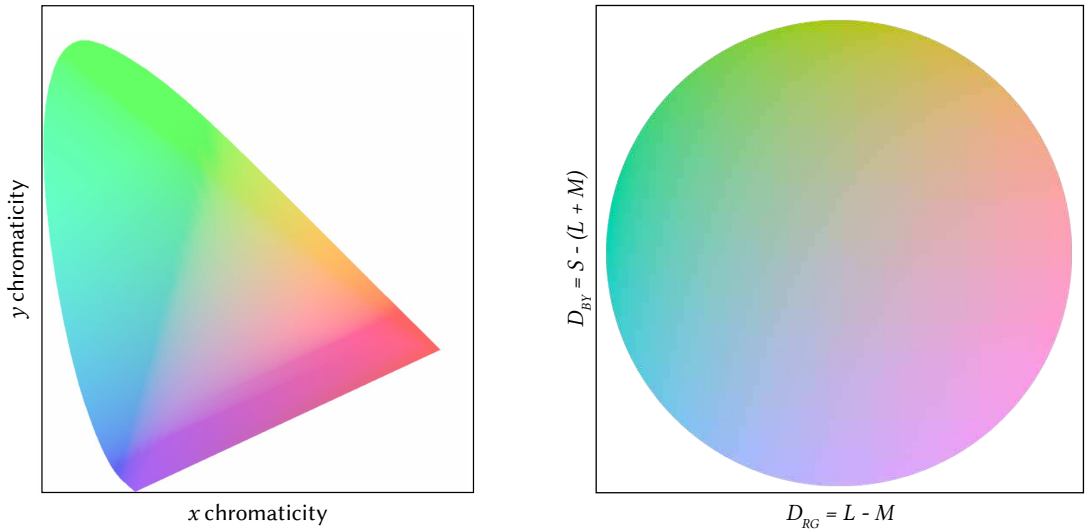
A variety of models have been proposed to characterize different aspects of peripheral visual perception and predict stimulus visibility as a function of visual eccentricities. Many of these models are based on measurements of contrast sensitivity across the visual field [Barten, 1999a; Cajar et al., 2016; Daly, 1992; Kelly, 1979]. While most metrics focus primarily on spatial image characteristics [Rimac-Drıje et al., 2010; Rimac-Drlje et al., 2011; Wang et al., 2001], more recent approaches have incorporated spatio-temporal aspects of perception [Krajancich et al., 2021; Mantiuk et al., 2021].

### 2.3.2 Colorimetry

**Color Perception.** The study of color perception predates much of our physiological understanding of visual signal processing. As a result, early research in color perception was primarily based on psychophysical color-matching experiments, in which participants adjusted the brightness of different colored light sources to match a given reference color. These experiments led to the development of the CIE 1931 *RGB* [Guild, 1931; Wright, 1929] and CIE 1931 $XYZ$ color spaces [Fairman et al., 1997]. The CIE $XYZ$ color space, visualized in Figure 2.6a, has since become the foundation of modern colorimetry, systematically mapping the entire gamut of visible light and providing a hardware-independent framework for quantifying colors [Fairman et al., 1997].

Using the $XYZ$ color space, early psychophysical threshold measurements (as discussed in Section 2.2.1) revealed that human color sensitivity is non-uniform within this space [MacAdam, 1942]. Notably, the three-dimensional nature of the $XYZ$ color space suggested that the human visual system encodes colors using three basis functions—long before physiological recordings confirmed the peak spectral sensitivities of different cone cell families [Bowmaker and Dartnall, 1980; Dartnall et al., 1983].

Due to the technical challenges of directly measuring the full spectral sensitivities of




(a) the $XYZ$ color space (b) the $DKL$ color space

Figure 2.6: *Color spaces.* (a) and (b) depict an equiluminant slice of the $XYZ$ and $DKL$ color spaces respectively.

individual photoreceptor cells, such physiological recordings remain unavailable today [Sincich et al., 2009]. However, psychophysical color-matching experiments involving individuals with congenital color blindness—who lack one of the three cone cell types—have enabled estimations of the spectral sensitivities of the long ($L$), medium ($M$), and short ($S$) wavelength-sensitive cone cells [Stockman and Sharpe, 2000] (visualized in Figure 2.3b).

Beyond the observation that spectral information is encoded through three types of cone receptors, further color-matching experiments have revealed the existence of color opponency mechanisms. Color opponency suggests that certain colors are perceptually opposite to one another—for example, increasing the intensity of red light (or decreasing the intensity of green light) causes an opposing shift in the observed hue [Jameson and Hurvich, 1955]. A similar opponency relationship was identified for blue and yellow hues [Jameson and Hurvich, 1955]. These findings led to the development of the CIE

*LAB* color space, which models perceptual color differences more uniformly than the CIE *XYZ* space [Schiller and Logothetis, 1990].

Just as psychophysical experiments predicted the existence of *L*, *M*, and *S* cone cells, the discovery of color opponency foreshadowed a corresponding physiological mechanism known as cone opponency. Neural recordings have shown that cone responses are compared in a feed-forward manner within the Lateral Geniculate Nucleus (LGN) (see Figure 2.4) before reaching the visual cortex, with opponent channels forming early in the visual pathway [De Valois et al., 1966; Krauskopf and Karl, 1992]. These findings led to the development of the *DKL* color space named after its originators [Derrington et al., 1984a], which provides a physiologically relevant, perceptually uniform representation of color (cf. Figures 2.6a and 2.6b).

In *DKL* color space, the two primary opponent mechanisms correspond to:

1. The difference between *L* and *M* cone activations ($L - M$ channel), which encodes red-green opponency.

2. The difference between combined $L + M$ and *S* cone activations ($(L + M) - S$ channel), which encodes blue-yellow opponency.

Thus, the DKL color space is defined as a linear transformation of the LMS color space. Assuming a D65 gray background the transformation between the color spaces can be expressed as

$$\begin{bmatrix} D_{Ach} \\ D_{RG} \\ D_{BY} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & -2.3112 & 0 \\ -1 & -1 & 50.9875 \end{bmatrix} \begin{bmatrix} L \\ M \\ S \end{bmatrix}, \tag{2.8}$$

where the *DKL* color space vectors representing the achromatic, red-green, and blue-

yellow channels respectively. With its strong physiological and psychophysical foundations, the *DKL* color space is widely used in studies of color discrimination [Ashraf et al., 2024; Conway et al., 2018; Hansen et al., 2009].

**Eccentricity Effects.** As discussed in Section 2.3.1, human color perception exhibits eccentricity effects, meaning that color sensitivity decreases with increasing retinal eccentricity, much like visual acuity [Cohen et al., 2020; Hansen et al., 2008, 2009]. For example, given a reference color, discrimination thresholds in *DKL* color space form ellipses, and the size of these sub-threshold regions expands significantly as retinal eccentricity increases. At 50° eccentricity, the ellipse radii are approximately 4.5 times larger than at 5° eccentricity, indicating a substantial decline in color discrimination ability in peripheral vision [Hansen et al., 2009].

**Cognitive Effects.** Beyond low-level perceptual limitations, color sensitivity is also influenced by cognitive factors and task demands. For instance, during fixation shifts (saccades), color sensitivity decreases uniformly and significantly [Braun et al., 2017]. Furthermore, prior studies have found that color discrimination is notably less sensitive in comparison to color detection [Vingrys and Mahon, 1998]. More recently, Cohen et al. [2020] demonstrated that color sensitivity is further diminished during active and natural viewing tasks—to the extent that peripheral color desaturation becomes imperceptible to observers. These findings underscore the complexity of color perception, a field that continues to be actively explored in contemporary vision science.

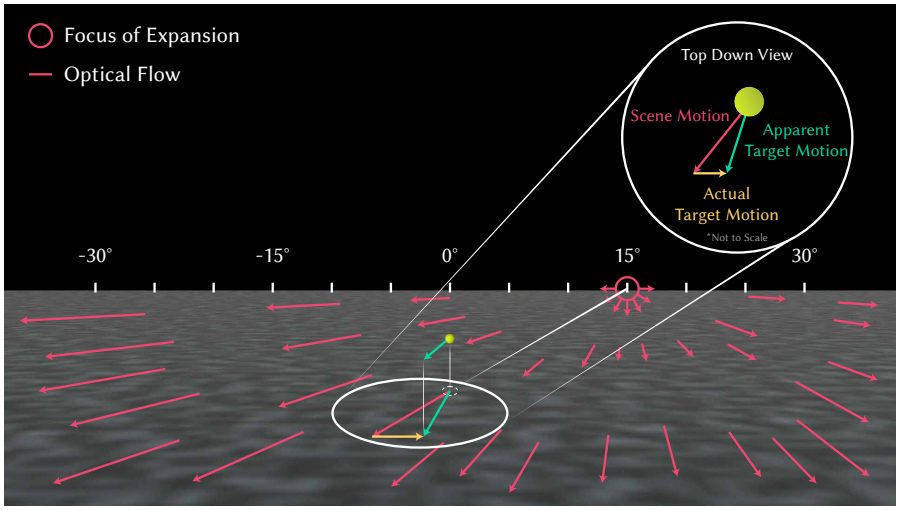



Figure 2.7: *Optical flow processing.* Optical flow of a moving scene is a vector field representing the motion of local patterns across the visual field, indicated by red lines. The focus of expansion (FOE) is the divergence point of this vector field. Human observers compensate for global motion effects when estimating the scene-relative motion of targets.

### 2.3.3 Motion Processing

As perceptual judgments become more complex—such as in motion perception—our current understanding of neurophysiology and psychology encounters significant limitations. In this dissertation, we investigate how humans process local and global motion cues to form a cohesive perception of their surroundings. This review focuses specifically on literature relevant to this topic.

**Optical Flow Processing.** As illustrated in Figure 2.7, human perception of object and environmental motion relies on the optical flow it generates within the visual field [Jain, 1983; Warren Jr and Hannon, 1988]. For local optical flow patterns, observers can visually track moving targets through pursuit eye movements, which minimize retinal slip. At the global level, optical flow resulting from self-motion is crucial for scene understanding and locomotion, serving as the primary visual cue for interpreting complex motion patterns, such as those seen in video displays, and for perceiving 3D

spatial layouts [Talukder and Matthies, 2004]. Accordingly, optical flow is widely used in research to characterize the spatiotemporal motion of dynamic visual stimuli [Huang et al., 1995; Neumann, 1984].

Optical flow patterns generated by rigid translational motion exhibit a stationary on-screen point, known as the focus of expansion (FOE), from which flow vectors radiate outward [Jain, 1983]. The location of the FOE serves as a critical visual cue that humans use to infer the direction of motion for both objects and scenes [Jain, 1984; Warren Jr and Hannon, 1988]. Consequently, prior research has explored how the dynamics of the FOE, which encode observer-relative scene motion in 3D space, influence human motion perception [Jain, 1984; Lappe et al., 1999; Warren Jr and Hannon, 1988].

**Motion Estimation of 3D Visual Targets.** Understanding how humans process motion cues to differentiate between self- and object motion is crucial for assessing the effects of artificially induced motion sensations. This is especially relevant for vection—the sensation of self-motion in a stationary setting [Howard and Howard, 1994; Hu et al., 2019]—as well as for motion-rich display environments, such as racing video games or films featuring dynamic camera movements.

In most real-world scenarios, self-motion perception relies on a combination of visual and non-visual cues to accurately estimate 3D movement vectors within the environment [Xie et al., 2020b; Xing and Saunders, 2022]. Notably, when certain cues—such as vestibular signals [DeAngelis and Angelaki, 2012] and visual depth information from stereopsis [Didyk et al., 2011] and accommodation [Murray, 1994]—are experimentally removed, motion perception becomes increasingly biased warping perceptions of both self-motion [Dokka et al., 2019; Layton and Fajen, 2016; Li et al., 2018; Xie et al., 2020b; Xing and Saunders, 2022] as well as object motion [Xing and Saunders, 2022]. That

is, perceptual estimates of the *Actual Target Motion* in Figure 2.7 become inaccurate. Investigations into neural correlates of self- and object motion dissociation suggest that vestibular signals contribute to fine-tuning these perceptions [Sasaki et al., 2017]. Further studies on the effects of visual complexity and the semantic structure of environments indicate that the congruency of optical flow from self- and object motion affects the accuracy of object motion perception [Lutwak et al., 2022].

### 2.3.4 Eye Motor Control

Due to the foveated nature of human vision, as discussed in Section 2.3.1, the eyes are in constant motion, sampling the visual field to accurately analyze the surrounding environment. As a result, the human visual system is highly dynamic and capable of executing various types of eye movements, including smooth pursuit, vestibulo-ocular, saccadic, and vergence movements [Leigh and Zee, 2015].

In this work, we investigate the characteristics of these eye movements and employ eye-tracking technology to develop computational models that capture their dynamics. This section reviews the relevant literature on the different types of eye movements examined in this study.

**Saccades.** Human eyes change visual fixation three to four times every second [Fabius et al., 2019] via rapid exploratory movements called saccades Leigh and Zee [2015]. Saccadic eye movements allow for frequent shifts of attention to better understand one's surroundings and to localize objects of interest, e.g., potential dangers [Purves et al., 2008]. These movements are ballistic and follow a predictable trajectory [Bahill et al., 1975a; Kowler, 2011], with amplitude, velocity, and duration exhibiting nonlinear relationships. Short saccades, in particular, display an asymmetric, bell-shaped velocity

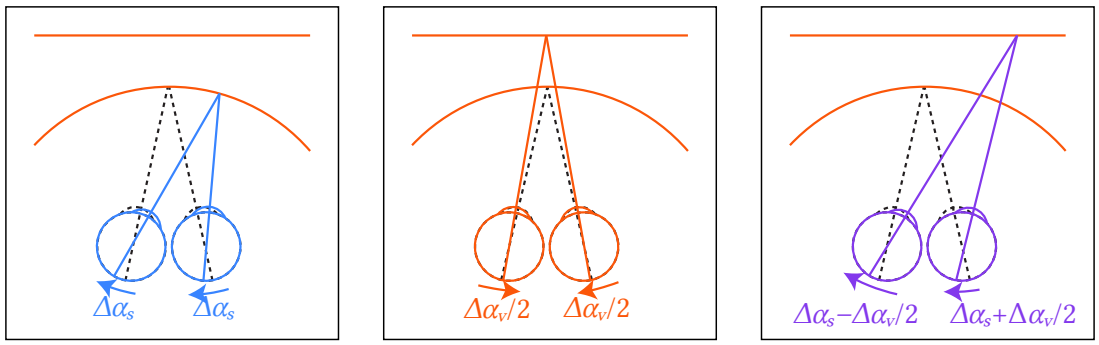
profile [Bahill et al., 1975a], a characteristic that enables the modeling of saccadic profiles even with partial observations.

During a saccade, and for a brief period afterward, the visual system undergoes a temporary perceptual blindness known as *saccadic suppression* [Burr et al., 1994; Diamond et al., 2000; Ibbotson and Cloherty, 2009; Matin, 1975]. This phenomenon naturally helps gaze-contingent graphics tolerate higher eye-tracking latencies [Albert et al., 2017] and has also been utilized in virtual reality redirected walking [Sun et al., 2018].

Saccadic eye movements are often inaccurate, frequently undershooting their target [Becker and Fuchs, 1969; Deubel et al., 1982]. The magnitude of this error depends on factors such as the target location uncertainty, sensory noise [van Beers, 2007], and adaptation processes [Cotti et al., 2009]. Researchers have proposed modeling this uncertainty based on the visual characteristics of the target [Carpenter, 2004; Lisi et al., 2019].

**Smooth Pursuit.** While the primary function of saccades is to correct positional errors in gaze direction and move visual targets of interest into the fovea, pursuit eye movements are used to maintain the fixation on those visual targets [Leigh and Zee, 2015]. The pursuit system is driven by velocity differences between the gaze and the target, relying on the retinal slip signal of the target's image [Blohm and Lefèvre, 2010; Lisberger et al., 1987]. As a result, pursuit eye movements operate as a feedback system that synchronizes the eye and the target motion, ensuring that a foveated moving target remains within the fovea throughout the tracking process [Robinson, 1965].

However, pursuit eye movements do not initiate immediately [Lisberger et al., 1987]. Thus, when a foveated visual target moves suddenly, a catch-up saccade is triggered




(a) saccadic movement (b) vergence movement (c) combined movement

Figure 2.8: *Illustration of various eye movements.* (a) In saccadic movements, both eyes rotate by the same amount in the same direction. (b) In vergence movements, the eyes move symmetrically in opposite directions—away from each other or converging toward each other. (c) In combined movements, each eye moves by a different amount. The curvature of the iso-vergence circle is exaggerated and is not to scale.

to correct the accumulating positional error until the pursuit system is fully engaged [Missal and Heinen, 2017]. The maximum speed for smooth pursuit eye movements in most humans does not exceed 30 deg/s, meaning that stable tracking of faster-moving targets necessitates additional catch-up saccades. The predicted positional error has been shown to be a reliable indicator of whether such catch-up saccades will be triggered [Nachmani et al., 2020]. Additionally, target visibility has been found to influence the performance characteristics of both pursuit and catch-up saccade movements [Spering et al., 2005].

**Vergence.** Vergence movements are best understood in contrast to saccadic movements, which were described earlier. During saccades (Figure 2.8a), both eyes move conjugately, shifting gaze along a circle of iso-vergence (or the geometric horopter), which is determined by the centers of the two eyes and the fixation point (Figure 2.8) [Gibaldi and Banks, 2019]. In contrast, pure vergence movements (Figure 2.8b) are slower and disconjugate, adjusting gaze to a new depth and thereby defining a new

geometric horopter [Gibaldi and Banks, 2019; Yang et al., 2002]. In essence, while saccades enable rapid side-to-side eye movements within the same depth plane, vergence movements allow shifts between different depths.

In stereo displays that lack accommodative cues, the displacement of images presented to each eye provides a critical depth cue, driving vergence eye movements. A common issue in VR/AR settings arises from the conflict between variable vergence cues from stereo displacement and the static accommodation cue corresponding to the display depth. This mismatch leads to discomfort known as the vergence-accommodation conflict [Julesz, 1971]. The duration of pure vergence movements depends on travel distance, direction, and initial depth [Templin et al., 2014]. Measuring vergence movements is more challenging than measuring saccades due to their smaller amplitude [Yang et al., 2002; Yang and Kapoula, 2004], inconsistent performance [Welchman et al., 2008], complex neural coding [Cullen and Van Horn, 2011; King, 2011; Semmlow et al., 2019], and heightened sensitivity to external factors such as pupil dilation [Feil et al., 2017; Jaschinski, 2016; Nyström et al., 2016].

In natural 3D environments, saccadic and vergence movements are more commonly combined (Figure 2.8c) than executed in isolation, reflecting the 3D distribution of visual targets [Kothari et al., 2020; Lang et al., 2014]. Prior research has shown that the addition of saccades accelerates combined eye movements compared to pure vergence alone [Collewijn et al., 1995; Coubard, 2013; Erkelens et al., 1989; Pallus et al., 2018; Yang and Kapoula, 2004]. Competing theories seek to explain the neurological pathways governing vergence and combined movements, but no single theory has achieved consensus [Mays, 1984; Quinet et al., 2020; Zee et al., 1992]. This lack of consensus contrasts with the well-established theories for saccadic movements [Bahill et al., 1975b].

## 2.A Deriving Equation (2.7)

For a Brownian motion process as described by Equation (2.3), the joint probability distribution of an evidence value $a$ observed at time $t$ is described by the Fokker-Plank equation:

$$\frac{\partial f}{\partial t} + r\frac{\partial f}{\partial a} = \frac{1}{2}\frac{\partial^2 f}{\partial^2 a}, \tag{2.9}$$

with boundary conditions

$$\begin{cases} f(0, a) & = \delta(a) \\ f(t, \alpha) & = 0 \end{cases} \tag{2.10}$$

where $p$ is the probability density function of particles behaving according to Equation (2.3), and $\delta$ is the Dirac delta function. The solution to the boundary value problem described by Equation (2.9), with boundary conditions of Equation (2.10), is

$$f(t, a) = \frac{1}{\sqrt{2\pi t}}\left(\exp\left[-\frac{(a - rt)^2}{2t}\right] - \exp\left[2r\alpha - \frac{(a - 2\alpha - rt)^2}{2t}\right]\right). \tag{2.11}$$

This probability density function describes the joint probability of observing any given pair of time $t$ and evidence $a$. Using this density function, we first compute the probability of the evidence being below the criterion, $\alpha$. For the distribution of first passage time, $T$, this probability is equivalent to the survival function. I.e.,

$$S(t) = P(T > t) = \int_{-\infty}^{\alpha} f(t, a)da. \tag{2.12}$$

Plugging in Equation (2.11) into Equation (2.12) we get,

$$S(t) = \Phi\left(\frac{\alpha - rt}{\sqrt{t}}\right) - \exp(2\nu\alpha)\Phi\left(\frac{-\alpha - rt}{\sqrt{t}}\right). \tag{2.13}$$

Finally, we are able to derive the probability density function of $T$ via the relation between the PDF function and the survival function:

$$\begin{aligned} h(t) &= -\frac{dS}{dt} \\ &= \frac{\alpha}{\sqrt{2\pi t^3}} \exp \frac{-(\alpha - rt)^2}{2t}. \end{aligned} \tag{2.14}$$

# Chapter 3

# Color Pathway Limitations and Display Design

In this chapter, we explore the limitations of color vision in human visual periphery (see Section 2.3.2 for relevant background), and how these limitations enable us to design more power efficient head-mounted display systems. Such power optimization opportunities are timely, as AR/VR devices are increasingly becoming untethered for portability, outdoor usage, and unrestricted locomotion to enable ultimate immersion. At the same time, as we've discussed in Section 2.1.1, the display specifications are far from reaching the full requirements for highest fidelity that humans can perceive. Thus, the demands for higher resolution, framerate, and dynamic range are steadily increasing, which is directly at odds with the limited energy capacity of untethered AR/VR devices.

For example, when fully charged, both the Oculus Quest 2 and Hololens 2 can actively run only for 2-3 hours[1]. Since the total energy capacity increases only marginally

[1] https://docs.microsoft.com/en-us/hololens/hololens2-hardware

because "there is no Moore's law for batteries" [Schlachter, 2013], power consumption has become a primary concern in the design process of AR/VR devices [Debattista et al., 2018; Wang et al., 2016; Zhang et al., 2021b]. In our measurement of HTC Vive Pro Eye and Oculus Quest 2, the display consumes as much as half of the total power consumption by comparing the power when the display is on vs. off. The results are consistent with data reported in other measurement studies [Leng et al., 2019; Yan et al., 2018]. Display power will only become more important in the cloud rendering paradigm, where the computation is offloaded to the cloud, heightening the contribution of display to the total device power.

Conventional display power optimizations are geared toward smartphones, which, when directly applied to VR devices, lead to significant visual quality degradation. This is because smartphone display optimizations are fundamentally gaze-*agnostic*, rightly so because smartphone displays have very narrow field-of-view. These optimizations either modulate pixels *uniformly* across the display [Shye et al., 2009; Yan et al., 2018] or are purely based on the content (e.g., UI elements) [Dong et al., 2009; Dong and Zhong, 2011; Ranganathan et al., 2006]. Classic gaze-contingent optimizations in AR/VR such as foveated rendering, while reducing the rendering load [Krajancich et al., 2021; Patney et al., 2016], do not (directly) reduce the display power.

We present a gaze-contingent rendering approach that reduces the power consumption of untethered VR displays by as much as 24% while preserving visual quality during active viewing. We achieve this by only modulating the chromaticity of the display output without changing luminance.

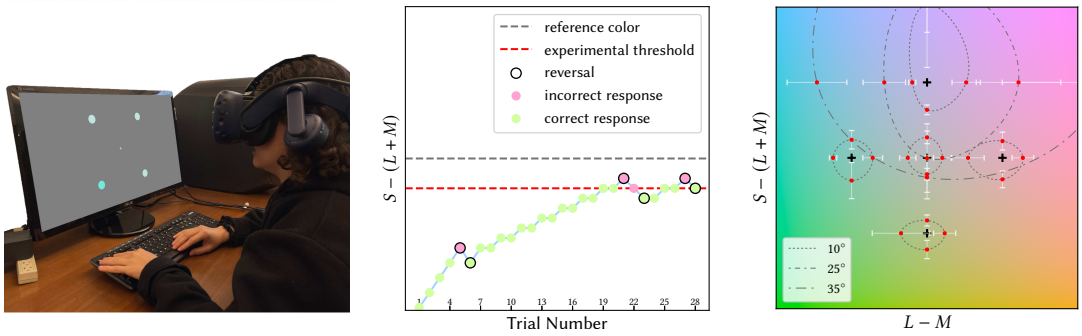
This method is jointly motivated by hardware research that revealed the variation of power consumption of displaying different colors on LEDs [Dong and Zhong, 2011], as well as limitations of human peripheral color vision, as discussed in Section 2.3.2.

That is, given an original frame such as in a 360 video, we seek a computational model that guides a gaze-contingent color shift that (1) requires the minimal power cost, and (2) preserves the perceived fidelity.

To accomplish this, we conducted two studies. First, we quantitatively model how our color sensitivity degrades with higher retinal eccentricities. Second, we physically measure the LED display power consumption as a function of the displayed color. Given the perceptual and the power model, the system performs a constrained optimization that identifies, for each pixel, an alternative color that minimizes the power consumption while maintaining the same perceptual quality. Critically, the optimization problem has a *closed-form* solution because of the judicious design decisions we made in constructing the perceptual and power models. As a result, this perception-perserving color modulation can be implemented as a real-time shader.

We validate this method with both subjective studies on panoramic videos, as well as an objective analysis on large-scale natural image data. We demonstrate the model's effectiveness in display power reduction and perceptual fidelity preservation, relative to an alternative luminance-based "power saver". Our objective analysis concludes that this model shows generalizability to a large variety of natural scenes and save, on average, 14% power.

Complementary to prior work on reducing the rendering power, this work reduces the *display* power—by modulating the display color while preserving perceptual fidelity. We show that significant power saving is readily obtainable by adjusting only color; combining color and luminance modulation would conceivably lead to higher power savings, which we leave to future work (see Section 3.4.3). Source code and data for this chapter's contents are available at `www.github.com/NYU-ICL/vr-power-saver`.




(a) task setup (b) example staircase sequence (c) established thresholds

Figure 3.1: *Study for quantifying human color discrimination.* (a) depicts the psychophysical study setup described in Section 3.1. (b) The sequence of $BY = S - (L + M)$ axis color contrasts in the $DKL$ color space that were displayed during an adaptive staircase are shown. Correct and incorrect responses are color coded, and staircase reversals are outlined (see legend). The established perceptual threshold is visualized via the dotted red line. (c) The overall established color discrimination thresholds (red dots) for the 5 sampled reference colors (black crosses) are displayed in the $DKL$ color space. White bars indicate 75% confidence intervals of the measurements. We only visualize the thresholds of the 25° and 25° eccentric discrimination tasks for one of the reference colors to avoid visual clutter. Cubic splines were used to connect discrimination thresholds to improve plot readability.

## 3.1 Eccentricity Effects on Color Perception

We aim to exploit how human perception of color varies across the visual field, so that we can adjust the appearance of visual stimuli in our peripheral vision in an advantageous way. Hansen et al. [2009] showed that while our ability to discriminate colors significantly deteriorates at high retinal eccentricities, we still maintain some ability to discriminate colors at eccentricities as high as 45°. Drawing inspiration from this work, we designed and performed a psychophysical study on the perceptual *discrimination* thresholds of colors, given various reference colors (5 total) and retinal eccentricities (from 10° to 35°). The experimental data later transforms to a computational model in Section 3.3.1.

The $BY = S - (L + M)$ and $RG = L - M$ axes are axes in the $DKL$ color-space, which

compares the difference between $S$ vs $L + M$ and $L$ vs $M$ cone activations [Derrington et al., 1984a].

### 3.1.1 Experimental Design

**Setup.** We perform our study with the HTC Vive Pro Eye head-mounted display as shown in Figure 3.1a. Participants remained seated during the duration of the study, and interacted with the user study software via the keyboard.

**Participants.** We recruited 5 participants (ages 20-32, 2 female) for a series of four-alternative forced choice (4AFC) adaptive staircase experiments (see Section 2.2.1) to determine the discrimination thresholds. All participants had normal or corrected-to-normal vision and exhibited no color perception deficits as tested by the Ishihara pseudo-isochromatic plates. In this pilot study, we chose 5 participants due to the long duration of our staircase experiment. This is also practiced for similar threshold-determination psychophysical experiments [Krajancich et al., 2021; Sun et al., 2020]. All experiments were approved by an ethics committee and all participants' data was de-identified.

**Stimuli.** As shown in Figure 3.1a, the stimuli were four colored disks (with a diameter of 5 degrees). They were rendered simultaneously on top of a neutral gray background (i.e., $[0.5, 0.5, 0.5]$ in linear *sRGB* space, or 71.5 nits). The azimuth position of the disks remained constant throughout the entire study, located at 45°, 135°, 225°, and 315° (i.e. the four diagonals in the participant's visual field), while the radial position (i.e., the retinal eccentricity) varied across sequences to be either 10°, 25°, or 35°. Three of the disks have the same "reference" color, and the fourth has a "calibration" color which

changes throughout a sequence of trials. The space of colors that the disks can obtain is visualized as a color-space in Figure 3.1c. The luminance of all disks is maintained at the same level as the background's luminance.

**Tasks.** The task was an 1-up-2-down 4AFC adaptive staircase procedure targeting a performance level of 71% correct. The study was conducted in a single session split into 60 staircase sequences (= 5 reference colors $\times$ 3 eccentricities $\times$ 4 color space dimensions, as specified below) of trials.

During each trial participants were instructed to identify which one of the 4 colored disks appeared different. The participant was instructed to fix their gaze on a white crosshair at the center of the screen for the duration that the stimuli were shown. We used eye tracking to ensure participants maintain their gaze at the central crosshair. We automatically rejected a trial if the user's gaze moves beyond 3° eccentricity, randomized the trial order again, and notified them. At the start of each trial of a sequence, we shuffle the four colored disks, and display them for 500 ms (the same stimulus duration used in prior color discrimination literature [Hansen et al., 2009]). Once the stimuli disappear, we prompt the participant to identify and select the disk with the calibration color, using the keyboard. Depending on their answer, the calibration disk's color was made easier or harder to discriminate in the subsequent trial by adjusting the chromaticity of the calibration disk to be closer/farther from the reference color while preserving its luminance. After 6 reversals of this staircase procedure (or a maximum of 50 trials), the sequence terminates, and the next sequence begins. We visualize the progression of an example staircase-procedure in Figure 3.1b.

Across the sequences, we present 5 different reference colors, as visualized with black crosses in Figure 3.1c, each presented at 10°, 25°, and 35° retinal eccentricities. For each

reference color, we adjust the color from four directions along the two equi-luminant cardinal axes in the *DKL* color-space (see Section 2.3.2 for details).

The entire study took approximately 1.5 hours for each participant and they were encouraged to take breaks in between sequences. At the beginning of each user study, the participants completed 1 sequence to familiarize with the procedure and equipment.

### 3.1.2 Results and Discussion.

**Results.** In total, $8,123$ trials were obtained from our participants (5 participants each with 60 sequences consisting of $\approx 21$ trials each on average). We record the color values at each reversal in *DKL* coordinates, and average the last 3 reversals (out of 6 total) to determine the final discrimination threshold for each participant. The average thresholds across all participants are visualized in Figure 3.1c in red, along with the 75% confidence interval error bars. As we approach the reference color from four directions in *DKL* space, we obtain four different thresholds for each color at each eccentricity. The lines connecting the four thresholds do not represent the shape of the overall threshold, and is only served as a visual guide to group each set of thresholds together. To avoid visual clutter, we plot discrimination thresholds at 10° eccentricity for each reference color and 10°, 25°, and 35° eccentricity thresholds for one reference color. Refer to Section 3.4.3 for all the measured threshold values separated by each participant.

**Discussion.** For our work, we only sampled colors on a single equi-luminant plane. In *DKL* space, this corresponds to keeping the $D_{Ach}$ dimension of the color space constant. First, we observed unequal thresholds with different reference colors even if they were displayed at the same eccentricity. That motivates us to develop our computational perceptual model considering the reference color as one of the inputs.

Prior work which utilizes the *DKL* color-space suggests that discriminative thresholds measured with respect to a specific adaption luminance can be extended to arbitrary adaptation luminances due to the linearity of the cone-opponent process [Larimer et al., 1974, 1975]. We use these results in this research and only conducted discriminative threshold measurements at a single adaptation luminance of 71.5 $cd/m^2$ as mentioned above.

In the scope of our work, we did not study how spatial frequencies of stimuli affect discriminative thresholds. Our experimental data provides the thresholds for a stimulus with a dominant frequency equal to 0.2 cpd corresponding to the stimulus size used throughout the experiment.

Unsurprisingly, our data shows a decrease of ability to discriminate chromatic discrepancies as the retinal eccentricity increases. The trend agrees with past experiments [Hansen et al., 2009], and is intuitive given the higher density of retinal receptors in the fovea [Song et al., 2011]. Figure 3.1c shows that the fall-off of discriminative sensitivity is very sharp, and the region of sub-threshold chromaticities at 35° can take up as much as a third of the observable hues. Some participants noted that at high eccentricities, all four disks appeared to be different, even though three of the disks were colored identically. As such, the amount of noisy thresholds at high eccentricities attribute to the larger uncertainty for the overall threshold measurements as shown in Figure 3.1c. Further investigations into this surprising phenomenon is an interesting future work.

We also observe inter-subject variation in the measured thresholds, as shown in Section 3.4.3. While this could be due to a number of reasons (e.g., observer metamerism [Xie et al., 2020a], pre-receptoral filtering [Norren and Vos, 1974], calibration, experimental setup, etc.), further study is required to understand the reason for these differences. Nevertheless, for developing a computational model, we use the most conservative

thresholds across participants, instead of an average fit. This assures generalization to a larger population considering individual variances (see Section 3.3.4).

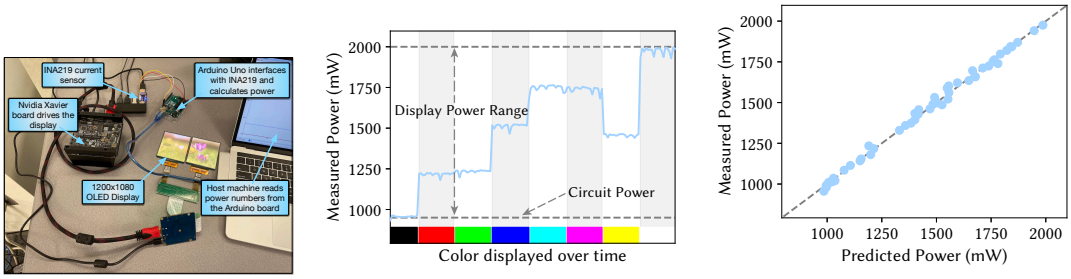
Lastly, it is notably critical that those thresholds only hold for discriminative tasks. Using the observed thresholds, we performed a preliminary validation with a *sequential detection task* and two-alternative-forced choice (2AFC). In this study, the same group of participants was instructed to observe pairs of stimuli and identify whether they appear identical. Some of the trials consist of one non-altered image, with the other containing peripheral color altering within the identified thresholds. We observed that a majority of users can successfully identify the altered condition, suggesting the distinct perceptual thresholds between discrimination and detection tasks. Nevertheless, during active vision tasks where an observer is instructed to freely observe natural visual content, their sensitivity may significantly reduce [Cohen et al., 2020]. We hypothesize that the color sensitivity during active vision is also lower than during discriminative tasks. We investigate and validate the hypothesis in more detail in Section 3.4.1.

## 3.2 Measuring Display Power with Varied Colors

To measure the power consumption characteristics of VR displays and how it varies depending on the images displayed on them, we conduct a hardware study, and later use the collected data to derive a model for predicting the power consumption of a display given the image displayed on it.

### 3.2.1 Experimental Setup.

For our power study, we use the Wisecoco H381DLN01.0 OLED. The display module has two identical displays, each with a resolution of 1080×1200, matching the aspect




(a) equipment setup (b) measurement output (c) model accuracy

Figure 3.2: *Display power measurement and modeling.* (a) The OLED display power measurement hardware rig is visualized (see Section 3.2.1 for details). (b) Voltage and current readings are multiplied to measure power consumption at each timestep and plotted as a function of the the displayed color which was cycled through every 5 seconds. The colors shown are the eight vertices of the *sRGB* color cube. (c) The measured power consumption values for various displayed colors are compared to the linear power model predictions. The power model was regressed by randomly sampling 52 colors in the *sRGB* color space and resulted in a mean relative error of 0.996%. The dashed line indicates the line of perfect measurement and prediction agreement.

ratio of HTC Vive Pro Eyes, which is what we use for perceptual studies.

We do not use the native display modules in Vive Pro Eye HMD and Oculus Quest 2 for power studies, because their displays are physically tightly integrated into the headsets; thus, the display power cannot be easily isolated from the rest of the system. In the case of the Oculus Quest 2, the headset is powered by a battery that is tightly integrated into the headset, which prevents us from using methods used in studying smartphone display power, where the battery is unplugged and replaced with an external power supply that has internal power sensing capabilities [Dash and Hu, 2021; Dong et al., 2009; Halpern et al., 2016].

Figure 3.2a shows the experimental setup to measure display power. We intercept the display power supply with a SwitchDoc PowerCentral board, which has an on-board INA219 module (with a 0.1Ω shunt resistor) to measure the current. The INA219 module is connected to an Arduino board through the I2C interface. We develop a driver that

runs on the Arduino board to get the display current and voltage, from which we can calculate the power.

The driver running on the Arduino board configures the INA219 sensor to output a new power measurement every $\sim$ 68 ms; each power reading is internally averaged over 128 samples, resulting in an effective power sampling rate of $\sim 1,882$ Hz.

### 3.2.2 Measurement and Discussion.

As a preliminary test, we measure the power consumption of the eight vertices of the *sRGB* color cube. For each color, we set all the display pixels to that color, display it for five seconds, and calculate the average power. Figure 3.2b shows the measured power trace. It is clear that the display power consumption is sensitive to the color.

We make two observations from Figure 3.2b. First, even when the display is showing black pixels, i.e., when the LEDs are not emitting light, there is a non-trivial amount of *static* power consumption. The power beyond the static portion is consumed by the LEDs, which we dub the dynamic display power. This static power is consumed by the peripheral circuitry that drives the LEDs, such as the per-pixel transistors and capacitor as well as the addressing logic [Huang et al., 2020]. The contribution of the static power is about 50% in display white and is about 80% when displaying red and green.

The trend of semiconductor technology is that the circuit power is decreasing over time with better fabrication technologies [Bohr, 2007], but the LED power is much harder to reduce because the display must sustain certain luminance levels to meet brightness requirements, which arguably do not change dramatically over time. Our work aims to reduces the (color-sensitive) dynamic power of the display, which will become more important as the static power reduces in the future.

Second, the dynamic power consumption of red and green colors are roughly

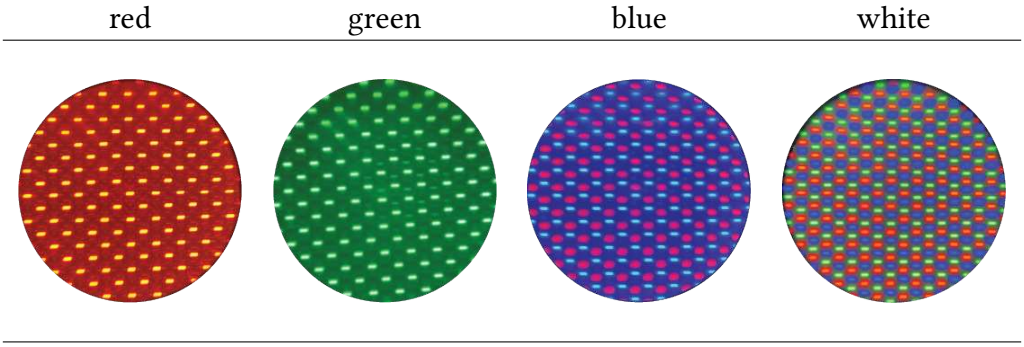
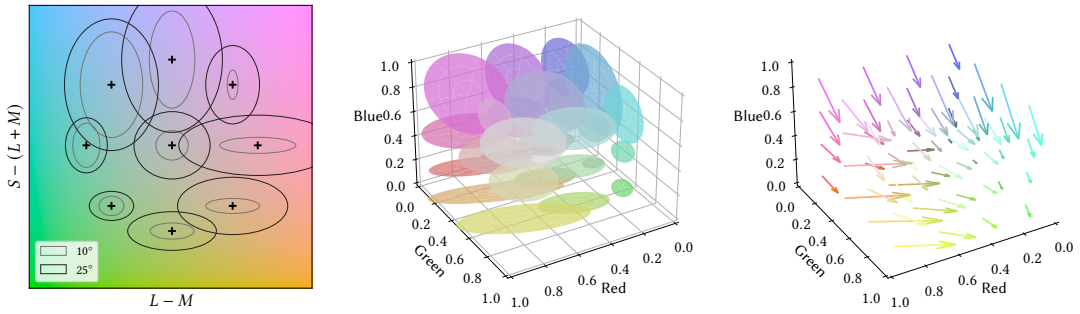



Figure 3.3: *Microscopic photos of the OLED panel.* We image the display under *sRGB* red, green, blue, and white colors using a Carson MicroFlip mircoscope with a magnification of 120x. Note that the display red and green primaries roughly match their corresponding primaries in the *sRGB* color space, but *sRGB* blue requires contributions from both the blue and red sub-pixels from the display panel.

half that of blue. This is because displaying the *sRGB* blue on our display requires contributions from both the blue and red sub-pixels (due to the primaries used by this display) as confirmed by examining the microscopic images of the display (Figure 3.3). As a result, if we expect to see any energy wins, we anticipate that green-, and/or, red-shifting images can decrease the power consumption of the image. We will leverage the measured data to obtain a computational power-vs-color model in Section 3.3.2.

## 3.3 Perceptually Guided Power Optimization

Using the results of our perceptual user study, and hardware power measurements, we develop a display power optimization model under the constraint that the change in the images observed by human subjects is not perceptible. In Section 3.3.1, we first derive a computational model of human color discrimination (Figure 3.4) using the data obtained from Section 3.1. In Section 3.3.2, we build a linear power consumption model regressed from the physical measurement data in Section 3.2. Finally, in Section 3.3.3, we integrate the two models above (as a constrained convex optimization) toward a closed-form




(a) predictions in *DKL* space (b) predictions in *sRGB* space (c) predicted power optimization

Figure 3.4: *Color perception model and power-aware chromaticity optimization.* (a) We illustrate the perceptual model-predicted discrimination thresholds of nine equi-lumianant reference colors in *DKL* color space at two eccentricities. Areas within individual ellipses are predicted to be perceptually indistinguishable from their reference colors (black cross). (b) Thresholds in *DKL* space can be transformed into linear *sRGB* space via linear transformation. We visualize thresholds in *sRGB* space when eccentricity equals 25° by shading sub-threshold color sets with their corresponding reference colors. (c) The model-guided chromaticity shifts at 25∘ eccentricity that minimize power consumption are shown as a vector field. The original and power-optimized colors correspond to the tail and head of the vectors respectively.

display color modulation function. It aims to minimize the display's power consumption while ensuring the modulation within the human discriminative thresholds.

### 3.3.1 Perceptual Model for Color Discrimination

We develop a computational framework for quantifying the discriminative threshold of any given color at different eccentricities. The set of colors which are indistinguishable from some *test* color by human observers are modeled as ellipse shaped regions defined over equi-luminant color-spaces [Hansen et al., 2008, 2009; Krauskopf and Karl, 1992]. Notably, the MacAdam [1942] ellipses are the first to model discriminative thresholds as such. Additionally, Krauskopf and Karl [1992] show that the sizes of these ellipses are best described in the *DKL* color-space [Derrington et al., 1984a].

It is customary to discrimination thresholds using the color contrast; that is, colors

are defined relative to a reference (a.k.a., the *adaptation*) luminance. For a test color, $\mathbf{t}$, and an adaptation color $\mathbf{b}$, expressed in the *DKL* color space, the color contrast of the test color with respect to the adaptation color equals

$$\boldsymbol{\kappa}(\mathbf{t};\mathbf{b}) = \frac{\mathbf{t}-\mathbf{b}}{b_{Ach}}. \tag{3.1}$$

where *Ach* is the achromatic channel of the *DKL* color space (cf. Equation (2.8)).

In this work, we use the *LMS* color space as defined by Smith and Pokorny [1975], which is what the original *DKL* color space is based on [Derrington et al., 1984a]. The particular *LMS* cone fundamentals are so defined that the coordinate $t_{Ach}$ of a color is strictly equal to the luminance of the color, i.e., the $Y$ coordinate in the $XYZ$ space.

**Modeling ellipse level sets.** In our model, we represent the set of all equi-lumiant colors which cannot be discriminated from a test color, $\mathbf{t}$, relative an adaptation color, $\mathbf{b}$, using an ellipse-shaped region centered around the color contrast of the test color. The boundary of this ellipse region corresponds to the discriminative threshold of $\boldsymbol{\kappa}(\mathbf{t},\mathbf{b})$. The set of color coordinates which represent this threshold, $\mathbf{x}$, fulfill the system of equations:

$$\begin{cases} x_{Ach} & = b_{Ach} \\ \mathcal{E}(\mathbf{x};\mathbf{t},\mathbf{b},\boldsymbol{\alpha}) & = 0. \end{cases} \tag{3.2}$$

The first constraint ensures that all the color coordinates on the threshold are equi-luminant to the adaptation color. The second constraint ensures that all $\mathbf{x}$ are on the edge of the ellipse region with major and minor semi-axes equal to $\boldsymbol{\alpha} = (\alpha_{RG}, \alpha_{BY}) \in \mathbb{R}^2$.

Formally, the function $\mathcal{E}(\cdot)$ is defined as

$$\mathcal{E}(\mathbf{x};\mathbf{t},\mathbf{b},\boldsymbol{\alpha}) = \sum_{i=\{RG,BY\}} \left(\frac{\kappa(x_i;b_i)-\kappa(t_i;b_i)}{\alpha_i}\right)^2 - 1, \tag{3.3}$$

**Model Regression.** Equation (3.3) requires the knowledge of the ellipse-size parameters, $\alpha_i$. Prior work has shown that $\alpha_i$ relates to the color contrasts of various test colors, $\boldsymbol{\kappa}(\mathbf{t},\mathbf{b})$, as well as the retinal eccentricity, $e \in \mathbb{R}^+$, at which a colored stimulus is displayed [Hansen et al., 2009; Krauskopf and Karl, 1992]. We leverage our user study data from Section 3.1 to learn the relationship

$$\Phi : \ (\boldsymbol{\kappa}, e) \mapsto \boldsymbol{\alpha} \tag{3.4}$$

where $\boldsymbol{\kappa} \in \mathbb{R}^2$ are the *RG* and *BY* coordinates of the test color in *DKL* space computed using Equations (2.8) and (3.1). Specifically, we use our data to optimize a shallow neural network, which estimates the discrimination thresholds, using least-squares regression:

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\Phi} \|\Phi(\boldsymbol{\kappa}, e) - \boldsymbol{\alpha}\|_2^2. \tag{3.5}$$

The $R^2$ value of the regression is 0.58 (adjusted $R^2$ value of 0.51), indicated an acceptable regression accuracy. Note that our raw data from Section 3.1 is intentionally pre-processed as described in detail in Section 3.3.4. Briefly, we aim to cover more *conservative* thresholds that are generalizable to broad users instead of an "average fit".

**Neural Network Architecture.** We chose the Radial Basis Function Neural Network (RBFNN) with a sigmoid output layer to ensure local smoothness, as well as a positive,

localized output range. Mathematically, the network is summarized as

$$\Phi(\boldsymbol{\kappa}, e) = \boldsymbol{\eta} \odot S\left(\sum_{j=1}^{N} \boldsymbol{\lambda}_j \rho\left(\left\|\begin{bmatrix}\boldsymbol{\kappa}\\ e\end{bmatrix} - \mathbf{c}_j\right\|_2, \sigma_j\right) + \boldsymbol{\nu}_j\right), \tag{3.6}$$

where $\odot$ is the term-wise multiplication operator. The RBFNN takes the input, and computes the weights of the effect each of the $N$ nodes of the latent representation have on the input. It does so by applying a Gaussian Radial Basis function, $\rho$, centered at $\mathbf{c}_j$ with std of $\sigma_j$, for each node, $j$. The weights of each node is scaled by a scaling constant $\boldsymbol{\lambda}_j$, incremented by the linear bias $\boldsymbol{\nu}_j$, summed up, and passed to the sigmoid function $S$ and mutliplied by a scaling factor $\boldsymbol{\eta}$ to produce the final prediction. The trainable parameters of this network are the centres, $\mathbf{c}_j$, sizes, $\sigma_j$ of the radial bases, as well as the final scaling factors $\boldsymbol{\lambda}_j$, and linear biases $\boldsymbol{\nu}_j$. $\boldsymbol{\eta}$ is a normalization constant and chosen to be the maximum possible value of contrasts within the capability of the display used in our work, and hence does not change. For our work we keep the number of nodes $N = 5$ low to maintain smoothness of the outputs. Please refer to our source code for more details on the model specifications.

**Ellipse re-parameterization.** Since the adaptation color, $\mathbf{b}$, is the same for all variables in Equation (3.3), we simplify the function by re-parameterization as $a_i = \alpha_i b_{Ach}$ for $i \in \{RG, BY\}$:

$$\mathcal{E}(\mathbf{x}; \mathbf{t}, \mathbf{a}) = \sum_{i=\{RG,BY\}} \left(\frac{x_i - t_i}{a_i}\right)^2 - 1. \tag{3.7}$$

While the original formulation in Equation (3.3) relates the ellipse to in terms of color contrast, and are ultimately the variables used to regress the model, as we'll see in Section 3.3.2, it's helpful to reformulate the model with respect to raw color space

intensity values: $\mathbf{x} - \mathbf{t}$, as well as the new parameter, $\mathbf{a}$, are both expressed in terms of *DKL* intensity values. Ultimately, the obtained model is visualized in Figures 3.4a and 3.4b.

### 3.3.2 Power Model for Display Illumination

In this section we derive a computation model that correlates an OLED's power consumption with the pixel color. The display power is modeled as the sum of the LED power, which consists of the powers of its three sub-pixels, and the power of the peripheral circuitry (e.g., the thin-film transistors) [Huang et al., 2020]. It is known that the power of an OLED sub-pixel is roughly proportional to its current, which is proportional to the numerical value of the corresponding channel [Tsujimura, 2017]. Thus, given the *RGB* value of the three sub-pixels, $\mathbf{x}_{disp} \in$ *disp-RGB* (i.e., the pixel value in the display native color space), its total power consumption is

$$\mathcal{P} = \left( \sum_{i=\{1,2,3\}} p_i x_i \right) + p_{circ} = \mathbf{p}_{disp}^T \mathbf{x}_{disp} + p_{circ}, \tag{3.8}$$

where $\mathbf{p}_{disp} \in \mathbb{R}^3$ is the vector of unit powers of each sub-pixel, and $p_{circ} \in \mathbb{R}$ is the static power consumption (consumed by the peripheral circuits) when all the pixels are black, i.e., the LEDs do not emit light and, thus, do not consume power.

In most computer graphics applications, it is impractical to use the display's native color space because it varies depending on the manufacturer specifications, and could be unknown. Color-spaces that are commonly used, such as (linear) *sRGB*, can transform to a display's native color-space via some linear transformation, $M \in \mathbb{R}^{3\times3}$. Without loss of generality, using this transformation, we can rewrite Equation (3.8) in terms of

the (linear) *sRGB* pixels as

$$\begin{aligned}\mathcal{P}(\mathbf{x}_{srgb}) &= \mathbf{p}_{disp}^{T} M_{srgb2disp} \mathbf{x}_{srgb} + p_{circ} \\ &= \mathbf{p}_{srgb}^{T} \mathbf{x}_{srgb} + p_{circ},\end{aligned} \tag{3.9}$$

where $M_{srgb2disp}$ is the transformation matrix from (linear) *sRGB*'s color-space to the display's, and $\mathbf{x}_{srgb} \in$ sRGB denotes the pixel color in linear *sRGB* space. For convenience, we define $\mathbf{p}_{srgb}^{T} = \mathbf{p}_{disp}^{T} M_{srgb2disp}$, which intuitively denotes the power consumption of the three display sub-pixels under unit *sRGB* simuli.

$\mathbf{p}_{srgb}$ depends on the specification of a particular display. In our work, we study an OLED display module from Wisecoco that has two 1080×1200 displays, as described in Section 3.2. Critically, our methodology is not unique to the specific display at study and, thus, can be extended to build power models for any other three-primary display.

**Power model regression.** To build an analytical power model, we must find the parameter, $\mathbf{p}_{srgb}$. We do so by physically measuring the power consumption of 52 randomly sampled colors in the *sRGB* space, including the eight colors that correspond to the eight vertices of the *sRGB* color cube, as described in Section 3.2, and solving an over-determined linear system,

$$\mathcal{P}^{(color)} = \mathbf{p}_{srgb}^{T} \mathbf{x}_{srgb}^{(color)} + p_{circ}, \tag{3.10}$$

where *color* is the 52 sampled colors, via the classic linear least squares method. Figure 3.2c shows the measured power of these sampled colors ($y$-axis) and the regressed model outputs ($x$-axis). The mean relative error of the regression is 0.996%, indicating an accurate model.

### 3.3.3 Optimizing Display Energy Consumption under Perceptual Constraints

Finally, using Equation (3.2) and Equation (3.9) we can minimize the power consumption function of a display, $\mathcal{P}(\mathbf{x})$, while constrained within the perceptual limits set by $\mathcal{E}(\mathbf{x})$. Qualitatively, we notice that the power function is a linear function of the input, $\mathbf{x}$, so the minimizing power will be on the surface of the discriminative threshold ellipse (as opposed to its interior). Notice that in this optimization problem, it is more convenient to use *DKL* intensities instead of color contrasts (cf. Equations (3.3) and (3.7)).

Formally, we define the optimization process as:

$$\begin{aligned} \mathbf{x}^*_{dkl} &= \arg\min_{\mathbf{x}_{dkl}} \mathcal{P}(M_{dkl2srgb}\mathbf{x}_{dkl}) \\ &\text{subject to: } \mathcal{E}(\mathbf{x}_{dkl}; \mathbf{t}_{dkl}, \mathbf{a} = \boldsymbol{\alpha} b_{Ach}) = 0, \end{aligned} \tag{3.11}$$

where the original color of the pixel is $\mathbf{t}$ and, the adaptation color of the display is $\mathbf{b}$. In our work we choose $\mathbf{b}$ to be equal to a color with a chromaticity equal to the CIE D65 Standard Illuminant (i.e., the reference white in the *sRGB* color space) and a luminance equal to the luminance of the test pixel $\mathbf{t}$. While the choice of adaptation color is an interesting question to explore, it is beyond the scope of this work and is left as future work.

Due to the convexity of both the cost and constraint functions, we can apply the method of Lagrange multipliers to find the output color, $\mathbf{x}^*_{srgb}$, which minimizes the

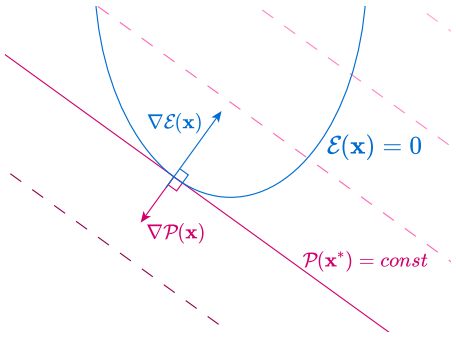



Figure 3.5: *Closed-form derivation of optimal chromaticity.* Level sets of the power function, $\mathcal{P}$, and the boundary of the constraint $\mathcal{E}(\mathbf{x}) = 0$ are visualized in red and blue, respectively. Constrained optimizations of convex functions can be determined in closed form via the method of Lagrange multipliers. Specifically, the optimal chromaticity, $\mathbf{x}^*$, which is guaranteed to be at the boundary of the constraint surface due to convexity can be found by observing that the gradient of both the constraint and optimization functions must be collinear (see red and blue vectors).

total power consumption in closed form:

$$\mathbf{x}^*_{srgb} = M_{dkl2srgb} \begin{bmatrix} t_{Ach} \\ \frac{p_{RG} a^2_{RG}}{\sqrt{p^2_{RG} a^2_{RG} + p^2_{BY} a^2_{BY}}} \\ \frac{p_{BY} a^2_{BY}}{\sqrt{p^2_{RG} a^2_{RG} + p^2_{BY} a^2_{BY}}} \end{bmatrix}. \tag{3.12}$$

Figure 3.5 visually illustrates how this optimal color is found using the derivatives of $\mathcal{E}$, and $\mathcal{P}$. Refer to Section 3.A for the derivation of the above result.

### 3.3.4 Implementation Details

**Perception Study Data Pre-processing.** We take two steps to pre-process the perception study data. Both steps are meant to keep the model's threshold estimation conservative, which is necessary for two reasons. First, there are natural variances across participants (Section 3.1.2) and, thus, a conservative estimation allows our model to generalize to large populations. Second, our model is built to modulate the displayed

colors to preserve the visual fidelity in active viewing, which we hypothesize to have a lower threshold than that in discriminative tasks.

First, we use the smallest thresholds across participants, instead of an average fit. Second, we observed small asymmetries in the collected thresholds, and we confirmed that this is also the case in Krauskopf and Karl [1992]. We made the engineering decision to keep our model's thresholds more conservative; thus each threshold is chosen to be the narrower one from the two thresholds approached from opposing sides along a *DKL* axis. That is, given a threshold approached from the positive *RG* side, $\alpha^{+}_{RG}$, and one from the negative *RG* side, $\alpha^{-}_{RG}$, the discrimination threshold we pick for model regression is:

$$\alpha_{RG} = \min(\alpha^{+}_{RG}, \alpha^{-}_{RG}), \tag{3.13}$$

and similarly for the *BY* axis.

**Eccentricity Extrapolation.** In our perceptual model regression, we restricted the range of valid input eccentricities to be between 10° and 35° because we had only measured discriminative thresholds within this range of eccentricities. We avoided color-shifting content at eccentricities $< 10°$ due to the low power-saving payoffs for foveal and para-foveal regions. Meanwhile, eccentricities $> 35°$ were clamped down to 35° as a conservative estimate.

**Shader.** We implement a post-processing image-space shader in the Unity ShaderLab language to compute per-pixel power-minimizing color. Figure 3.6 outlines the pseudocode of our shader. We tested our shader on the HTC Vive Pro Eye (relevant specs shown below) powered by an NVIDIA RTX3090 GPU, and observed that processing

Table 3.1: *HTC Vive Pro Eye Specifications.*

| Feature | Specification |
|---|---|
| Display Resolution | $1440 \times 1600$ pixels per eye |
| Display Refresh-rate | 90 Hz |
| Peak Luminance | 143 cd/m$^2$ |
| Eye-tracker Accuracy | $0.5^\circ - 1.1^\circ$ |
| Eye-tracker Frequency | 120 Hz. |

each frame takes less than 11 ms, which ensures no loss of frames in the displays.

## 3.4 Power Optimizer Evaluation

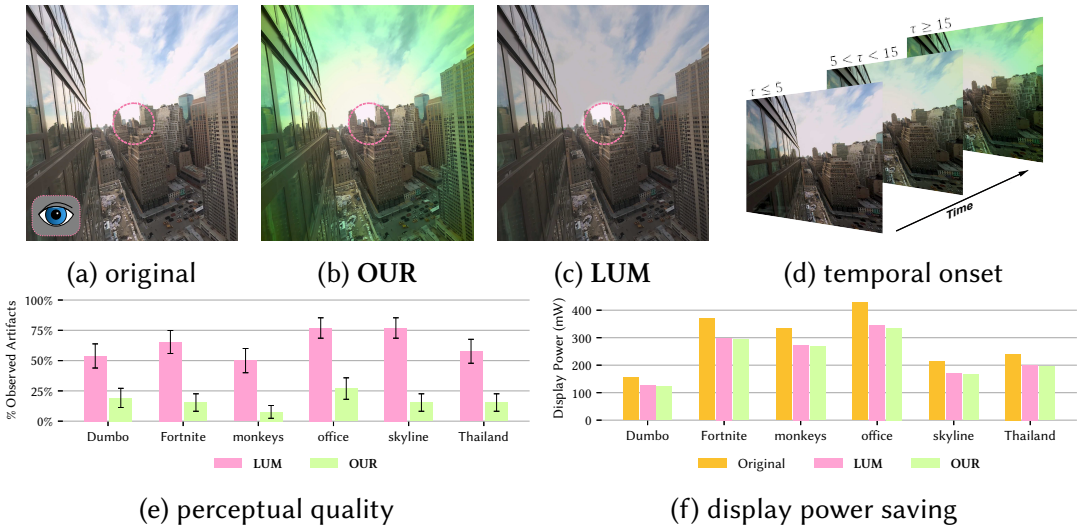
In this section we evaluate the performance and applicability of our model. In Section 3.4.1, we first conduct a psychophysical experiment to assess the perceptual fidelity between our method and an alternative luminance-based power reduction approach. The experiment design follows previous literature [Cohen et al., 2020]. Then, in Section 3.4.2, we measure the model's generic benefits in broad applications by further analyzing the display power saving ratio with a large scale natural image dataset, ImageNet.

### 3.4.1 Psychophysical Study for Perceptual Quality

Motivated by the experiment of Cohen et al. [2020], we conduct a psychophysical user study to measure participant-experienced fidelity deterioration, as well as the corresponding power-saving level during active and real-world viewing. "Active and real-world" is notably a condition where participants may freely rotate their head/eyes and naturally investigate an immersive scene.

1: $\text{POS}_{\text{PIXEL}} = \text{GETPIXELPOS}()$
2: $\text{POS}_{\text{GAZE}} = \text{GETGAZEPOS}()$
3: $\mathbf{t}_{\text{SRGB}} = \text{SAMPLETEXTURE}(\text{MainTex}, \text{POS}_{\text{PIXEL}})$
4: $e = \text{GETECCENTRICITY}(\text{POS}_{\text{GAZE}}, \text{POS}_{\text{PIXEL}})$
5: **if** $e < \text{ECC}_{\text{MIN}}$ **then**
6: **return** $\mathbf{t}_{\text{SRGB}}$
7: **else if** $e > \text{ECC}_{\text{MAX}}$ **then**
8: $e = \text{ECC}_{\text{MAX}}$
9: **end if**
10: $\mathbf{t}_{\text{DKL}} = M_{\text{SRGB2DKL}}\mathbf{t}_{\text{SRGB}}$
11: $\text{LUM} = \mathbf{t}_{\text{DKL}}.\text{x}$
12: ▷ Adaptation color
13: $\mathbf{b}_{\text{DKL}} = M_{\text{SRGB2DKL}}[\text{LUM}, \text{LUM}, \text{LUM}]^T$
14: $\boldsymbol{\kappa} = (\mathbf{t}_{\text{DKL}} - \mathbf{b}_{\text{DKL}})/\mathbf{b}_{\text{DKL}}.x$
15: $\text{INPUT} = \begin{bmatrix} \boldsymbol{\kappa} \\ e \end{bmatrix}$ {Model input}
16: INIT RBF[5], LINEAR[2] {Model output}
17: ▷ Lines 14-22: RBFNN from Equation (3.6)
18: **for** $i$ in $[0, 5)$ **do**
19: $\text{RBF}_i \leftarrow \rho\left(\left\|\text{INPUT} - \mathbf{c}_i\right\|_2, \sigma_i\right)$
20: **end for**
21: ▷ $\odot$: element-wise multiply
22: **for** $i$ in $[0, 2)$ **do**
23: $\text{LINEAR}_i \leftarrow \boldsymbol{\lambda}_i \odot \text{RBF} + \boldsymbol{\nu}_i$
24: **end for**
25: $\boldsymbol{\alpha} \leftarrow \boldsymbol{\eta} \odot \text{SIGMOIDLINEAR}$
26: $\mathbf{a} = \boldsymbol{\alpha} \odot \mathbf{b}_{\text{IDKL}}$
27: ▷ Compute power-optimal color
28: $\mathbf{p} = \text{GETPOWERMODELCOEFFS}()$
29: $\mathbf{x}^*_{\text{DKL}} = \mathbf{t}_{\text{DKL}} + \mathbf{a}^2 \odot \mathbf{p}\frac{1}{\sqrt{\sum_i a_i^2 p_i^2}}$
30: **return** $M_{\text{DKL2SRGB}}\mathbf{x}^*_{\text{DKL}}$

Figure 3.6: *Shader implementation pseudocode.* The shader routine optimizes an input color into the optimized color as described by our method from Section 3.3.




(a) original (b) **OUR** (c) **LUM** (d) temporal onset

(e) perceptual quality (f) display power saving

Figure 3.7: *User study stimuli and results.* (b) and (c) show the results of applying the two gaze-contingent shading conditions **OUR** and **LUM** to an example frame (a) in the video sequence stimuli, with the dashed circles indicating the user's gaze. (d) We gradually increase the strength of the filter over the course of 10 seconds (see Section 3.4.1 for details). (e) shows the percentage of trials where users identified "artifacts" with error bars indicating *Standard Error Mean* (SEM). Across the 6 scenes, **OUR** exhibits significantly lower values than **LUM**, evidencing our method's benefit in preserving perceptual fidelity. (f) Physically measurements of the dynamic display power consumed by the OLED panels for each scene using each rendering method are shown. Power consumption of the peripheral circuitry are excluded. The naive **LUM** and **OUR** both achieve comparable power savings compared to the original.

**Setup and participants.** We recruited 13 participants (ages 21-32, 3 female). None of the participants were aware of the research, the hypothesis, or the number of conditions. All participants have normal or corrected-to-normal vision. We used the same hardware setup as our preliminary user study in Section 3.1. Before each experiment, we ran a five-point eye-tracking calibration for each participant.

**Stimuli and conditions.** The stimuli were 6 panoramic video sequences as shown in Figure 3.8. For broader coverage, the tested scenes contain natural/synthetic, static/-dynamic, bright/dark, and indoor/outdoor content.

We studied the perceptual quality by applying two gaze-contingent and power-saving shading approaches to the scenes: a baseline luminance-modulated shader, **LUM**; and the shader with our chromaticity modulation method, **OUR** (Section 3.3.4). Specifically, in **LUM**, we applied a constant scaling factor to all peripheral (eccentricity > 10°) pixels' colors. That is, **LUM** can be understood as a gaze-contingent version of the "power-saving mode" on mobile devices. The scaling factor was determined in such a way that the power saving (estimated using the power model) of **LUM** is similar to that of **OUR**. An example frame of the original stimulus, **OUR**, and **LUM** are shown in Figure 3.7.

Similar to Cohen et al. [2020], we *temporally* inserted one of the two shaders to the original stimulus during each trial. More formally, let $I_o$ be the original video and $I_p$ be the power optimized version. Then starting at timestamp $\tau_o = 5$s, we linearly interpolate (i.e., $lerp$) between $I_o$ and $I_p$ over a course of 10 seconds. At $\tau_p = 15$s, the transition completes and the power optimized video is played henceforth. That is,

$$I(\tau) = lerp\left(I_o, I_p, \frac{\tau - \tau_o}{\tau_p - \tau_o}\right) \tag{3.14}$$

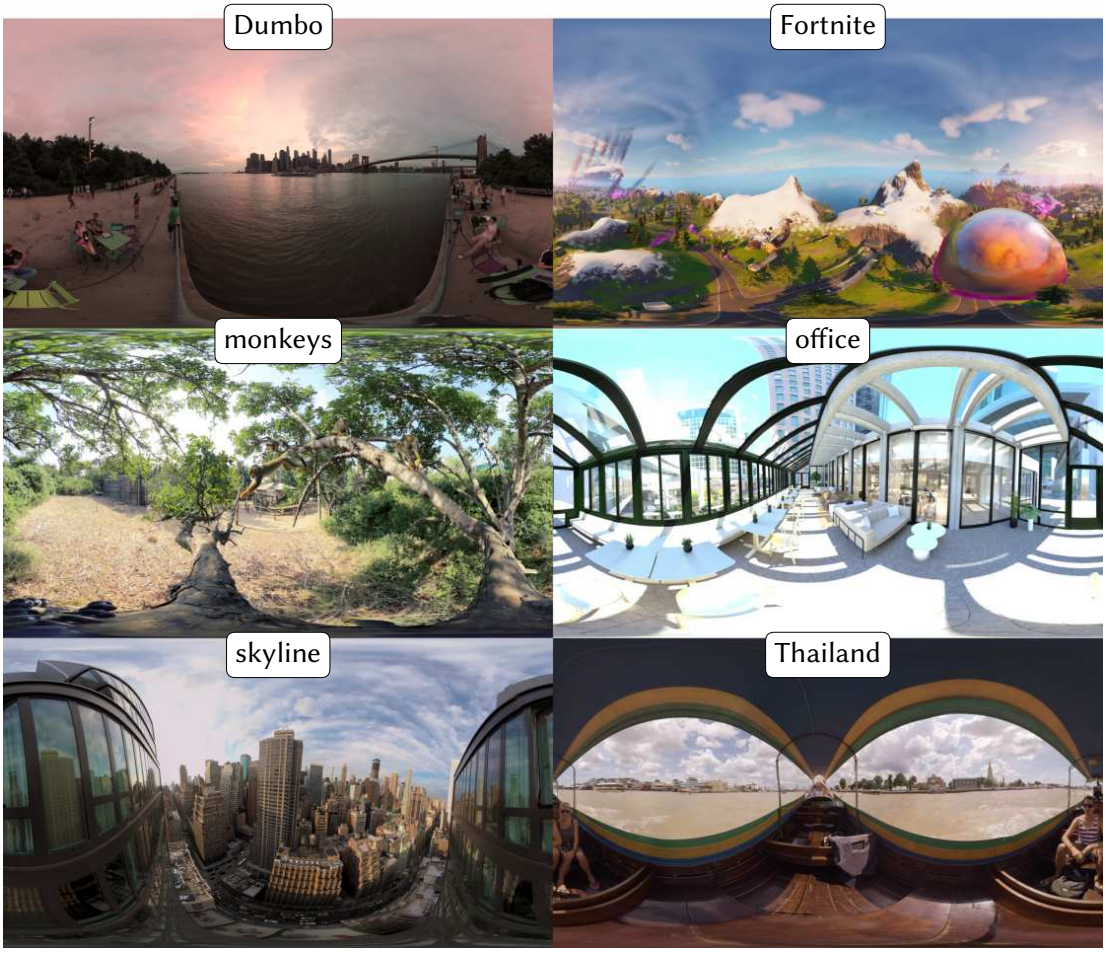



Figure 3.8: *Panoramic Video Scenes.* Representative panoramic frames captured from 360 degree monoscopic video scenes are used in the evaluation user study described in Section 3.4.1. Image credits: *Dumbo, monkeys, skyline* by Humaneyes Technologies, *office* by Rabbitt Design, *Thailand* by VR Gorilla, and *Fortnite* by AmiramiX.

The process is illustrated in Figure 3.7d. Note that the temporal insertion also implicitly compares the original frame with each of the two power-saved conditions.

**Tasks.** Our experiment consisted of 24 trials (6 scenes × 2 condition × 2 repetitions), lasting approximately 15 minutes for each participant. Before the experiment started, we first displayed 2 trial runs to familiarize the participant with the setup. Afterwards, six 20-second video sequences (with representative frames displayed in Figure 3.8) were shown to the participant in a counter-balanced randomized order.

During each trial, the participant was instructed to perform a scene-specific task, such as "count the number of chairs" to ensure they were actively viewing the scene. After each trial, the participant was instructed to answer both the scene-specific task and a two-alternative forced choice (2AFC, similar to Cohen et al. [2020]) question "did you notice any visual artifacts?". Before beginning the experiment, we show the participants static frames from the "skyline" video as visual examples of "artifact" stimuli, including one original frame and the two corresponding conditions, **OUR**, and **LUM**. The example images were displayed on a computer monitor (as opposed to the VR headset); thus, the participants' retinal image was significantly different from the stimuli shown during the study. This is to ensure that the participants are not biased when shown artifacts.

**Metrics and results.** We use the percentage of trials where participants noticed artifacts as the metric of perceptual quality. Lower values indicate better quality, i.e., less noticeable visual modulation. Figure 3.7e plots the user-reported values of each scene and each condition. As visualized in Figure 3.7e, the average percentage of observed artifacts in **LUM** is 63.5 ± 9.4% (STE) and in **OUR** is 16.7 ± 7.3%. The lowest percentage of observed artifacts in scenes with **OUR** applied occurred in the monkeys scene, a scene with large amounts of green, whereas the highest percentage occurred in

the office scene, a very bright and uniformly colored scene relative to the other scenes. A one-way repeated ANOVA analysis showed that the shading condition (**OUR** vs. **LUM**) has a significant effect on the perceptual quality ($F_{(1,24)} = 18.42, p = .00025$).

As plotted in Figure 3.7f, we also measured the display power consumption of each power-saved shading condition for each scene. The average savings between **OUR** and **LUM** are similar ($20.8 \pm 1.2\%$ vs. $18.6 \pm 1.4\%$ (95% confidence)). **OUR** exhibits the highest power saving in the skyline scene, due to higher relative uniform distribution of blue colors.

**Discussion.** The results reveal our method **OUR**'s significant out-performance on preserving perceptual quality over a gaze-contingent luminance-reduction-based approach (**LUM**), even though both conditions achieve a comparably similar power-saving scale. Note that, under the same power, there are infinite ways of constructing **LUM**, including smoothing the edge but darkening the farther periphery. Our implementation of **LUM** is partially inspired by Pöppel and Harvey [1973], which suggests that human luminance change detection thresholds remain relatively constant beyond 10° eccentricity. The design, however, may not be perceptually optimal. Therefore, studying and modeling the luminance-induced effects may not only provide a stronger baseline condition, but improve our model that is currently restricted to colors only.

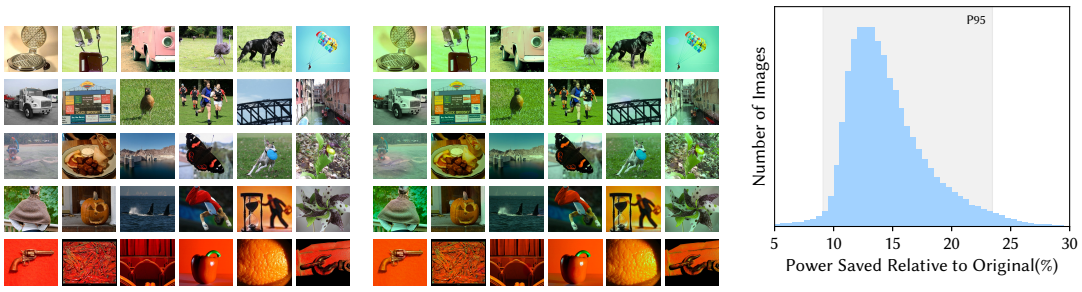
Our perceptual fidelity and power-saving capabilities are also content-based. For example, we notice high power savings in the "office" scene, but the average % of observed artifacts is higher than other scenes. This is hypothetically because the scene has a significantly higher brightness compared to the others. On the other hand, the "monkey" scene has relatively high density of green colors, and thus has the lowest perceived average % of artifacts. The observations motivate us to investigate, in the

future, the chromaticity-luminance joint effect (Section 3.4.3) beyond the first model that guides color-perception-aware VR power optimization.

While detection tasks are commonly leveraged in the foveated rendering literature [Patney et al., 2016; Sun et al., 2017], we opted to validate our gaze contingent filter with active and natural viewing, similar to Cohen et al. [2020]. The design choice is two-fold. First, we attempt to simulate the conditions of real-world VR applications where the users, with various tasks in mind, make head and eye movements to explore the environment. Examples include gaming and video-watching. Second, prior literature suggests unequal color detection and discrimination thresholds. Vingrys and Mahon [1998] discovered that chromatic sensitivity for detection is significantly greater than for discrimination. However, by leveraging our model and shader in this experiment, we verify our hypothesis that our color sensitivity during natural and active vision is, in fact, lower than that of discrimination, and thus enable the method's applicability for broad VR scenarios.

### 3.4.2 Measuring Power-Saving Capability for Broad Content

In our psychophysical study Section 3.4.1, we observe that the possible power savings are dependent on the displayed content. For example, colors that are highly saturated have little room in their equi-luminant plane that is within the bounds of the *sRGB* cube. Therefore, they have less power saving potential as any potential power-saving chromaticity shifts are clipped by the *sRGB* bounds. To study how much power can be saved in practical applications where users may observe arbitrary imagery, we conduct an objective evaluation by applying our method to a large sample of the ImageNet dataset [Russakovsky et al., 2015]. We then measure the distribution of power savings using our power model.




(a) original (b) ours (c) power savings histogram

Figure 3.9: *Power Savings Estimation.* We estimate potential power savings when our model is applied to natural images sampled from ImageNet [Russakovsky et al., 2015]. The amount of power saved is content dependent. Images in (a) and (b) are grouped based on how much power-savings were achievable (top to bottom each row saves better than 99.9%, 55%, 45%, 0.1% of the entire dataset considered; bottom row are the worst 0.1% performers). (c) We show the distribution of potential power savings in this evaluation, and annotate the percent power saved for 95% of images to be within 9.1 – 23.5%.

**Setup.** We simulate how an image would be observed in a VR setting by resizing the image to be displayed at 90° field-of-view, and randomly sample a location within the image and select that as the gaze location. The randomization of the gaze location was applied to prevent bias in the power estimation. Specifically, we observed that many images in ImageNet have a foreground object centered on the image; selecting random gaze locations allows us to include images where the foreground objects will sometimes have the filter applied to them. We repeat this process for 10% (randomly sampled) of the ImageNet dataset, totaling in over $\sim 120k$ images, to collect original and power-optimized image pairs.

**Metrics.** For each image pair, we measure the estimated power consumption using our model from Section 3.3.2, and compute the relative decrease in power consumption by applying our filter with respect to the ground-truth condition.

**Results.** We observe that the mean display power saving recorded across the entire dataset is 13.9%, and guarantee 9.1 − 23.5% savings with *P95* confidence. Please refer to Figure 3.9c for the detailed histogram of the estimated power savings. We visualize a small sample of the images we applied the filter to in Figures 3.9a and 3.9b.

**Discussion.** Sample images from different percentiles of power-saving as shown in Figure 3.9 show that images with the highest power savings are commonly bright and/or blueish scenes, and vice versa. Intuitively, bright scenes provide larger *percentage* changes in LED luminance, and thus unlock larger space for power-saving. Since blue colors on the LED consume the most power, as demonstrated in Section 3.2, images rich in blue/green colors can be optimized most effectively. Meanwhile, images which are already saturated with red colors cannot be optimized for higher power-saving because the space of power-wise "cheaper" colors is narrower.

### 3.4.3 Limitations

**Active vision vs. discrimination vs. detection.** While our evaluation on active/-natural viewing tasks in Section 3.4.1 is representative of real-world VR scenarios [Cohen et al., 2020], our initial perceptual data are collected using a more conservative *discrimination* task. It is also common in the foveated rendering literature to evaluate using *detection* tasks [Patney et al., 2016; Sun et al., 2017]. Our conducted preliminary detection-based experiments, which showed that sensitivity to color changes in detection tasks is significantly greater that that in discrimination tasks, are consistent with prior work [Vingrys and Mahon, 1998].

An exciting future direction is, thus, to investigate an adaptive model that accommodates for color sensitivity under all three tasks (detection, discrimination, active/natural

viewing). That way, our color modulation algorithm can be dynamically configured according to the specific viewing scenario of a VR user.

**Perceptual model.** Our current perceptual model is constructed with respect to per-pixel colors. An interesting future extension is to consider inter-pixel, potentially higher-dimensional, features such as spatial frequency and local contrast. Performing these analyses (e.g., frequency domain analyses as in Tursun et al. [2019]), however, increases the computational overhead. How to best balance the level of details in perceptual analysis and display power saving is an open question we leave to future work.

**Luminance adjustment.** In our work, we model and modulate pixel *chromaticity* to reduce display power consumption while preserving *luminance*. This design choice reduces the dimension of our perceptual model and, thus, yields a convex constrained optimization problem with a closed-form solution. Investigating the luminance-chromaticity joint modulation is an interesting future research direction that would conceivably lead to higher power savings [Vingrys and Mahon, 1998].

Jointly adjusting luminance and chromaticity, nevertheless, comes with a few challenges. First, it would require sampling a new dimension in constructing the perceptual model. Second, prior literature suggests the weak eccentricity-dependent effect [Metha et al., 1994] in detecting and discriminating absolute luminance. Finally, the perceptual level sets, when considering the luminance dimension, might not be convex, which might complicate the optimization, cause false local minima, and reduce the shading speed.

**Color Temperature Adaptation.** Another interesting direction for future research is to leverage chromatic adaptation [Fairchild, 2013] to reduce display power by adjusting the color temperature of the display white point. The advantage of this approach is that it is not gaze-contingent: it can potentially reduce the power of the entire display without requiring eye tracking. Adaptation to display color has been long investigated [Fairchild and Reniff, 1995; Peng et al., 2021], but such studies in VR displays are relatively new and rare [Chinazzo et al., 2021] and lack a comprehensive computational model. Note that the chromatic adaptation benefits are additive: our model can be seen as a sub-space initial attempt (by exploiting spatial color perception) under a *given* adaptation state.

**Display Persistence.** VR displays usually have low persistence to reduce motion blur [Hainich and Bimber, 2016]. A common solution is to hold a frame for only a short period of time during each display refresh[2]. As a consequence, the display power is relatively low to begin with (compared to displaying a frame throughout a refresh cycle). Nevertheless, our work demonstrates significant display power saving opportunities even with reduced displayed times. In addition, reducing the display period leads to low average luminance, which limits the applicability of a luminance-based approach to reduce power — another reason we choose to maintain the luminance.

**Implementation.** There is room for improving the speed of our shader, which currently is bottlenecked by the atomic `for`-loop. Deferred shading techniques may shed light on alleviating the bottleneck. One promising solution is to evaluate the optimization problem (Equation (3.11)) offline (e.g., sampling colors and eccentricies) and save the results as a 3D texture, which the shader simply looks up at rendering time.

[2] `https://developers.google.com/vr/discover/fundamentals#display_persistence`

Due to the tight integration of the display, computation module, and battery in commercial AR/VR devices, our display power measurement has to be done on a 3rd party display module that has the identical aspect ratio of the VR device we use for perceptual studies. Investigating physical means to measure the exact display power as in an AR/VR device would reveal the real-world energy savings concerning the battery equipped with the device. It would also be interesting to see how our perception-conserving color modulation idea can be applied to smartphone displays, which have much narrower field-of-views.

**Follow-up Work.** Since our initial publication of this work, our perceptual model has been leveraged to implement a frame-buffer compression system that alleviates DRAM traffic by up to 67% and outperforms existin frame-buffer compression mechanisms by up to 20%, while preserving the perceptual fidelity of the displayed content. While such a compression scheme is numerically lossy, our perceptual model assists in ensuring that the color artifacts generated caused perceptually lossless results [Ujjainkar et al., 2024].

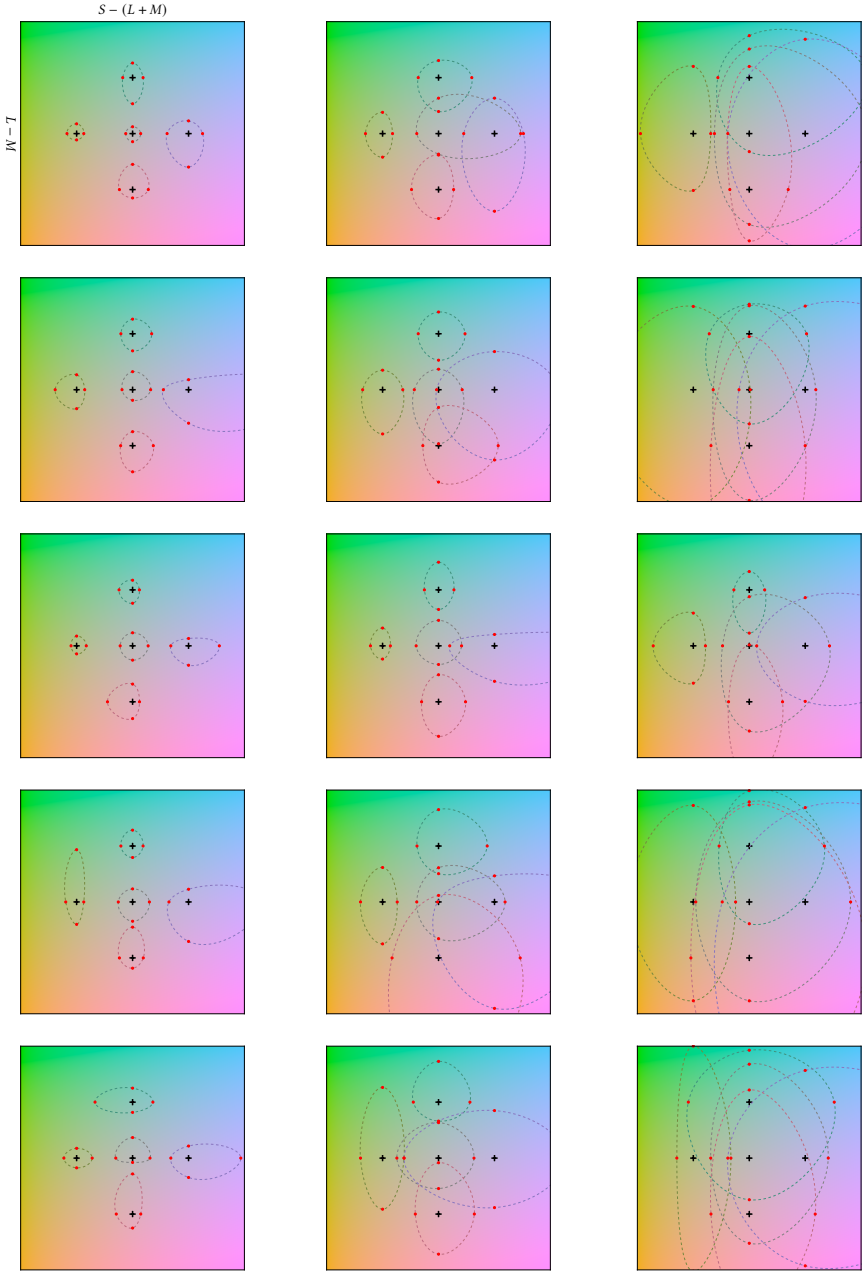
| Subject ID | Eccentricity = 10° | Eccentricity = 25° | Eccentricity = 35° |
|---|---|---|---|
| subject_01 |  | | |
| subject_02 | | | |
| subject_03 | | | |
| subject_04 | | | |
| subject_05 | | | |



Table 3.2: *Pilot Perceptual Study Threshold Data.* See Figure 3.1 for plot description.

## 3.A Optimal Color Modulation Derivation

Given an ellipse constraint function

$$\begin{cases} x_{Ach} = t_{Ach} \\ \mathcal{E}(\mathbf{x}) = \left(\frac{x_{RG}-t_{RG}}{a_{RG}}\right)^2 + \left(\frac{x_{BY}-t_{BY}}{a_{BY}}\right)^2 - 1 = 0, \end{cases} \tag{3.15}$$

and a power cost function

$$\mathcal{P}(\mathbf{x}) = p_{Ach}x_{Ach} + p_{RG}x_{RG} + p_{BY}x_{BY} + p_{circ}, \tag{3.16}$$

we may solve the power optimizing $\mathbf{x}^*$ via the method of Lagrange multipliers.

First, we notice that $x_{Ach}$ should not change. Intuitively, this effectively reduces the dimensionality of the optimization onto the plane $x_{Ach} = t_{Ach}$. Formally, we may rewrite the constraint and power functions in terms of a 2-dimensional variable $\vec{y} = (y_{RG}, y_{BY}) = (x_{RG}, x_{BY})$:

$$\mathcal{E}(\vec{y}) = \left(\frac{y_{RG} - t_{RG}}{a_{RG}}\right)^2 + \left(\frac{y_{BY} - t_{BY}}{a_{BY}}\right)^2 - 1 = 0, \tag{3.17}$$

and

$$\mathcal{P}(\vec{y}) = p_{RG}y_{BY} + p_{BY}y_{BY} + const. \tag{3.18}$$

The minimizing vector $\vec{y}^*$ satisfies the condition that the gradients of $\mathcal{E}$, and $\mathcal{P}$ are

co-linear. So the system of equations we need to solve for $\vec{y}^*$ is

$$\begin{cases} \nabla \mathcal{E}(\vec{y}^*) = \phi \nabla \mathcal{P}(\vec{y}^*) \\ \mathcal{E}(\vec{y}^*) = 0, \end{cases} \tag{3.19}$$

for some scalar constant $\phi$.

Computing the gradients, we get

$$\begin{cases} \frac{2}{a_{RG}} \frac{y^*_{RG} - t_{RG}}{a_{RG}} = \phi p_{RG} \\ \frac{2}{a_{BY}} \frac{y^*_{BY} - t_{BY}}{a_{BY}} = \phi p_{BY} \\ \left( \frac{y^*_{RG} - t_{RG}}{a_{RG}} \right)^2 + \left( \frac{y^*_{BY} - t_{BY}}{a_{BY}} \right)^2 - 1 = 0. \end{cases} \tag{3.20}$$

Finally, we solve for $\vec{y}^*$ using this system of equations to get the optimal color, $\vec{x}^*$:

$$\begin{aligned} x^*_{Ach} &= t_{Ach} \\ x^*_{RG} &= \frac{p_{RG} a^2_{RG}}{\sqrt{p^2_{RG} a^2_{RG} + p^2_{BY} a^2_{BY}}} \\ x^*_{BY} &= \frac{p_{BY} a^2_{BY}}{\sqrt{p^2_{RG} a^2_{RG} + p^2_{BY} a^2_{BY}}}. \end{aligned} \tag{3.21}$$

# Chapter 4

# Motion Processing Error Correction

When driving on the road, we must accurately estimate and respond to the motion of various objects in a dynamic environment, including other vehicles and pedestrians. How users perceive object motion is also a universal metric in computer graphics applications, such as guiding camera trajectories in video playback [Kang and Cho, 2019], controlling game difficulties [Caroux et al., 2013], compressing videos [Furht et al., 2012], and reducing simulator sickness [Hu et al., 2019; Park et al., 2022]. In these real-world scenarios, both the objects and we ourselves may move within dynamic 3D environments. In such situations, extracting scene-relative object motion solely from the mixed and anisotropic optical flow on the screen can lead to misinterpretations due to its ambiguous nature [Dokka et al., 2019]. Therefore, we ask, "How accurately can we perceive moving objects in scenes featuring different motion dynamics?".

In Section 2.3.3 we discussed that prior studies have observed that perceptual errors can occur when estimating object movements during self-movements [Dokka et al., 2019; Xing and Saunders, 2022] and in 3D scenes [Cornilleau-Pérès and Gielen, 1996; Van den Berg and Brenner, 1994a]. However, to provide design guidance in downstream

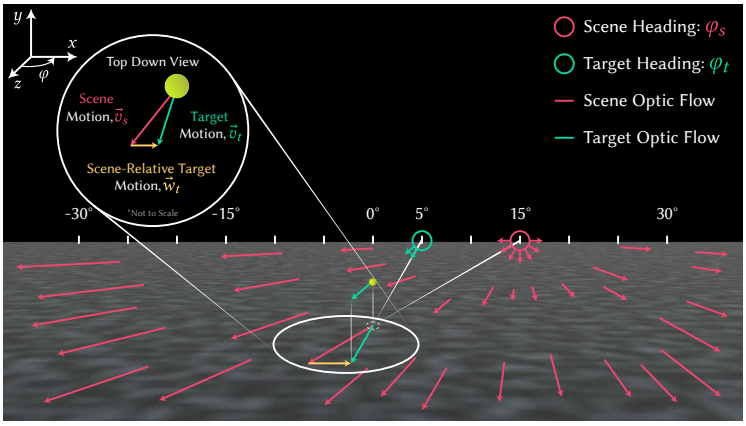
graphics applications, a quantified understanding of the variability of these errors across different scene dynamics is still missing. Filling this knowledge gap poses a remarkable challenge due to the need of sampling a diverse range of conditions, conducting repeated experiments, and involving a wide population to account for variations in individuals' sensory and perceptual variances [Xing and Saunders, 2022].

In this chapter, we measure and analyze the errors in our visual perception of screen-displayed object motion, particularly in relation to concurrent global scene movements which result in dynamic environments. To this aim, we present a series of large-scale psychophysical studies comprising over 10,000 trials, which correlate object motion perception and scene dynamics characterized by scene movements and content depths. We employ and validate a crowdsourcing approach to tackle the unique challenges posed by the need for large sample sizes in both population and trial repetitions.

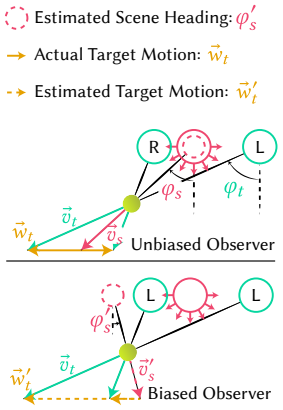
Additionally, we also showcase how the model can guide animation and game design to reduce perceived errors in object motion by viewers. We hope this work will contribute to a new frontier in the computer graphics community, focusing on understanding the visual performance limitations introduced by displays and exploring design strategies to compensate for them. Source code and data for this chapter's contents are available at `www.github.com/NYU-ICL/motion-estimation`.

## 4.1 Studying Object Motion Perception

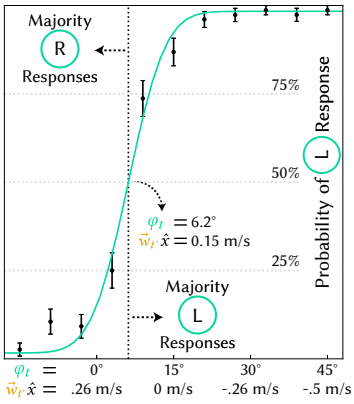
In a dynamic scenario, a target object moves in the scene ($\vec{w}_t$), which simultaneously appears to be moving to the observer who is also in motion ($\vec{v}_s$), as visualized in Figure 4.1a. Figure 4.1b (top) illustrates that an unbiased "perfect" observer can accurately understand $\vec{w}_t$ and $\vec{v}_s$ by analyzing their vector combination, $\vec{v}_t$, as it appears on-screen.




(a) object movement and self movement




(b) unbiased vs biased observer




(c) left/right threshold measurement

Figure 4.1: *Illustration and analysis of biased perception during self-motion.* (a) Accurate reconstruction of the scene-relative target motion $\vec{w}_t$, requires observers to subtract their percept of scene motion $\vec{v}_s$, from the observed on-screen target motion $\vec{v}_t$. The divergence point of optical flow fields due to scene and target motions, a.k.a., FOE, denoted as circles at the horizon. (b) Unbiased "perfect" observers can perfectly estimate the scene heading, $\varphi_s$, to determine the direction of scene-relative target motion. Observer L(eft)/R(ight) responses are annotated inside the FOE circle for each target motion condition. Biased human observers make judgment errors due to mis-estimation of the scene heading, $\varphi'_s \leq \varphi_s$. Biased estimations denoted as dashed arrows. (c) The psychometric curve visualizes the probability of observers L/R responses for various target motion conditions. The curve indicates that when the target moves through the scene at a speed of 0.15 m/s to the right (equivalent to an observed target heading of $\varphi_t = 6.2°$) observers believe the object to not be moving sideways, on average. Data used for curve fitting is shown as a scatter plot (with SEM error bars).

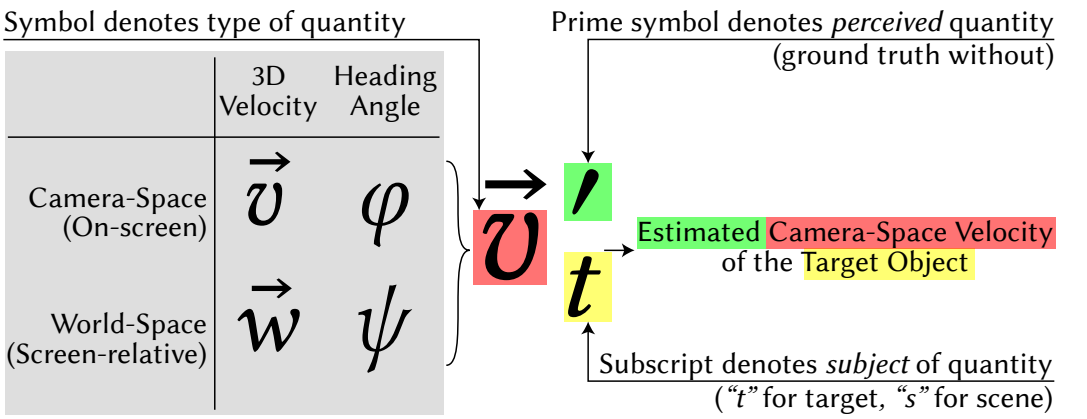



Figure 4.2: *Motion-related variable notation used throughout Sections 4.1 and 4.3 and Figs. 4.1 and 4.8.*

Refer to Figure 4.2 for a reference to all target and scene motion-related symbols used throughout the manuscript. However, this ideal scenario may not reflect reality. As depicted in Figure 4.1b (bottom), we are imperfect in estimating either motion due to the decomposition ambiguity [Xie et al., 2020b; Xing and Saunders, 2022]. First, depending on scene dynamics, our perception of scene and target heading often exhibits a "central bias", meaning an *under*-estimation [Xie et al., 2020b; Xing and Saunders, 2022]. Second, when observers lack visual cues to determine the target distance, the ambiguous optical flow further exacerbates the mis-estimation [Van den Berg and Brenner, 1994a]. For example, in Figure 4.1a, it is ambiguous whether the ball is large and moving at a farther depth or small and moving at a closer depth. Therefore, we study (Section 4.1.1), quantify, and model (Section 4.1.3) the perceptual bias scale of target motions under various scene dynamics and content.

### 4.1.1 Experimental Design

**Participants.** We recruited subjects for the study through the crowdsourcing platform *Prolific*. A strict screening protocol was enforced to mitigate potential confounds

arising from task misinterpretation and attention lapses, ensuring high-quality data (see *Filtering*). As such, we consider the data from $n = 38$ subjects (ages $20 - 56$, 21 male) screened from an initial pool of 78. All study protocols were approved by an institutional review board (IRB), and subjects were compensated at a rate of $\$15/h$. Refer to the supplementary video for animated visualizations of all study procedures.

**Stimuli and procedure.** The study was conducted via a web-based application on a computer screen. A screen calibration procedure ensured that all subjects viewed the stimuli at approximately 50° fov. After calibration, they received a text-based introduction to the stimuli and task.

Subjects initiated each trial by pressing a button. As shown in Figure 4.3a, they were presented a fixation cross at the screen center for .5 s at the beginning of each trial and instructed to maintain their gaze stationary. After the cross disappeared, a 2 s video (recorded at 60 fps) was shown. Initially, a flat ground surface with Perlin noise texture is visible, conveying forward scene motion with variable speed, $v_s$, and heading direction, $\varphi_s$, to an observer at variable height $h_s$. The ground texture was chosen to avoid tuning to specific spatial frequency ranges, and instead incorporate a broad spectrum of frequencies, similar to Xing and Saunders [2022]. After 1 s, a yellow probe (target object) was introduced at a height, $h_t$, positioned 6 m in front of the observer at 5° eccentricity below fixation ($h_s - h_t = .52$ m). The target object then moved either left or right relative to the scene at various speeds, $w_t$, for the rest of the clip (1 s). The object remained visible throughout all trials.

At the end of the video, subjects were prompted to indicate, via button press, whether the probe was moving left or right *relative to the scene.* If they didn't respond after 10 s, the trial expired and prompted a *screening* trial before retrying. No feedback was

provided during trials to prevent learning effects.

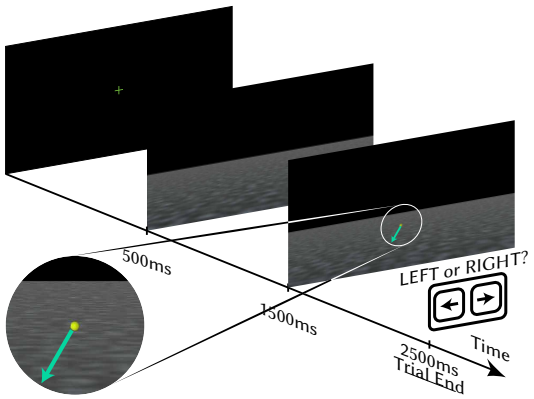
Prior to the study, subjects participated in an interactive *training* session to familiarize themselves with the task and interfaces. The session comprised eight unique trials of the same protocol. During training, subjects were provided with feedback on their performance after each trial and shown a top-down visualization (see Figure 4.3c). Subjects were required to respond correctly to all training trials before being allowed to progress. Training conditions were selected to prevent external bias (see *Conditions*).

**Metrics.** The procedural goal of the study was to determine the threshold heading of the target object, $\mu$, at which subjects perceive the target's scene-relative velocity to be zero: $\vec{w}'_t = 0$ (a.k.a., bias and inaccuracy). During each trial, the subject is presented with targets of different velocities, $\vec{w}_t$, which appear on-screen to be moving along

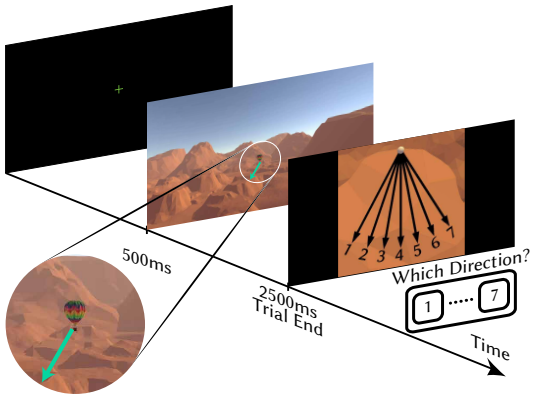
$$\vec{v}_t = \vec{w}_t + \vec{v}_s, \tag{4.1}$$

as illustrated in Figure 4.1a. By aggregating subject responses for different target velocities, $\vec{w}_t$, each corresponding to a different target heading direction, $\varphi_t$ (see Figure 4.1b), we fit a psychometric curve, $p$ (see Equation (2.2)). This allows us to determine the response bias in heading target heading judgments, $\varphi_t = \mu$, at which observers perceive that the target is neither moving left nor right. The experimental data used to fit the response bias, $\mu$, and slope, $\sigma$ consisted of 11 target headings, $\varphi_t$, stimulus levels uniformly sampled between $[-\varphi_s, +3\varphi_s]$ (see Figure 4.1c).

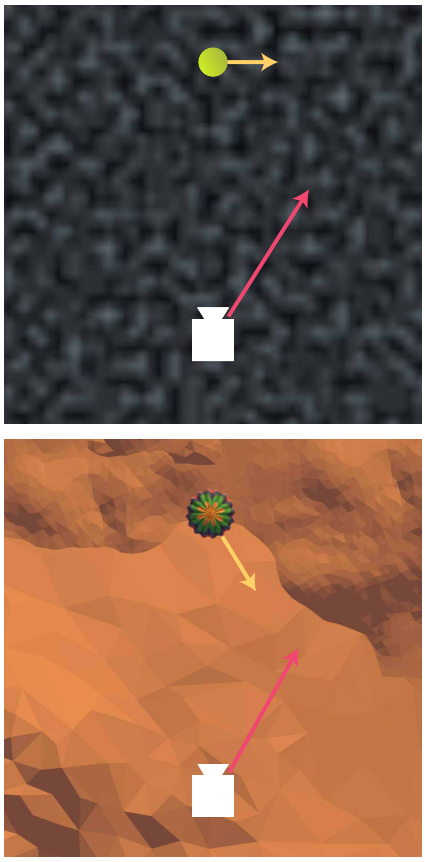
**Conditions.** Beyond determining the psychometric parameters of a single condition, we aim to investigate how these parameters vary with scene motion, and depth. To this aim, we anchor our measurements to a reference condition, where




(a) pilot study protocol: judge left/right




(b) application study protocol: choose direction



(c) top-down view

Figure 4.3: *Study protocols.* (a) In the psychophysical study, a fixation cross is displayed for .5 s at the beginning of each trial. Subsequently, a video plays depicting a scene moving towards the observer at a non-zero heading angle (arrow in (c)). After 1 s, a moving yellow probe (green arrow) is added to the screen. Once the 2 s video finishes, the subject is asked whether the probe was moving left or right. The probe does not have a forward velocity (top of (c)). (b) In the application study, the protocol is near-identical, with three differences. The target object is added at the start of the trial, it has forward velocity (bottom of (c)), and the subject is asked to choose one of seven options to indicate the direction of the object's motion.

$\{v_s = 1 \text{ m/s}, \varphi_s = 15^\circ, h_s = 1.75 \text{ m}\}$, and explore test conditions where only one attribute of the reference changes. These test conditions vary in scene dynamics in speed, $v_s \in \{0.5 \text{ m/s}, 3 \text{ m/s}\}$ and heading, $\varphi_s \in \{5^\circ, 25^\circ\}$, as well as scene content in height, $h_s \in \{.55 \text{ m}, .74 \text{ m}, 5.22 \text{ m}\}$, resulting in a total of 8 study conditions. Note that we vary the observer height $h_s$ to examine the corresponding scene's depth disparity to the target. To provide a more intuitive representation of depth disparity, we henceforth express these conditions via a dimensionless target-scene depth disparity coefficient: $d = h_t/h_s \in \{.05, .3, .9\}$ for each scene height condition, and $d = .7$ for the reference.

Lastly, in the training session, to avoid introducing external bias to subjects' judgment, the trials were deliberately designed as (1) significantly different from trials in the study, and (2) sufficiently easy for classification, yet difficult enough to mitigate potential misinterpretation of the task. So, we selected four trials with $\varphi_s = 40^\circ$, and $\varphi_t \in \{\pm 30^\circ, \pm 40^\circ\}$. The trial with $\varphi_s = 40^\circ$ and $\varphi_t = 30^\circ$ satisfied the requirement (2) above and thus was reused as a *screening* trial to identify subjects who misinterpreted the task even after the training. The *screening* trial was repeated 24 times throughout the study, Each trial was mirrored to ensure left/right balance, resulting in a total of $(11 \times 8 + 24) \times 2 = 224$ main trials (median completion in 21 min).

**Filtering.** To ensure high-quality data from crowd subjects, we employed a two-layer statistical screening. First, we screened inattentive subjects who only made random guesses. An informal pre-pilot study suggested that subjects almost always gave correct responses when $\varphi_t = 3\varphi_s$ as these were easy-to-answer trials. We leveraged this observation and required an accuracy of $\geq 90\%$, or a guess rate of $\lambda < 10\%$, to pass this screen (random guess accuracy is 50%). Second, we screened for subjects who misinterpreted the task and indicated object motion directions relative to the

*observer*. To this end, we required an accuracy of $\geq$ 50% on *screening* trials (where observer-relative accuracy is 0%). Refer to Section 4.A for study results reported without screening trial-based filtering.

### 4.1.2 Results and Discussion.

**Results.** From the initial 78 subjects, we removed 4 (5%) from the attentiveness screen and 36 (46%) from task understanding screen, within a normal range for such crowdsourcing studies [Brühlmann et al., 2020]. In total, 6, 688 trial results were used for further analysis. Prior to combining the left and right heading conditions, we conducted a one-way *Analysis of Variance* (ANOVA) which showed that the direction of heading did not have a significant effect on the subject-aggregated responses ($F_{1,174} = .1, p = .75$).

As described in *Metrics*, we statistically summarized study responses by fitting psychometric curves, extracting the low-dimensional parameters of the threshold, $\mu$, and slope, $\sigma$, for each condition separately (with a fixed $\lambda = 1.6\%$ across all conditions found via the attentiveness screen guess rate). Curve parameters for each series of conditions that varied along a single attribute were interpolated via polynomial regression (quadratic for $\mu$, and linear for $\sigma$). The results are visualized in Figure 4.4. See Section 4.B for individual curve parameters and polynomial term coefficients.

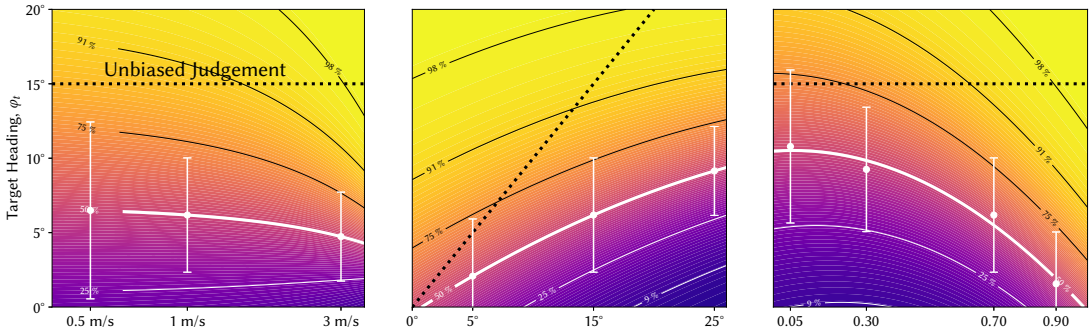
**Discussion.** The statistical analysis demonstrates that we can safely aggregate heading directions in a left-right agnostic manner. The central bias persists across all studied conditions, as evidenced by the measured thresholds below the "unbiased judgment" line in Figure 4.4. This suggests that objects moving to the *right* at a heading angle between the 50% threshold and the unbiased judgment line will be perceived as moving to the *left* by most observers.

We observe other notable trends from the visualization. **From Figure 4.4a**, we observe a steady increase in both bias and consistency. That is, at higher scene speeds, judgments across subjects become more consistent, yet inaccurate. **From Figure 4.4b**, the threshold for the scene heading model intersects at zero degrees, indicating that for forward headings, our perception of lateral motion directions becomes accurate due to the lack of asymmetric optical flow cues. Comparing the unbiased judgment line with the threshold fit suggests that the scale of motion estimation bias is roughly proportional to the scene heading, $\varphi_s$. **From Figure 4.4c**, our perceptual errors increase with the depth disparity between the target and the surrounding scene (i.e., $\uparrow d$). Intuitively, this reveals that if the scene content is too far (e.g.,, the sky), it no longer appears to move nor offer cues to target motion. Conversely, if the scene overlaps with the target (i.e., $d \rightarrow 0$), we still observe a significant bias.

Our 2D-monitor-based study results notably reveal stronger bias compared to prior literature with similar stimuli but in VR ($12^\circ$ when $\varphi_s = 15^\circ$ [Xie et al., 2020b; Xing and Saunders, 2022]). This aligns with previous findings of stereo cues on motion perception [Burlingham and Heeger, 2020; Van den Berg and Brenner, 1994a,b]. The stronger bias observed in 2D displays underscores the crucial need to thoroughly measure, predict, and compensate for human errors in the prevailing computer graphics medium today. This also motivates the future development of 3D displays. In the following section, we utilize our study data to establish a perceptual model predicting human errors in target and scene heading judgment.

### 4.1.3 Modeling Target Motion Errors

**Model Extrapolation.** In Section 4.1.1, we conducted three separate polynomial fits to distinct subsets of the study data, each sharing only the reference condition of

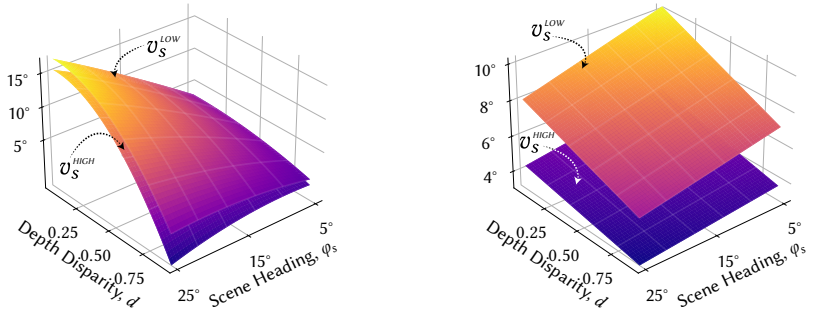



(a) scene speed, $v_s$ (b) scene heading, $\varphi_s$ (c) target-scene depth disparity, $d$

Figure 4.4: *Psychophysical Study Results.* Psychometric curves along (a) scene speed, (b) scene heading, and (c) target-scene depth ratio are fitted from the study data, and interpolated via polynomial regression. Yellow colors indicate majority left responses in the left/right study protocol described in Section 4.1.1. Each curve's threshold is denoted as a scatter with error-bars indicating the jnd offset, or stimulus levels at 25/75% response probability. Contour lines represent jnd step-sizes. "Perfect" unbiased observer's thresholds, as depicted in Figure 4.1b, are visualized as comparison via dotted black lines. Refer to supplementary video for user study conditions which correspond to various points across the heatmaps.

$\{v_s = 1 \text{ m/s}, \varphi_s = 15°, d = .7\}$. By factoring out the parameters of the reference from the fitted models, we express each model as $\mu(v_s) = \mu_r k_v(v_s)$, $\mu(\varphi_s) = \mu_r k_\varphi(\varphi_s)$, and, $\mu(d) = \mu_r k_d(d)$, where $\mu_r$ represents the psychometric threshold of the reference; $k_{v/\varphi/d}$ denote the three individually fitted polynomial models with $\mu_r$ factored out. That is, these models show how the threshold changes due to a change in condition from the reference, meaning, $k_v(v_s = 1 \text{ m/s}) = k_\varphi(\varphi_s = 15°) = k_d(d = 0.7) = 1$. To integrate these individual models into a unified holistic one, we employ a first-order approximation and assume the absence of cross-condition effects. Then, we express the overarching model as:

$$\mu(v_s, \varphi_s, d) = \mu_r k_v(v_s) k_\varphi(\varphi_s) k_d(d). \tag{4.2}$$




(a) threshold, $\mu$ (b) slope, $\sigma$

Figure 4.5: *Full model parameters.* The combined model parameters are visualized as 2D surface slices at two different scene speeds, $v_s^{HIGH} = 3$ m/s and $v_s^{LOW} = 0.5$ m/s. The threshold, $\mu$ indicates the critical heading of observed targets, $\varphi_t$, at which observers, on average indicate that the target is moving neither left nor right toward the observer. The slope, $\sigma$ indicates the confusability between different target headings (i.e., higher $\sigma$ indicates that the ability to discriminate two target headings are poorer). As reported in Section 4.1.1, increasing the scene movement speed increases the perceptual bias (meaning lower threshold) for observers, while decreasing the confusability between targets moving along different heading directions.

This formulation ensures that the trends of each model are extended across a broader spectrum of conditions without compromising the predictive accuracy of the existing conditions. We acknowledge that closer analysis of cross-condition effects could reveal more intricate trends in motion perception errors and is an interesting direction of study, but in the scope of this work, we aimed to determine only the first-order effect, and explore the interesting applications that such a model can enable.

In Figure 4.5, we present a visualization of the predicted psychometric parameters of the combined model. The extended model features combinations of prominent features discussed in Section 4.1.1 such as the decrease in estimation errors as the target-scene depth disparity, $d$, decreases, and the proportional errors with heading direction, $\varphi_s$.

**Predicting Scene-Relative Target Heading.** Thus far, our psychophysical study, and analysis have concentrated on measuring motion judgment errors under the simple condition where the scene-relative target's motion, $\vec{w}_t$, was constrained along a single

axis leftward or rightward (illustrated by dashed yellow vectors in Figure 4.1). But how do these results generalize to conditions where target objects can move in various directions? In order for our model to be applicable for any practical scenarios, it is imperative to establish a framework for extending our perceptual model to accommodate target motions beyond simple lateral movements.

As shown in Figure 4.1b and supported by the relation in Equation (4.1), the poor estimation of the two motions—the scene motion ($\vec{v}_s$) and scene-relative target motion ($\vec{w}_t$)—are dependent on each other. This relationship is expressed as $\vec{w}_t = \vec{v}_t - \vec{v}_s$, where $\vec{v}_t$ represents the target's observer-relative velocity. Hence, an observer's misjudgment of scene-relative target movement corresponds to an opposite misjudgment of scene movement:

$$\vec{w}'_t = \vec{v}_t - \vec{v}'_s. \tag{4.3}$$

In our study, the psychophysical thresholds indicate the critical value $\vec{v}_t$, with a corresponding heading of $\varphi_t = \mu(v_s, \varphi_s, d)$, at which $\vec{w}'_t = 0$. By incorporating these results into Equation (4.3), we conclude that our model yields the perceived heading of scene motion, which our study has shown to deviate from the actual heading:

$$\varphi'_s = \mu(v_s, \varphi_s, d). \tag{4.4}$$

Ultimately, by combining Equations (4.3) and (4.4), we derive an expression for estimating the perceived scene-relative target motion:

$$\vec{w}'_t = \vec{v}_t - \vec{v}'_s = (\vec{w}_t + \vec{v}_s) - \vec{v}'_s = \vec{w}_t + \vec{v}_s - (R_\mu \hat{z}) v_s \tag{4.5}$$

where $R_\mu \hat{z}$ represents the *forward* unit vector (see Figure 4.1a) laterally rotated by $\mu(v_s, \varphi_s, d)$. We visualize this vector sum in Figure 4.8a.

## 4.2 Model Validation

### 4.2.1 Measuring Model Robustness

To ensure model robustness, we conduct a numerical validation by fitting the model to half of the experimental data, and measure its goodness-of-fit to the other half of the data unseen by the fitted model. Specifically, each of the $n = 38$ subjects' data is randomly partitioned into either a model fitting or evaluation group. We then assess the model's prediction accuracy compared to the observed data using the $R^2$ coefficient for each study condition. Due to the arbitrary nature of the subject partitioning operation, we repeated this procedure $N = 20$ times, and observed that the lowest score recorded was .61, while the mean score across all conditions and repeats to be .95, compared to the full model's self-fitting score of .98, indicating acceptable fits [Ozili, 2023].

### 4.2.2 Generalizability Over Population

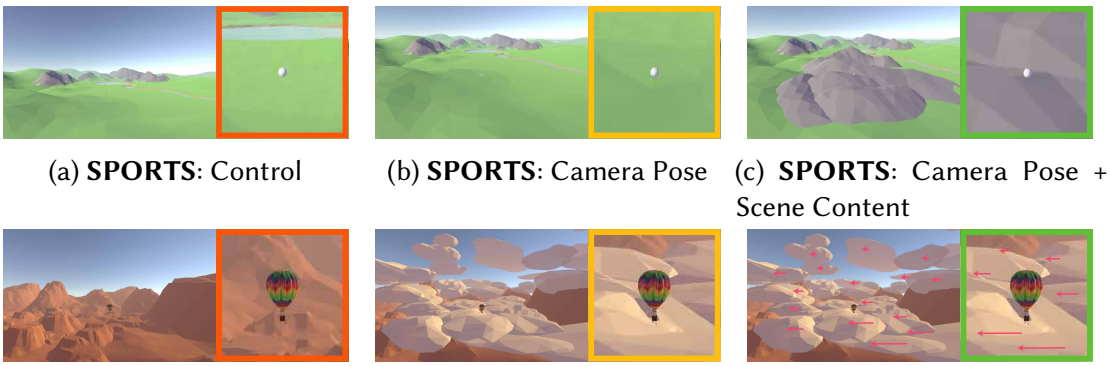
We validate whether the psychometric curves fitted from the sample population in Section 4.1.1 can generalize to unseen subjects. To this aim, we conducted a smaller-scale user study featuring only the *reference* condition from our main study in Section 4.1.1 on a new subject group ($n = 23$, ages $22 - 52$, 11 males). This study replicated the study protocol, stimuli, and crowdsourcing-based recruitment methods of Section 4.1.1.

**Conditions.** Our goal in this study was to investigate the variability of motion judgment errors across different subjects and to use the results to validate our main

study in Section 4.1.1. To keep the study duration and cost feasible, we only studied the reference condition from the main study (i.e., $\{v_s = 1 \text{ m/s}, \varphi_s = 15^\circ, d = .7\}$) and increased the number of repetitions for each trial (10 repeats) to sufficiently fit corresponding psychometric curves for individual subjects. Step sizes between target heading levels, $\varphi_t$, were decreased to $4.2^\circ$ to ensure higher precision measurements. Overall, the study consisted of 80 *measurement* trials, 20 *filler* trials featuring random conditions to prevent categorical judgments [Xing and Saunders, 2022], and 48 *screening* trials (see Section 4.1.1 for details) for a total of 148 trials completed in 15 min by the median subject.

**Results and discussion.** We fit individual psychometric curves to each of the subjects' aggregated study responses, and observed a mean threshold, $\mu_{\text{avg}} = 4.6^\circ \pm 1.1^\circ$ *Standard Error Mean* (SEM) and mean slope, $\sigma_{\text{avg}} = 6.2^\circ \pm 1.4^\circ$ SEM for the condition identical to the reference of our main study. A single sample $t$-test indicates that the mean threshold and slope from the main study $\mu = 6.2^\circ$ and $\sigma = 5.7^\circ$ is not significantly different from the distribution of thresholds and slopes in the evaluation study, $t(22) = -1.4, p = .18$ and $t(22) = .35, p = .73$, respectively.

The statistical analysis demonstrates that the psychometric threshold found for the reference condition in our main study lies within acceptable limits of thresholds of out-of-population individuals. While the approach for establishing representative psychometric curve parameters utilized in this evaluation study are more robust due to the larger volume of samples we collect per-subject, we note that conducting a main study of similar scale in terms of different conditions studied becomes unfeasible in practice due to prohibitively high study durations and costs.



(a) **SPORTS**: Control (b) **SPORTS**: Camera Pose (c) **SPORTS**: Camera Pose + Scene Content

(d) **FLIGHT**: Control (e) **FLIGHT**: Static Scene (f) **FLIGHT**: Dynamic Scene

Figure 4.6: *Application case study protocols and scenes.* (a)/(d) shows the original animations of the target and camera simultaneously moving in a 3D scene. Both the model prediction and our study results indicate that the animation design induces significant perceptual errors in users' perceptual error of target motion. To reduce such errors, our model enables predictive suggestions for design optimizations, such as adjusting camera poses (b), as well as adding static (c)/(e) and dynamic (f) background geometries.

## 4.3 Application Case Study: Animation Design Guidance

Scene dynamics, including camera and object motion control [Hsu et al., 2013], as well as scene content, such as depth [Kellnhofer et al., 2013], are crucial factors in animation design [Jiang et al., 2021; Lino and Christie, 2015], video editing [Kang and Cho, 2019], and game development [Caroux et al., 2013]. Traditionally, the design of these factors has been implicitly driven by aesthetics or storytelling.

We investigate observers' perceptual errors in the target dynamics with two 3D animations. Subsequently, we propose model-guided design alterations, including optimizing camera pose, adjusting the placement of scene objects, and introducing subtle motions to them, to mitigate the predicted perceptual errors. We evaluate the effectiveness of these scene design improvements by conducting multiple-choice user studies.

### 4.3.1 Experimental Design

**Participants and procedure.** We conducted two user studies via crowdsourcing and recruited $n = 22$ subjects (ages $20 - 64$, 10 male) for each. Unlike the two-alternative forced choice (left vs. right judgment) tasks in Section 4.1.1, subjects in this study directly indicated perceived scene-relative directions of target motion. As shown in Figures 4.3b and 4.8, they chose from one of seven options, each representing a scene-relative target heading of $\psi_t \in \{\pm 30^\circ, \pm 20^\circ, \pm 10^\circ, 0^\circ\}$. After viewing a 2 s video featuring a moving target within a moving scene, subjects referred a top-down view presented at the end of each trial and pressed a button to indicate their choice.

**Stimuli.** Two realistic scenes, along with corresponding target objects, were used to simulate common gaming and simulation animations: (1) sports gaming with golf (**SPORTS**), and (2) flight simulation (**FLIGHT**), as shown in Figure 4.6. In both scenes, as depicted in Figure 4.3c, the scene moves towards the observer at a heading of $\varphi_s = 25^\circ$ with a speed of $v_s = 1$ m/s and .5 m/s for **SPORTS** and **FLIGHT**, respectively (the scene and target sizes were re-scaled to align with the scaling of our model).

Each scene features a target object: a golf ball, and a hot-air balloon. At the start of each trial, the target object appears at a random location within $10^\circ$ from the fixation point, and a distance of $12 - 14$ m and $6 - 7$ m from the observer for each scene. The target moves towards the observer along the $3^{\text{rd}}$ trajectory in Figure 4.8 at a heading of $\varphi_t = 10^\circ$ and a speed of $v_t = 2.8 \times v_s$. The observer-relative motion of the target is equivalent to a scene-relative motion along the $6^{\text{th}}$ trajectory in Figure 4.8, or $\psi_t = -20^\circ$.

Each subject completed 10 repetitions of these trials as well as 5 more *filler* trials with random target object headings to prevent categorical responses. We provided mirrored motions for each trial to ensure left-right balance for a total of 30 trials per

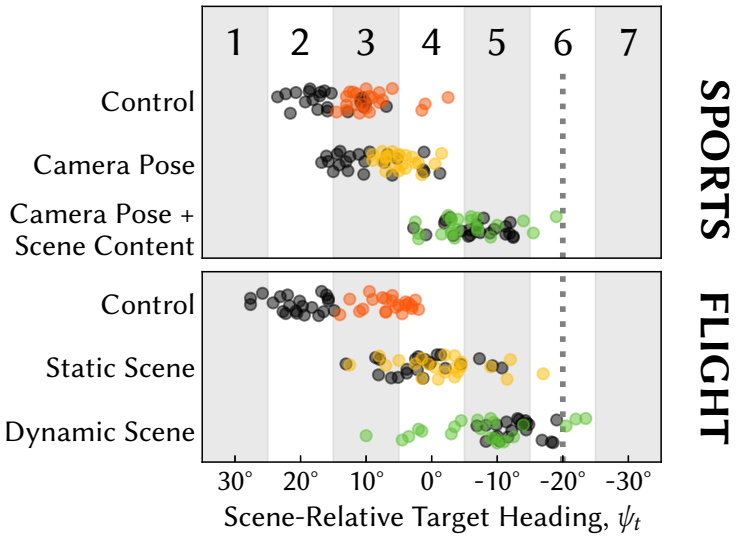



Figure 4.7: *Results of the application case study.* The x-axis shows the scene-relative target heading angles corresponding to individual options (1-7) provided in the study. The red and yellow/green points represent the distribution of per-subject aggregated mean response data in control and our model-suggested re-designed animations, respectively. The black points represent the corresponding response distribution simulated from our model prediction. The points ($\psi_t'$) are vertically jittered for plot visibility.

study condition. Similar to our psychophysical study in Section 4.1.1, subjects also completed a pre-study *training* session with a straight-ahead heading $\varphi_s = 0°$, and targets moving along 1$^{st}$, or 7$^{th}$ trajectory (i.e., $\psi_t \in \pm 30°$). The median completion time was 15 min.

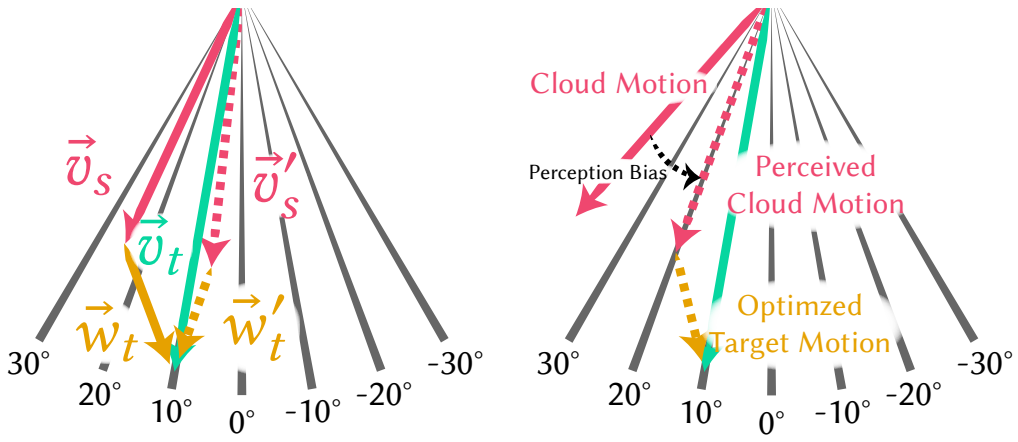
**Conditions.** For each scene, we prepared two content re-design "treatment" conditions without changing the original camera motion trajectory, when compared to the control conditions shown in Figures 4.6a and 4.6d. As evidenced in Figure 4.4c, decreasing target-scene depth disparity, $d$, reduces perceptual errors. Thus, to address this issue, in **SPORTS**, the first re-design elevates the camera height, and lowers the viewing angle for a more "birds-eye" view (Figure 4.6b). As a more aggressive re-design, we also added scene elements behind the target golf ball to further decrease depth disparity

(Figure 4.6c). Across these three scenes, the average scene-target depth disparities were $d = .1/.6/.7$, respectively. Using our model and target heading prediction framework of Section 4.1.3, we determined $\psi'_t = 16°/10°/-5°$ for the three conditions respectively.

Similarly, for **FLIGHT**, we first added static cloud objects into the scene to decrease the depth disparity from $d = .8$ to $d = .4$ as shown in Figure 4.6e. For the second treatment, we took a different approach by attempting to simulate a different scene heading by adding a horizontal drift velocity, $v = .25$ m/s, to the clouds relative to the rest of the scene to reinforce the lateral direction of optical flow and induce a higher perceived scene heading angle of $\varphi_s = 37°$ (see Figure 4.8b). In effect, our model predicts that the perceived scene-relative target heading for the target hot-air balloon was $\psi'_t = 22°/-1°/-12°$, respectively.

### 4.3.2 Results and Discussion

**Results.** For both studies, we summarize the mean response of each subject and each condition by aggregating across the 20 recorded trials. Figure 4.7 compares the acquired distributions of target headings $\psi'_t$ with the model-prediction. Across subjects, in **SPORTS**, the measured mean and SEM target headings were $\psi'_t = 9.1° \pm .91°$, $4.8° \pm .60°$ and $-5.5° \pm 1.2°$ for the control, camera pose and additional scene content conditions respectively, while in **FLIGHT**, the measurements were $\psi'_t = 6.5° \pm .71°$, $-1.8° \pm 1.5°$ and $-7.5° \pm 1.8°$ for the control, static scene and dynamic scene conditions, respectively. Across all conditions, the ground-truth scene-relative target heading was $\psi_t = -20°$. A repeated measured ANOVA shows that the conditions within each study had a significant effect on the mean responses for both **SPORTS** ($F_{2,42} = 94.0, p < .01$) and **FLIGHT** ($F_{2,42} = 65.6, p < .01$) scenes.




(a) baseline prediction (b) design optimization

Figure 4.8: *Predicting and compensating target motion estimation in animation design.* (a) Similar to the illustration in Figure 4.1a, an observer may erroneously perceive the target motion $\vec{w}_t$ as $\vec{w}'_t$ by judging from $\vec{v}_t$ on screen. As shown in Figure 4.6, we leverage our model to alter the scene designs in various ways to reduce the error. (b) We take the "Dynamic Scene" condition in **FLIGHT** (Figure 4.6f) as example. The model-guided cloud motion alters observers' perception so that $\vec{w}'_t$ becomes closer to $\vec{w}_t$ (as evidenced in Figure 4.7).

**Discussion.** As shown by the ANOVA results, the model-guided content re-design significantly improved the accuracy of target heading judgments for the subjects. Our model was able to predict the overall trend of heading judgment errors, although the exact numerical predictions were slightly inaccurate. We attribute this performance regression to the introduction of higher-order cognitive cues in the more realistic stimuli and discuss its implications further in Section 4.4. Nevertheless, our model is still capable of providing a first-order approximation of the relationship between observer-relative scene and target velocities ($\vec{v}_s$ and $\vec{v}_t$) and the scene-relative target velocity ($\vec{w}_t$). In real-world applications, we can leverage these predictions to provide guidance and feedback on the overall estimation difficulty, and anticipated motion judgment errors users are likely to make when observing dynamic imagery.

## 4.4 Limitations

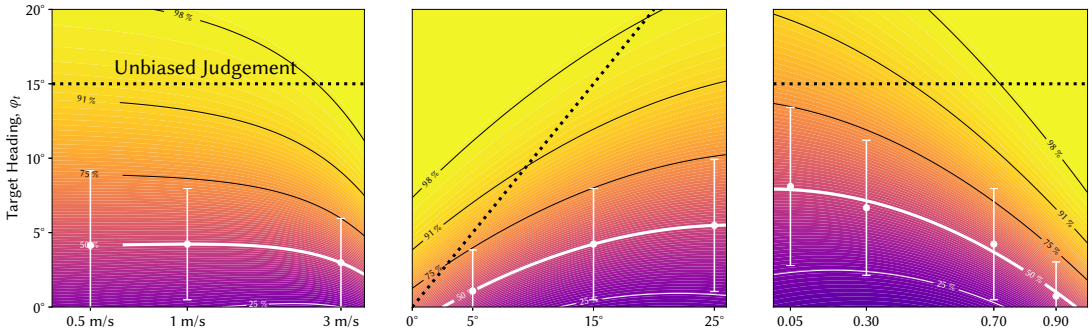
**Additional cues.** Beyond image space, stereo [Burlingham and Heeger, 2020] and vestibular [DeAngelis and Angelaki, 2012] cues from emerging 3D displays may also alter motion perception, together with semantic and cognitive influences, including human body pose [Blake and Shiffrar, 2007], visual path information [Li et al., 2009], and object shadows [Kersten et al., 1997]. Meanwhile, many of these phenomena rely on higher-order cognitive cues beyond low-level visual operators. For example, understanding the relationship between the motion of objects and the shadows they cast requires spatial reasoning and is a non-intrinsic, learned skill in humans [Van de Walle et al., 1998]. In this work, we chose to first establish a baseline for human perception at an abstraction level where all high-level cues were absent, and the only source of information was the optical flow derived from motion within a 3D environment. After confirming significant perceptual errors under these abstract baseline conditions, we then constructed a more realistic synthetic scene in Section 4.3 to determine whether any of the baseline estimation errors persist and to assess if our model can still mitigate these errors within the scope of our chosen parameterization, despite the introduction of high-level factors. We believe that these experiments successfully demonstrate the effective application of optimizing animation design pipelines as a first-order measurement and mitigation of human perceptual errors.

**Cross-conditions.** In Section 4.1, we characterize the scene dynamics with self movement (direction and speed) and content depths (with regard to the object). Exploring additional combinations of scene and object dynamics, such as rotations and vertical movements, leads to a prohibitively large number of trials. This poses challenges due to participants' limited attentive capacity for maintaining data accuracy, as well as

the associated financial costs or running long studies. Therefore, this research focuses on separately measuring the effects from individual dimensions. To study the cross-conditions while maintaining feasibility, we plan to first analyze a primary effect via a pilot study similar to In [2017], and extend the work towards a dimension-reduced study.

**Motion degrees of freedom.** We study perceptual errors for horizontal motion patterns along transverse (horizontal) planes—the more common human motion [Hummel et al., 2016]. However, both object and scene motions together form a complex 12 degrees of freedom (DoF) problem (6 DoF each for the self and the object) across all planes, including the coronal and sagittal. In such case, a rotating observer or object will elicit a moving FOE [Danz et al., 2020]. Therefore, introducing a temporal movement factor to the FOE, a.k.a., its *locus*, could be a key to modeling arbitrary motions [Rangarajan and Shah, 1992]. Additionally, camera motion analysis using a large-scale egocentric motion dataset (e.g., Ego4D [Grauman et al., 2022]) could establish a coordinate system tailored for the most prevalent human motion patterns.

**Perceptual attention and confidence.** In highly complex scenarios, various objects may move in different directions. The confounding optical flow may further compromise observers' perception in understanding the motion [Warren et al., 1988; Warren Jr and Hannon, 1988]. Moreover, because of humans' selective attention, the movement of multiple objects can also interfere with the visual sensitivity towards a specific target [Min and Corso, 2019]. Our current model assumes full attention to a single target. In the future, we plan to explore the influence from optical flow entropy toward a more content-aware probabilistic model.




(a) scene speed, $v_s$ (b) scene heading, $\varphi_s$ (c) target-scene depth disparity, $d$

Figure 4.9: *Unfiltered Study Data Analysis.* Results of processing the data without applying the task understanding filter are visualized for comparison with Figure 4.4. See the caption for Figure 4.4 for details on the visualization designs.

## 4.A Unfiltered Psychophysical Data Analysis

In this work, we rejected a significant number of subjects via our task understanding filter, as described in Section 4.1.1, to ensure high quality data acquired from crowdsourced study participants. Here, we present the psychometric curve fitting results for the unfiltered data to serve as a comparison to the results included in the main manuscript. In Figure 4.9, we replicated Figure 4.4 to serve as a direct comparison between the filtered and unfiltered data. The psychometric threshold for the reference condition was $\mu_r = 4.2°$ when compared to $\mu_r = 6.2°$ as reported for the unfiltered data. The *Discussion* about the trends and patterns of the psychophysical study results in Section 4.1.1 are largely unchanged for the unfiltered data, albeit with a much stronger bias effect.

Table 4.1: Psychometric parameters for different scene speeds, headings, and depth ratios.

| **Attribute** | **Value** | **Threshold, $\mu$ (°)** | **Slope, $\sigma$ (°)** |
|---|---|---|---|
| $v_s$ (m/s) | 0.5 | 6.5 | 8.8 |
| | 1 | 6.2 | 5.7 |
| | 3 | 4.7 | 4.4 |
| $\varphi_s$ (degrees) | 5 | 2.1 | 5.7 |
| | 15 | 6.2 | 5.7 |
| | 25 | 9.1 | 4.4 |
| $d$ | 0.05 | 10.8 | 7.6 |
| | 0.3 | 9.2 | 6.2 |
| | 0.7 | 6.2 | 5.7 |
| | 0.9 | 1.6 | 5.2 |

## 4.B Psychometric and Polynomial Fitting

Below, we list the parameters for all the psychometric curves fitted using the data collected from our psychophysical study of Section 4.1.1:

These psychometric parameters were then regressed to fit polynomial curves with fitted parameters $\mu_r = 6.2°$ and $\sigma_r = 5.7°$.:

$$\begin{aligned}
\mu(v_s, \varphi_s = 15^\circ, d = .30) &= \mu_r \times (.931 + .077v_s + .006v_s^2),\\
\mu(v_s = 1 \text{ m/s}, \varphi_s, d = .30) &= \mu_r \times (.045 + .054\varphi_s + .001\varphi_s^2),\\
\mu(v_s = 1 \text{ m/s}, \varphi_s = 15^\circ, d) &= \mu_r \times (.531 - .171d + 1.390d^2),\\
\sigma(v_s, \varphi_s = 15^\circ, d = .30) &= \sigma_r \times (1.486 - .302v_s),\\
\sigma(v_s = 1 \text{ m/s}, \varphi_s, d = .30) &= \sigma_r \times (1.093 - .011\varphi_s), \text{ and,}\\
\sigma(v_s = 1 \text{ m/s}, \varphi_s = 15^\circ, d) &= \sigma_r \times (1.308 - .459d).
\end{aligned} \tag{4.6}$$

# Chapter 5

# Decision-Making Latency Effects from Visual Signal Characteristics

Measuring, modeling, and predicting how humans perceive and act on displayed visual content are important tasks in computer graphics, with applications in cinematic, real-time rendering, virtual/augmented reality (VR/AR), display optimization, esports, video compression/streaming, and visual design [Dunn et al., 2020; Mantiuk et al., 2004; Patney et al., 2016; Serrano et al., 2017; Sitzmann et al., 2018]. Perceptual image quality metrics predict the likelihood of visibility of image artifacts that result from creative and technical design, or are a side-effect of rendering, processing, or transmission. While many such metrics already exist, research is primarily focused on modeling the spatial/temporal *acuity* of the human visual system (HVS), not on how viewers "react" *after* perceiving the stimuli. Although visibility may be closely related to behavior, learning the transfer function between the appearance of visual stimuli and the different reactions observers might exhibit. Since responses are critical in many interactive applications such as esports and user interfaces, metrics that predict user reactive

performance are arguably in emerging and crucial demand.

Researchers have so far exhaustively studied the acuity of the human visual system and established a significant body of perceptual image-quality metrics [Hore and Ziou, 2010], as well as perceptually-optimized computer graphics techniques [Krajancich et al., 2021; Patney et al., 2016]. As discussed in greater detail in Section 2.1, such methods have unlocked significant performance and memory optimizations, as well as quality improvements. While a great deal of work has focused on how human perception can affect how we design graphics systems, behavior-aware computing is a relatively new field (see Section 2.1.3), and thus doesn't feature comprehensive literature focusing on this topic.

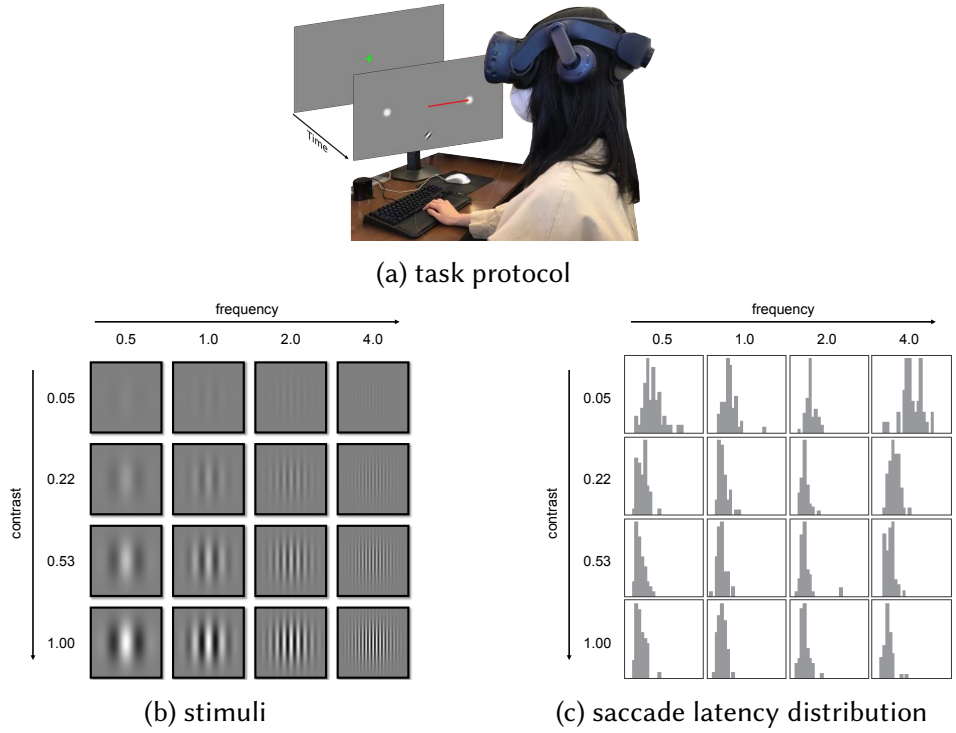
In this chapter, we study analytical models of user reactions based on the visual features of displayed content. Specifically, we explore how users make decisions by observing their eye movement behaviors. We propose an analytical model for a user's reaction time as evidenced by their eye movements. Specifically, across two main experimental designs, we study the temporal characteristics of saccadic eye movements (covered in further detail in Section 2.3.4) and use these measurements to infer human ability to make decisions based on the visual properties of the stimuli that they observe.

Saccadic reaction latencies, after the eye observes a stimulus, are closely tied to performance in a broad range of real-time applications. For instance, subtly (as low as 4ms [Kim et al., 2019]) altered saccade latency can significantly determine performance in competitive esports [Koposov et al., 2020]. Each saccade involves perceiving a stimulus, identifying the target [Lisi et al., 2019], sending oculomotor neural signals, and controlling the extraocular muscles to reorient the eyeballs. Due to these complex mechanisms, fully characterizing changes in saccade/fixation as a function of changes in visual stimuli remains an open problem in vision science and computer graphics.

Note that, unlike with visual quality metrics, both high and low visibility of a target could hypothetically induce a longer processing time for fine details or blurred content. That may lead to potential non-correlation between acuity and saccadic latency [Kalesnykas and Hallett, 1994]. Indeed, while visibility has been shown to be closely related to behavior, there is evidence that perceptually identical stimuli frequently result in significantly different reactions for observers [Mulckhuyse and Theeuwes, 2010; Spering and Carrasco, 2015]. Therefore, this work presents behavioral models which correlates the visual features of visual stimuli to the timing of humans' cognitive decisions to perform saccades. We adapt to different viewing tasks that involve reacting to both static and dynamically moving visual stimuli. Furthermore, we also demonstrate how these behavioral models can be leveraged to make predictions, and suggest optimizations toward customizing user action timings, measuring competitive fairness in video games, and predicting user performance across different display environments. Source code and data for this chapter's contents are available at `www.github.com/NYU-ICL/gaze-timing` and `www.github.com/NYU-ICL/pursuit-timing`.

## 5.1 Measuring Discrimination Latency

In our first experiment, we will study the decision-making latency in visual discrimination tasks. We begin by conducting a psychophysical experiment with parameterized stimuli to observe and measure the correlation between image characteristics and the time it takes to process and discriminate them in order to trigger a saccade, and whether/how the correlation differs from that of visual acuity. In the following section we will leverage our collected data to construct a predictive model of visual discrimination latency.




(a) task protocol

(b) stimuli

(c) saccade latency distribution

Figure 5.1: *Preliminary user study procedures and results.* (a) shows our setup and the study procedure: two *target* Gaussian patches are shown left and right from the initial fixation. After a brief delay of 300 to 500 ms, a *reference* Gabor stimulus appears in the inferior periphery. If the *reference* stimulus is oriented at 45° clockwise from the vertical axis, the correct *target* saccade location is on the right side, and vice versa for a *reference* stimulus with the opposite orientation (i.e. counter-clockwise orientation). The *latency* of the saccade response indicating the decision is recorded. Across trials, the contrast, frequency, as well as vertical eccentricity of the *reference* Gabor stimulus are varied as experimental parameters. *Target* Gaussian patches are unchanged across all trials. (b) visualizes all the stimuli used for this study. Chosen contrast values are $c = \{0.05, 0.22, 0.53, 1.0\}$ as measured by Weber contrast; frequency values are $f = \{0.5, 1.0, 2.0, 4.0\}$ cpd. All stimuli were shown at eccentricity values of $e = \{0°, 10°, 20°\}$. (c) histograms of saccade latencies for one sample subject when the reference stimulus was located at 0° eccentricity. The distributions exhibit a skewed asymmetrical shape, similar to other distributions of reaction time in related work (see Figure 2.2). With $\{c = 0.53, f = 2\text{cpd}, e = 10°\}$ as the reference stimulus, all stimuli images (from (b)) show high and similar FoVVDP scores ($9.52 \pm 0.03$), despite significant variances in their resulting saccade latencies.

Table 5.1: *Specifications of the HTC Vive Pro Eye display used in our studies.*

| Feature | Value |
|---|---|
| Display Resolution | $1440 \times 1600$ pixels per eye |
| Refresh Rate | 90 Hz |
| Peak Luminance | 143 $cd/m^2$ |
| Field of View | 110° diagonal |
| Eye Tracker Frequency | 120 Hz |

### 5.1.1 Experimental Design

**Setup.** The study was performed with an eye-tracked HTC Vive Pro Eye head-mounted display as shown in Figure 5.1a and implemented in the Unity Game Engine. The hardware details are specified in Table 5.1. During the study, participants remained seated and perceived stimuli through the stereo display. Before each experiment, a five-point eye-tracking calibration was applied on each individual.

**Participants.** The psychophysical study was performed with $n = 5$ participants (ages $22 - 28$, 3 female) with normal or corrected-to-normal vision. The participants were instructed to perform a series of two-alternative forced choice (2AFC) tasks for each trial. The experiment was conducted during a single session split into 10 blocks, with each block containing 225 trials, i.e., 11250 trials in total with all the participants. The procedure took around 2.5 hours for each participant, including breaks between blocks, a short training session preceding the experiment, and a debrief afterwards.

**Stimuli and Tasks.** Figures 5.1a and 5.1b illustrate the experiment procedure and stimuli. The task is to:

1. fixate at the center of the display,

2. when visible, identify the orientation (i.e., symmetry axis) of the Gabor pattern presented at some eccentricity in the visual field, and

3. make a saccade either to a left or a right target based on the orientation of the Gabor pattern.

We include Gabor patches for all combinations of contrasts ($c = \{.05, .22, .53, 1.0\}$), frequencies ($f = \{.5, 1.0, 2.0, 4.0\}$ pixels-per-degree), and eccentricities ($e = \{0°, 10°, 20°\}$). Three conditions (with $(c, f, e)$ values of $(.05, 4.0, 10°)$, $(.05, 4.0, 20°)$, $(.22, 4.0, 20°)$) were excluded due to the patches not being detectable by all participants. The eccentricity range was chosen to cover common scenarios since the human gaze does not typically go outside 10° from the center [Hatada et al., 1980], and most natural saccade sizes are less than 15° [Bahill, 1975]. Unless otherwise specified, we use Weber contrast in all our experiments and as input to our model.

At the beginning of each trial, the participants fixated at a cross shown in the center of the screen. Once they successfully fixated on the cross, it disappeared and a pair of Gaussian patches appeared at 10° eccentricities to the left and right of fixation. These patches served as the target locations to which the participants would saccade to indicate their decision about the stimulus. After a small delay—chosen randomly between 300 and 500 ms to avoid learning effects—the primary stimulus (Gabor patch) appeared either at the center of the screen (eccentricity=0°), or at some eccentricity in the inferior peripheral vision (eccentricity=10° or 20°). We instructed the participants to identify whether the Gabor stimulus was oriented at a rotation of 45° clockwise from vertical, as shown in Figure 5.1a, or 45° counter-clockwise from vertical. We further instructed them to saccade to the target patch corresponding to their determination, right for clockwise and left for counter-clockwise. During each trial we recorded the

subjects gaze at a rate of 120 FPS using the display's built-in eye tracker.

We varied the eccentricity, contrast, and frequency of Gabor patterns across trials such that each combination of variables was shown 5 times in each block for 10 blocks, yielding a total of 50 trials per condition. To ensure the participants were completing the task correctly, we discarded all trials where they do not complete the task correctly, and repeat all mistaken trials at the end of the block until all trials are completed. The order of these conditions was randomly shuffled within each block to eliminate any bias. Meanwhile, all features of the Gaussian target patches (only being used to cue the saccade direction) remained unchanged throughout the trials. For the practice session at the beginning of the experiment, each participant performed one block of the study with identical settings as in the actual study. Please refer to our supplementary video for an animated illustration.

**Saccade Analysis.** Our method of detecting reaction times for saccadic events is measured by the time of onset of the "primary" saccade that is used to move the gaze to the target location. We define the "primary" saccade as the saccade that is onset and offset within 3° of the intended gaze origin and target locations respectively. For saccade detection we use the method presented by Engbert and Mergenthaler [2006].

### 5.1.2 Results and Discussion

**Results.** Using this saccade detection method, we identify the saccadic latency as the duration between appearance of the primary stimulus (Gabor patch) and the first frame of a participant's saccade. We notice that the saccade latencies exhibit an asymmetrical distribution as shown in Figure 5.1c. As the various features of the stimulus are modulated, the overall shape of the distribution remained consistent while the mean saccade

latency varied by as much as 25% or 100 ms as shown in Figure 5.2. Increasing the contrast of the stimuli decreases reaction latency, while increasing the frequency increases the latency. Further, increasing the eccentricity does not always reveal a monotonic effect, but instead a U-shaped effect with the lowest mean latency values (265 ms) plateauing at 10°. For breakdown visualizations, please refer to Figure 5.2/Figure 5.14 for the effects of individual characteristics and participants.

**Discussion.** The above results and analysis reveal several remarkable discoveries on the relationships between visual characteristics and saccadic latency. The asymmetrical probability distribution agrees with the discoveries of prior work in measuring similar visual-oculomotor reactive latencies [Carpenter and Williams, 1995; Lisi et al., 2019; Palmer et al., 2011]. Additionally, at a given eccentricity, as the visibility of the stimuli improves (either by increasing contrast or by modulating the frequency), the latency decreases. Meanwhile, the latency rises toward infinity whenever visibility reduces and approaches the Contrast Sensitivity Function (CSF) threshold. Notably, when considering stimuli of equal contrast, the measured latency appears to scale at a similar scale as the contrast sensitivity corresponding to the frequencies of the visual targets. While not investigated further in this experiment's analysis, we return to these ideas in Section 5.3. Lastly, we observe a surprising effect that the saccade latencies for a stimulus at the fovea are longer than in mid-periphery. We hypothesize that the more analytic purpose of the fovea causes feature extraction to take longer, similar to the results reported by Kalesnykas and Hallett [1994].

The collected data and the observations drive our development of a closed-form probabilistic model inspired by the computational process of decision making, as detailed in the next section.

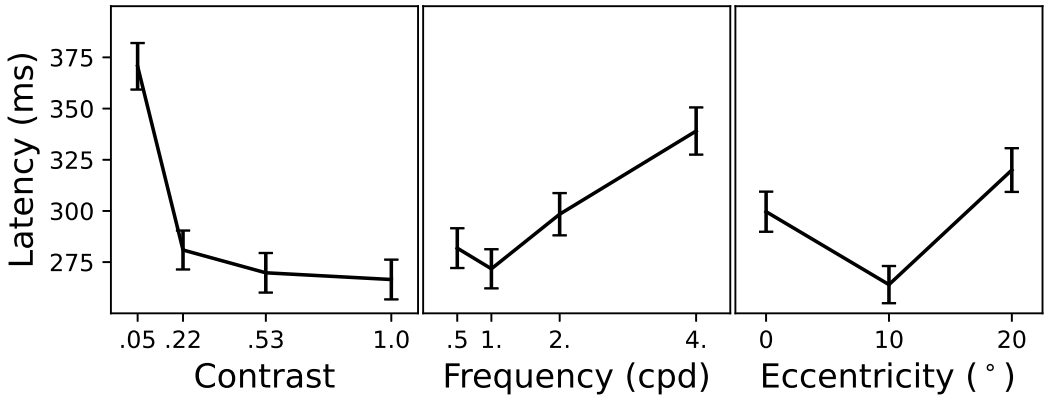



Figure 5.2: *Aggregate trends of our preliminary study dataset.* The pilot study raw data is aggregated using either contrast, frequency or eccentricity of the reference Gabor patch, and averaged across the other two variables. Error bars represent standard error of measurement. Reaction times decrease as visibility of the stimuli is improved, and vice versa. Surprisingly, the reaction latency when the stimulus is at the fovea is higher as compared to when it is in mid-periphery.

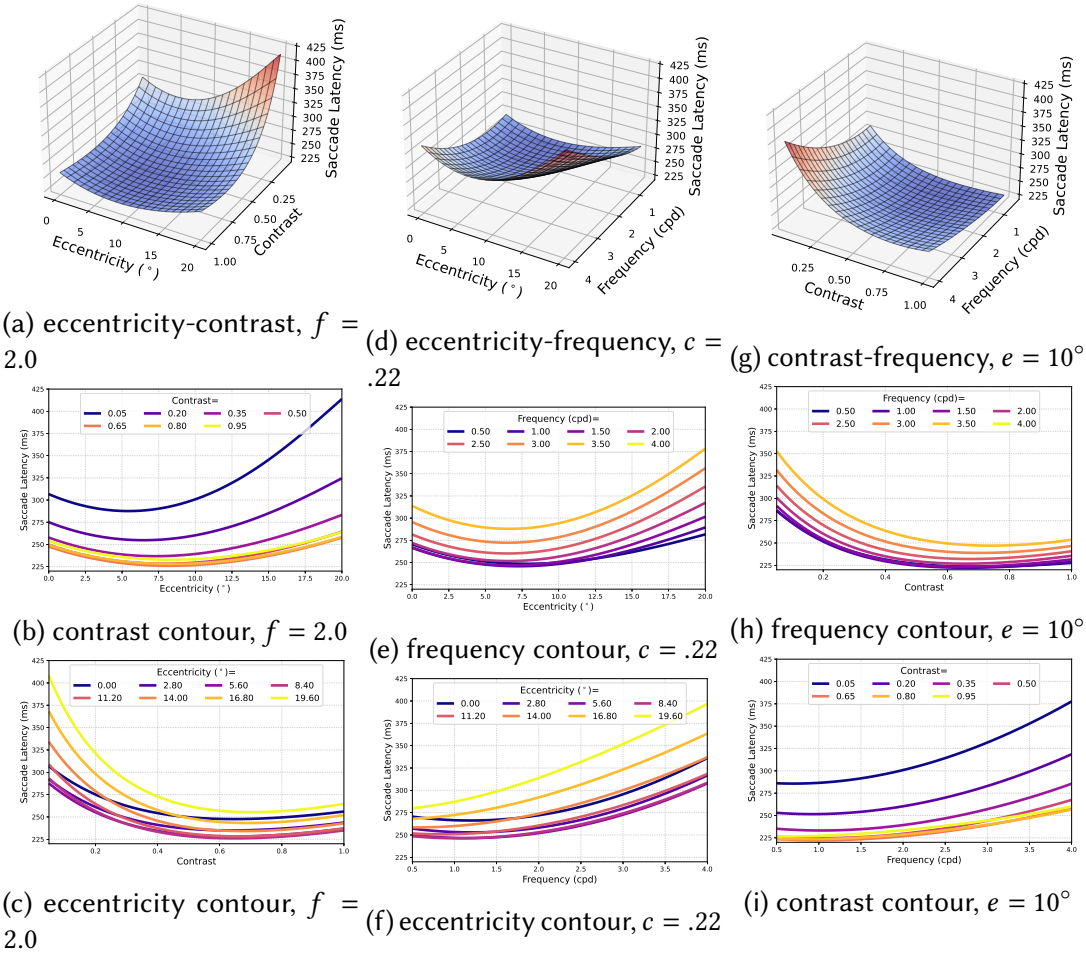
## 5.2 Behavioral Model of Discrimination Latency

### 5.2.1 Model Specification

In Section 2.2.2, we reviewed how speeded decision-making tasks are often modeled using the Drift Diffusion Model (DDM). Comparing the temporal distribution of saccade latencies observed in our study (Figure 5.1c) with those predicted by the DDM (Figure 2.2), we find that their shapes are similar—both exhibit bell-shaped distributions with positive skew.

This similarity aligns with the theoretical basis of our experiment: saccade onsets result from processing visual information from the target stimulus until sufficient evidence is accumulated to trigger a decision to shift the eyes. Given this process, research on saccadic decision-making frequently employs the DDM Myers et al. [2022].

In Equation (2.7), we established that, according to the DDM, the distribution of latencies is parameterized by the stimulus-dependent evidence accumulation rate, $r$,




(a) eccentricity-contrast, $f = 2.0$

(d) eccentricity-frequency, $c = .22$

(g) contrast-frequency, $e = 10°$

(b) contrast contour, $f = 2.0$

(e) frequency contour, $c = .22$

(h) frequency contour, $e = 10°$

(c) eccentricity contour, $f = 2.0$

(f) eccentricity contour, $c = .22$

(i) contrast contour, $e = 10°$

Figure 5.3: *Visualization of our model.* With a given task $D$, our model, defined in Equation (5.1), is $\mathbb{R}^3 \rightarrow \mathbb{R}$. The first row visualizes each two of the three dimensions $(c, f, e)$ as the variable to the saccade latency (z-axis). The second/third rows are the corresponding contours created by projecting the model to x-z/y-z axes. Note the U-shaped effects of $e$, and the inverse effects between $f$ and $c$.

and the response bias-dependent evidence criterion, $\alpha$. In this section, we examine how these parameters relate to the characteristics of visual stimuli—such as contrast, spatial frequency, and eccentricity—as well as response biases associated with the visual task of interest. Furthermore, we analyze how these parameters in Equation (2.7) are influenced by the nature of the visual content and the specific demands of the task.

**Evidence Criterion.** As discussed in Section 2.2.2, the evidence criterion, $\alpha$, reflects both an individual's response bias and the specific requirements of the task they are performing. Palmer et al. [2011] demonstrated that variations in visual signal strength do not significantly alter $\alpha$ within a given experimental procedure. However, it can be manipulated across different visual tasks, such as feature search, conjunction search, and spatial configuration search.

Furthermore, Reddi et al. [2003] showed that modifying task instructions—effectively biasing participants toward different goals—can adjust the evidence criterion without affecting the evidence accumulation rate, $r$. In our study, since experimental trials are fully randomized and participants are thoroughly familiarized with the study protocol and stimuli before data collection, we can reasonably assume that each individual's evidence criterion remains constant throughout the experiment.

**Evidence accumulation rate.** Palmer et al. [2011] also demonstrated that changes in the difficulty of visual content processing influence the evidence accumulation rate, $r$. Indeed, our study results (see Section 5.1 for details) reveal that visual characteristics affect the distribution of saccade latencies and exhibit a complex, non-monotonic relationship with stimulus processing difficulty.

These findings motivate us to model the evidence accumulation rate, $r$, as a function of the contrast ($c$), spatial frequency ($f$), and eccentricity ($e$) of the target stimulus—key

visual features that significantly impact perception. To summarize, we model saccadic decision-making latency, $T_{sac}$, using a DDM in which the evidence criterion, $\alpha$, is fixed based on the visual task specifications, while the evidence accumulation rate, $r$, is modeled as a function of the target stimulus's visual properties:

$$T_{sac} \sim \mathcal{IG}(\alpha, r(c, f, e)). \tag{5.1}$$

In our work, the relationship between $r$ and the visual characteristics of the stimulus (i.e., the tuple $c, f, e$) is modeled using a Radial Basis Function:

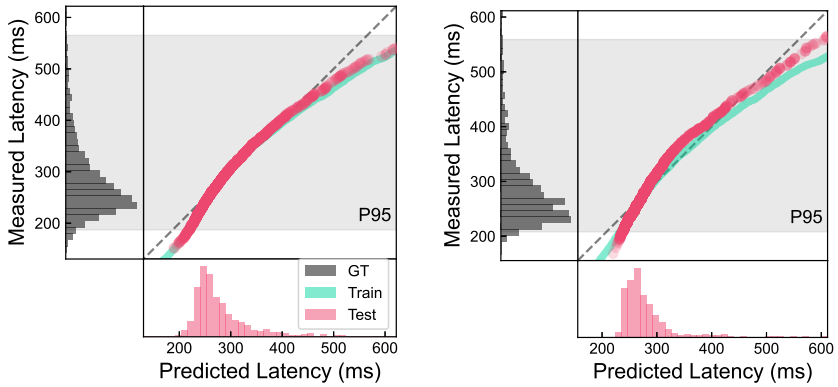
$$r(c, f, e) = \sum_{i=0}^{N} \lambda_i \rho \left( \left\| \begin{bmatrix} c & f & e \end{bmatrix} - \mathbf{b}_i \right\|, \sigma_i \right), \tag{5.2}$$

where $\mathbf{b}_i$ represents the individual radial basis centers, and $\rho$ is a Gaussian basis function. In our experiments, we set $N = 20$. Using the data collected in Section 5.1, we jointly fit the RBF parameters $\lambda$, $\mathbf{b}_i$, and $\sigma$, along with the evidence criterion $\alpha$, via gradient descent. Since baseline reaction times vary across individuals due to inherent differences, we first normalized each participant's reaction time data, ensuring that reaction times in the $c, f, e = 1, 1, 0$ condition were equalized across subjects before aggregating responses.

### 5.2.2 Numerical Evaluation

In this section we evaluate the model's performance by analyzing the model's robustness to alternate data fitting and testing partitions of the dataset.

**Protocol.** For each analysis, we reserve a different partition of the dataset from Section 5.1 for testing, and fit the model using the remaining data. We perform two types of partitioning protocols for reserving the test set:




(a) random partition (b) subject partition

Figure 5.4: *Model performance and generalization validation using preliminary user study dataset.* Histogram alignment between ground truth (gray) and model predictions are compared via Q-Q plots for different train/test splits in red/green respectively. The closer the Q-Q curves are to the diagonal, the more accurate the predictions are. P95 data volume intervals are highlighted in gray. As defined in Section 5.2.2, (a) and (b) show the results with random partition and `subject_01`'s data partition, respectively.

1. Random: a random selection drawn from all data points (20%),

2. Subject: all data from each individual subject (20%).

**Metrics and results.** We perform the Kolmogorov–Smirnov (K.S.) goodness-of-fit test between the reserved test data and our prediction [Massey Jr, 1951], and show the Quantile-Quantile (Q-Q) plot [Gnanadesikan and Wilk, 1968] in Figure 5.4. The Q-Q plot visualizes the correspondence of two probability distributions at each quantile. Data below the $y = x$ line in Figure 5.4 indicate an overestimation of saccade latencies and vice versa for data above the line.

Figure 5.4 shows the Q-Q plot for the training and testing sets across both partition protocols. The K.S. test fails to reject the null hypothesis that the observed user saccade latency distribution is drawn from our model-predicted distribution for (1) the random partition, $D = .2, p = .99$, and (2) the individual subject partitions:

| Subject ID | S1 | S2 | S3 | S4 | S5 |
|---|---|---|---|---|---|
| K.S. analysis | $D = .3$ | $D = .2$ | $D = .2$ | $D = .2$ | $D = .1$ |
| | $p = .79$ | $p = .99$ | $p = .99$ | $p = .99$ | $p = 1.0$ |

where $D$ is the K.S. Test statistic, and $p$ is significance value.

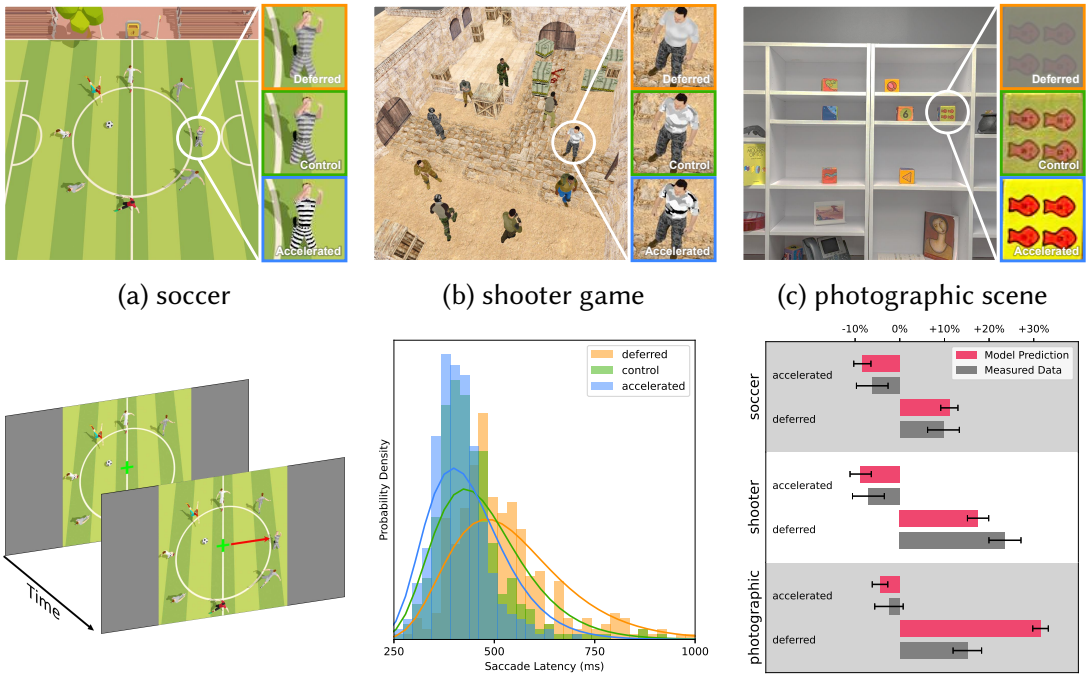
**Discussion.** The above analysis demonstrates that our model does not predict statistically different distributions compared to unseen observations across various partitioning protocols. The results of the randomly partitioned study (1) demonstrate the generalizability of our model across trials without observed overfitting. Analysis of the subject-partitioned study (2) verifies our model's applicability to unseen users, and thus general human saccadic behaviors.

### 5.2.3 Predicting Saccadic Behaviors in Natural Tasks

In Section 5.1, we observed that unnoticeably subtle visual changes may induce significantly varied reactive latencies, as was formulated and predicted by our model in Section 5.2. In this experiment, we evaluate our model's application in several realistic target search scenarios such as esports, and real-world photographs.

Via a series of psychophysical studies, we seek to determine: (1) whether our model can extend to predicting saccadic reaction latencies with natural task/stimuli; (2) whether we can imperceptibly alter the appearance of objects while still introducing enough reactive latency to materially influence real-world task performance. We answer these queries in our experiment and compare our findings to the model predictions.

**Participants and setup.** We recruited 14 participants (ages $22 - 33$, 3 female) for this series of 2AFC experiments. Two participants were excluded for inability to perform




(a) soccer (b) shooter game (c) photographic scene

(d) procedure (e) histogram comparison (f) predictions vs data

Figure 5.5: *Setup and Results of the Natural Task Evaluation.* Saccade latency modulation correlates with the contrast of stimuli as shown in the three distinct scenes (and target candidates) shown in (a)/(b)/(c). Each scene presents distinct visual characteristics including low polygon 3D scenes, dense geometries, or natural scenes. (d) illustrates the study procedure over time. With the **Control** condition as reference, all others show FovVideoVDP scores above 9.5, indicating identical perceptual appearance per [Mantiuk et al., 2021]. Using the shooter scene as example, (e) shows the user latency data in histograms, and our model predicted latency in curves. A significant agreement can be observed. Please refer to our supplementary videos for an animated visualization. (f) shows the mean relative durations (with **Control** as "0%" pedestal) of **Deferred**/**Accelerated**. The error bars indicate SEM. Full statistical analysis on all scenes can be seen in Section 5.2.3. Each individual's raw probabilistic distributions are provided in Section 5.B. 3D asset credits: haykel-shaba (a), and Slavyer (b) at Sketchfab Inc.

the tasks (self-reported difficulty perceiving peripheral stimuli and target identification accuracy greater than one standard deviation below the mean), resulting in 12 final participants. Two of the 12 participated in the preliminary study in Section 5.1. The study was conducted during a 10-minute sessions consisting of 153 trials per scene for each participant. The hardware and setup remain the same as in Section 5.1.

**Scenes and stimuli.** To simulate a broad range of applications, our user study stimuli consisted of three groups of images: (1) a synthetic soccer scene, (2) a synthetic first-person view as an analog for esports, and (3) digital photographs of an indoor shelf. Each group contained 51 different images; each has the target stimuli appearing at different locations (to avoid learning effects) on the visual field, and serve as a trial. The background and targets from each evaluation group are shown in Figure 5.5. Although shown in color in the manuscript for visual clarity, all images were rendered with grayscale on display to avoid bias from color cues.

**Tasks.** Participants were instructed to complete a similar 2AFC decision task across all trial. At the beginning of each trial, they were shown a background image containing several task-irrelevant objects. After a randomized $1 - 1.5$ second delay, an additional task-relevant stimulus, either a *target* or *non-target*, appeared on the scene as in Figure 5.5. Participants were shown both types of stimuli ahead of the experiment. The task was to saccade to targets, or remain fixated if the stimulus was identified as a non-target. This procedure allows us to measure the visual-oculomotor latency after which a subject identifies the discernible feature of interest from the stimulus. This emulates the common real-world scenarios where a new "intruder" of potential interest enters the subjects' visual fields. Please refer to our video for dynamic illustrations of the task.

**Conditions.** Across each image set, we tested three variations of the target stimulus in order to measure how changes in image features affect saccade latencies. In one variation the target stimulus had increased contrast and/or decreased frequency (**Accelerated**), in another variation the target had decreased contrast and/or increased frequency (**Deferred**), and a third unfiltered variation was used as a control group (**Control**). Each participant performed 51 images $\times$ 3 conditions $\times$ 3 scenes, resulting in 459 trials total, i.e., 5508 trials across the experiment. Measuring the precise frequencies affecting saccade latency is a complex task requiring pooling from multiband. Investigating a comprehensive pooling strategy is beyond the scope of this work. Therefore, we approximate the representative frequency as the Laplacian pyramid layer with the highest corresponding contrast, for those images without a uniform frequency pattern. Contrast and eccentricity computations were trivial to compute without requiring pooling operations.

**Results.** We present the results of our experiments in Figure 5.5. We again use the K.S. statistical test to evaluate alignment between predicted and measured histograms across the different scenes for each condition. We report the results of these tests below:

| | Deferred | Control | Accelerated |
|---|---|---|---|
| Soccer | $D = .2, p = .99$ | $D = .2, p = .99$ | $D = .2, p = .99$ |
| Shooter | $D = .2, p = .99$ | $D = .3, p = .79$ | $D = .1, p = 1.0$ |
| Photographic | $D = .3, p = .79$ | $D = .1, p = 1.0$ | $D = .2, p = 1.0.$ |

Please refer to Section 5.B for the collected saccadic latency distributions of individual participants and scenes.

Using the **Control** images as reference, we additionally calculate the FovVideoVDP values for all images in our dataset. We find the mean values to be above 9.5 for all

**Accelerated**/**Deferred** images, which indicates observers would be approximately at chance for detecting differences between them.

We also debriefed each participant after the experiment on their thoughts regarding the tasks, and most participants reported no self-awareness of reaction time difference.

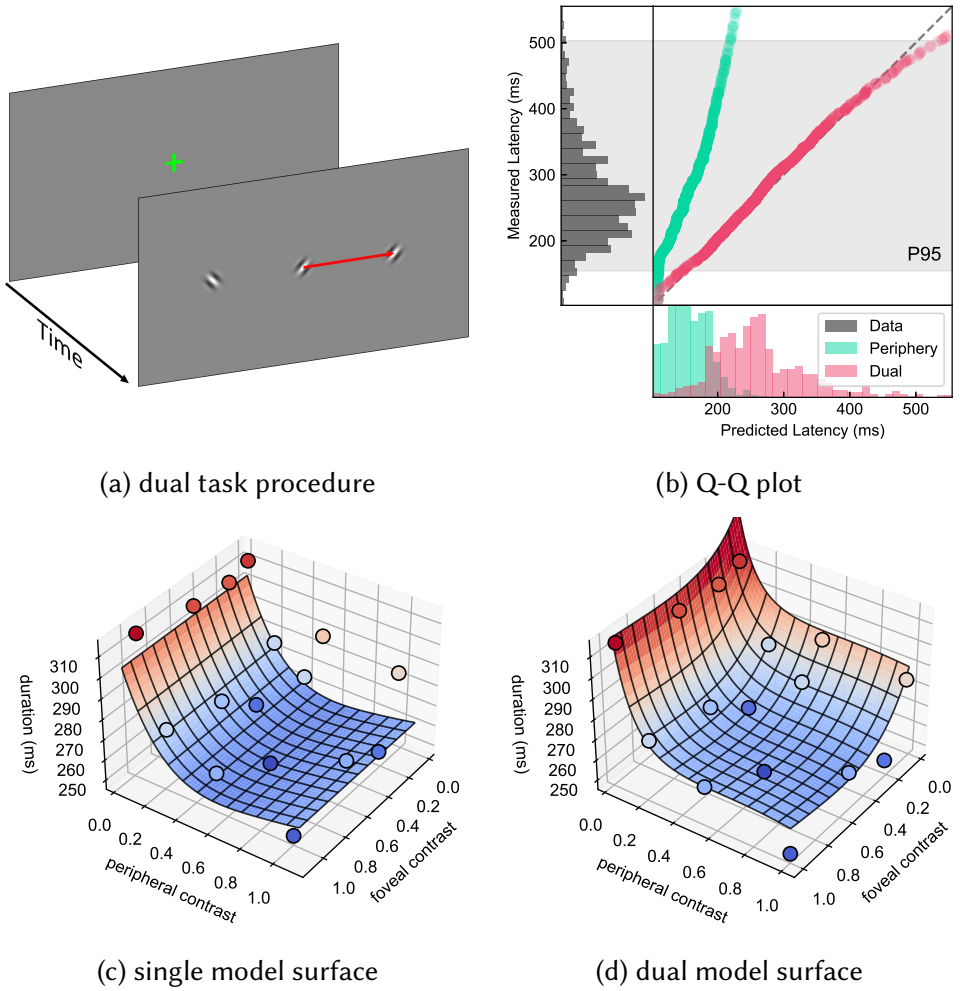
**Discussion.** Our results demonstrate agreement between the predictions made by our model and the observed saccadic latency distributions across 12 participants. We find significant differences in saccadic latency across conditions, despite identical perceptual appearance evidenced by the FovVideoVDP metrics.

Our prediction of the photographic scene results show correct trends and distribution ratios, albeit for a scaled absolute time (in ms) relative to the measured data. We attribute this scale variance to the fact that natural images contain wide frequency bands and our single-frequency pooling in the Laplacian Pyramid may discard significant frequency information. This motivates interesting future work on multi-frequency pooling models tailored for reaction time, see Section 5.5.

### 5.2.4 Predicting Foveal-Peripheral Dual Task Behavior

In various real-world scenarios, humans perform tasks by jointly analyzing both foveal and peripheral content, such as with reading, film watching, and architectural design. In this experiment, we extend and evaluate our model to such applications considering *dual* tasks.

**Modeling.** Our visual system processes foveal and peripheral stimuli independently and in parallel for a variety of tasks [Ludwig et al., 2014a]. That is, the foveal and peripheral pathways gather information concurrently, and the decision to trigger a




(a) dual task procedure

(b) Q-Q plot

(c) single model surface

(d) dual model surface

Figure 5.6: *Setup and results of the foveal-peripheral dual task evaluation.* (a) A dual foveal-peripheral task consists of two components: identification of both the foveal and peripheral Gabor patches. The subject was instructed to move their gaze to the peripheral patch with matching orientation to the foveal one. Please refer to our supplementary videos for an animated visualization. (c) We fit our periphery-only model ($T_p$, the surface) to data from the foveal-peripheral dual task (the sparse dots). A significant mis-alignment can be observed. (d) Considering maximum expected latency of both foveal and peripheral contrasts enables us to predict a more accurate relationship, $T_{dual}$, between the visual stimulus parameters and the observed saccade latency data. (b) Q-Q plot visualizing the goodness-of-fit of our model relative to the observed data. Alignment of the observed and predicted latency histograms shows that the dual model matches well with the experimental data (gray) within the P95 confidence interval (highlighted region). In contrast, the peripheral-only model fails to correctly predict saccade latencies. We omit visualizations of the foveal-only model $T_f$ to avoid duplication as it exhibits similar low performance in predictive quality to $T_p$. The full statistical analysis can be seen in Section 5.2.4.

saccade waits until both processes have finished. We hypothesize that these independent foveal and peripheral stimulus processing units operate using the *integration-and-action* process as described in Sections 2.2.2 and 5.2.

In this model, processing times for both the fovea, $T_f$, and periphery, $T_p$, follow Equation (2.6), and can be adapted to specific visual tasks and stimulus visual features as shown in Equation (5.1):

$$\begin{aligned} T_f &\sim \mathcal{IG}(\alpha_f, r_f) \\ T_p &\sim \mathcal{IG}(\alpha_p, r_p), \end{aligned} \tag{5.3}$$

where we create some shorthands for convenience:

$$\begin{aligned} r_f &= r(c_f, f_f, e_f = 0^\circ) \\ r_p &= r(c_p, f_p, e_p = 10^\circ). \end{aligned} \tag{5.4}$$

$e_p = 10^\circ$ because the peripheral stimulus for this experiment was at $10^\circ$ eccentricity. Then, as experimentally determined by prior literature on similar tasks [Ludwig et al., 2014a], we model the total saccade latency as the maximum value of these two random variables:

$$T_{dual} = \max(T_f, T_p). \tag{5.5}$$

**Setup.** To evaluate our hypothetical model for dual tasks, we conducted a user study to measure how saccade latency changes as we modulate foveal and peripheral stimuli independently. Unfortunately, it is not possible to individually determine the $\alpha_f$ and $\alpha_p$ values, because a user study for the dual task can only sample the *total* saccade latency from Equation (5.5). That is, the individual distributions, $T_f$, and $T_p$ are not measured

directly. Since finding these threshold values directly is not possible, we infer them via maximum-likelihood estimation (MLE) of the overall distribution of $T_{dual}$, given a dataset of size $n$:

$$\alpha_f, \alpha_p = \arg\max \sum_{i}^{n} \log L(\alpha_f, \alpha_p; t^{(i)}, \nu_f, \nu_p). \tag{5.6}$$

Please refer to Section 5.A for the derivation of the likelihood function for $T_{dual}$. The hardware setup in this experiment is the same as described in Table 5.1.

**Participants.** We recruited $n = 12$ participants (ages 22-33, 3 female) with normal or corrected to normal vision for a series of 2AFC experiments. The study was conducted during a single 10 minute session, including a total of 240 trials for each participant.

**Stimuli and Tasks.** At the beginning of each trial participants are shown three Gabor patches as illustrated in Figure 5.6a: one at the fovea, and two in the left and right peripheries at equal eccentricities of 10°. The foveal Gabor is tilted either 45° or −45° from the vertical axis; with chance probability, one of the peripheral Gabors is selected to have the same tilt as the foveal Gabor, while the other has the opposite tilt. The task is to identify and saccade to the peripheral Gabor of the same orientation as the foveal Gabor. For each trial, the central and peripheral Gabor contrast values are sampled from $[0.05, 0.22, 0.53, 1.0]$, drawn independently. That is, taking all combinations of central-peripheral Gabor contrast possibilities yields a total of 16 conditions. The frequency of all Gabors was fixed to 2.0 cpd (cycles-per-degree). Each participant also performed 15 randomly ordered practice trials before the start of the experiment.

**Results.** In Figure 5.6d, we show the relationship of both foveal and peripheral contrasts with saccade latency, as well as the ground truth data collected from our user study overlaid on top of the surface plot. The MLE regression produces threshold values of $\alpha_f = 3.21$ and $\alpha_p = 3.56$. Hence, the threshold ratio between the foveal and peripheral components is $1 : 1.04$. Similar to Section 5.2.2, we present the Q-Q plot comparing the data to our model predictions in Figure 5.6b. The K.S. statistical test again fails to reject the null hypothesis that the observed user saccadic latencies are drawn from our $T_{dual}-$predicted distribution ($D = 0.1$ and $p = 1.0$).

Models which consider only the peripheral contrast (shown in Figure 5.6c), or only the foveal contrast fail to accurately predict the saccade latencies. We run the K.S. test for both of these conditions and observe a significant difference between the data and the model predictions: $D = 0.9$ and $p = 0.002$ for the foveal-only model, and $D = 0.8$ and $p = 0.002$ for the periphery-only model.

**Discussion.** When humans perform tasks involving both foveal and peripheral analysis, we observe that a models considering only one eccentricity fails to predict saccade latencies, as illustrated in Figure 5.6b and demonstrated by the K.S. tests. By comparison, the joint model we propose in Equation (5.5), inspired by prior discoveries on visual mechanisms, successfully predicts the latency distribution.

### 5.2.5 Application Case Study: Esports Fairness Metric and Performance Optimization

A major application of our model is to measure and optimize human performance in competitive, real-time, or time-sensitive tasks such as defense, piloting, and esports. In this evaluation, we use esports as an example. In real-world professional gameplay, we

deploy our model to: 1) measure game fairness in terms of character skin design that triggers varied gaze motion performance between two teams; 2) measure and optimize the human target search performance under various screen resolutions, eye-display distances, and compare the performance with traditional and immersive displays.

**Data.** We collected professional replay videos from a popular esports game, Counter-Strike: Global Offense via YouTube. The data contains a ≈ half hour long video footage where we uniformly sampled 95 frames from beginning to the end. For each frame, we exploit the virtual human tracking model YoLO [Redmon et al., 2016] that predicts the team ID (Counter-Terrorist, **CT** and Terrorist **TR**), and bounding boxes. We assume the observers gaze lies in the middle of the screen, and apply our model to predict the time when the viewer reorients their gaze to each target. We measure the visual characteristics with a common display setting: a Samsung 32inch CH32H711 monitor, 2K 16:9 resolution, 70cm width, 300cd/$m^2$ brightness, and ≈ 1.33D (diopter = $\mathrm{m}^{-1}$) eye-display distance (50° FoV).

**Competition Fairness in Target Searching.** The game has two opposing teams of characters. Regardless of game task design and differences in tools, game fairness is an important concern in esports [Chen et al., 2014]. Using our model and the detected targets, we measure the average saccade latency of individual groups.

Figure 5.7b shows the results. We observed a significant difference between **CT** and **TR** groups: the average normalized latencies are 0.92 ± 0.02 for searching **CT**s and 0.95 ± 0.04 for searching **TR**s, indicating a 3.3% difference. Given previous literature indicating the mean saccade latency for CS:GO professional players to be about 282ms [Velichkovsky et al., 2019], 3.3% results in a 9.3ms reaction variance. One-way repeated measures ANOVA showed the group's significant main effect on the saccadic latency,

$F_{1,93} = 11.4, p = .001$.

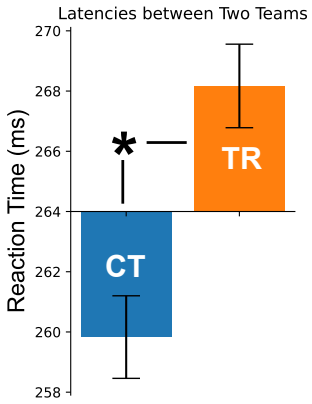
The results demonstrate a statistically significant difference between the two groups, in terms of them perceiving, processing, and reacting to appeared targets. That is, a **TR**/**CT** saccading to the other group is significantly faster/slower with no less than 2/1 frames on a 60/120FPS displays. The speed difference is remarkably higher than the minimum latency, as low as 4ms, that leads to altered esport performance among top-level competitors [Kim et al., 2019]. While this may have been one of the factors that contributed to the imbalanced competitive game performance (higher winning rate of **TR** on the map we analyzed) between these two groups [1], in practice the asymmetric weapon and task designs might also have played a role.

**Optimizing Player Performance.** A natural and extensively asked question is the role of eye-display distance (e.g., regular monitors vs. VR displays) and screen resolution in professional competitions. Using our model, we measure the statistical saccade latency as a function of displays with the same dataset as Section 5.2.5.

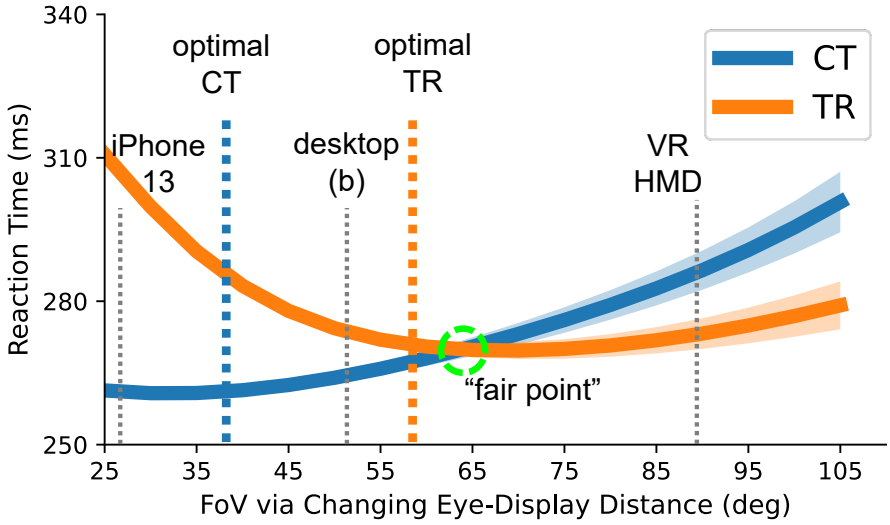
Figure 5.7c visualizes our results by observing the altered reaction performance. As before, we use the mean saccade latency for CS:GO professional players to be about 282ms [Velichkovsky et al., 2019]. First, both teams are not at their best performance with the initial 1.33D eye-display distance, with the faster reaction of **TR**s searching **CT**s. However, the teams reveal different trends by changing the eye's distance (thus FoV), which jointly alters target eccentricity ($e$) and frequency ($f$) (cf. Section 5.B). Particularly, the minimal saccade latency towards **CT** targets is 273.5ms at 34.6° FoV (0.9D eye-display distance). In comparison, the minimal latency towards **TR** targets is 266.1ms at 61.5° FoV (1.6D eye-display distance). The two curves intersect at 1.69D with

[1] `https://www.hltv.org/stats/teams/map/31/5995/g2`


(a) data




(b) team-wise performance




(c) performance w.r.t. varied FoV

Figure 5.7: *Results of esports video dataset analysis.* (a) illustrates our simulated eye-display spatial relationship and our CS:GO gameplay dataset (including the automated labeling of the teams). Note that changing the eye-display distances results in varied fovs, thus changing the perceived visual characteristics (eccentricity and frequency). (b) shows our model's approximation of the team-wise target searching performance. The X-axis indicates the team splits. The Y-axis shows the mean saccadic latency calculated by our model (with the annotated team as the target team being searched, i.e., the "opposite team"). The error bars show the standard error. (c) shows our analysis simulating various fovs by altering eye-display distances. The $x$-axis indicates the fovs in degrees. The $y$-axis shows the predicted mean latencies with the semi-transparent error bar as standard error. The point where the two group's mean latencies intersect is marked by the green circle. The lowest latencies of saccading for each team and the simulated fovs of non-desktop display environments are dash-labeled.

an identical latency of 266.3ms. We further simulate real-world use cases with different displays (for instance, gaming with mobile devices or training with VR displays). In this experiment, we use the measures from the iPhone 13 ($5.78 \times 2.53$ inches) with the commonly suggested 30cm (or 3.3D) eye-display distance, leading to a 25.7° FoV. Under this circumstances, the saccade time to **TR** becomes higher than **CT** (324.3ms vs 259.4ms) Similarly, the measurement with our virtual reality HMD (90° overlapped FoV), the relative trend is swapped: saccading to a **TR** becomes shorter than to a **CT**, with a 21.2ms difference (276.0ms vs 288.0ms).

The above analysis indicates the sensitivity of eye display correlation in determining performance and fairness in time-sensitive and competitive scenarios. Surprisingly, the statistical performance bias may swap with different eye-display relationships. For instance, with a mobile/VR setting, the visual stimuli may bias with **TR**/**CT** players in terms of reaction performance. In addition to the commonly referred measurement of visual similarity and task/map fairness, our model presents a novel perspective in competitive and highly dynamic scenario design, such as athletics, esports and defense.

## 5.3 Measuring Moving-Target Tracking Latency

In this second experiment, we examine decision-making latency in the context of visually tracking dynamically moving stimuli. Specifically, we manipulate target visibility by adjusting luminance and color contrasts, introducing external noise, and analyzing the temporal response to visual tracking behavior. Notably, in Section 5.1.2, we briefly discussed the possibility that contrast sensitivity-rescaled measures of visual signal strength could drive common trends in behavioral responses. That is, regardless of which visual features are modulated, expressing their visibility in terms of detection

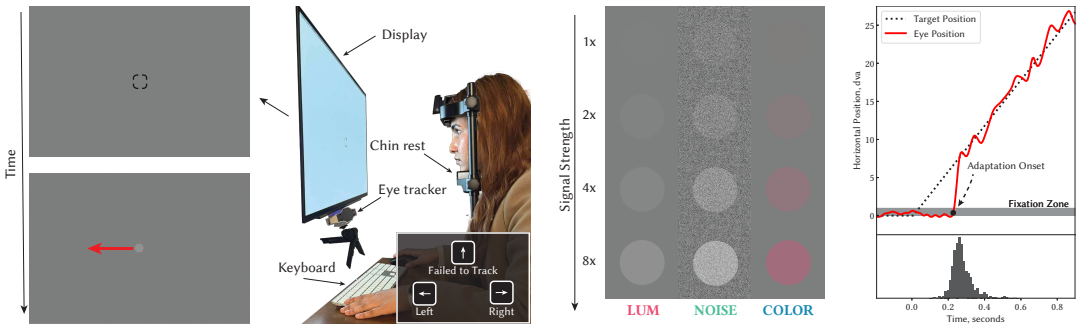
threshold contrasts (as done in many human vision computational models [Legge and Foley, 1980; Watson and Solomon, 1997]) may lead to similar behavioral performance patterns. By analyzing how changes in target visibility across different axes of image manipulation affect performance, we identify common behavioral trends and ultimately develop a computational model for predicting human performance in tracking dynamic visual content.

### 5.3.1 Experimental Design

**Participants.** Five participants (ages 22 - 28, 3 male) with normal or corrected-to-normal vision were recruited for a series of three psychophysical experiments. Each experiment was conducted during a separate session, and each session consisted of two blocks of 120 visual target tracking tasks each. Each experiment took a total of about 40 minutes to complete. All experimental protocols were approved by an institutional review board (IRB).

**Setup.** Experimental stimuli were displayed on a 27.5 inch OLED monitor (LG 27GR-95QE) with a refresh rate of 240 Hz, a spatial resolution of $2560 \times 1440$ pixels, and subtended a horizontal field of view of $55°$. As shown in Figure 5.8a, participants seated throughout the experiments, and their head positions were stabilized with a chin rest. Eye position signals were recorded using a 150 Hz eye tracker (GazePoint 3), and was calibrated before each block of trials.

**Stimuli and design.** Participants were instructed to visually track targets that appeared and moved across the screen. As illustrated in Figure 5.8b, targets were solid disks with a diameter of .5 dva (degrees of visual angle) placed on top of a neutral gray




(a) protocol (b) stimuli (c) example gaze trajectory

Figure 5.8: *Experimental protocol, stimuli, and data.* (a) Participants visually tracked target stimuli that appeared on the display, and indicated via button press its direction of movement. Eye position traces were recorded using an eye tracker. (b) Target visibility was modulated by adjusting different visual features across the **LUM**, **NOISE**, and **COLOR** experiments. Stimuli with signal strength = 1× were calibrated to be at threshold of visibility, and scaled up by signal strength. (c) Visuo-motor adaptation onset distributions were determined by detecting the latest saccades that moved away from the fixation zone during each trial.

background (with measured mean luminance of 54 cd/m$^2$), and moved at a constant speed. The visibility and speed of the targets varied across trials and constituted the different conditions across which we compared the tracking performance. We leverage the eye-tracking data collected as participants visually tracked the targets to quantify a measure of performance as further detailed in the *Analysis* paragraph.

One of the main factors that affect the performance of visually tracking moving targets is the speed of the target [Spering et al., 2005]. In our experiments, we varied the target speeds $v_t \in \{7.5, 15, 30\}$ dva/s.

Additionally, as discussed in Section 2.3.4, the visibility of visual targets significantly affects visuo-motor adaptation performance. Therefore, across our three experiments, we modulated target visibility within three different features. Specifically, we studied how performance is affected by luminance contrast in low external noise (**LUM** experiment), and high external noise (**NOISE** experiment), as well as by color contrast

(**COLOR** experiment).

In the **LUM** and **NOISE** experiments, the luminance contrast of the target varied between conditions with zero background noise in **LUM**, and a uniform additive noise with RMS contrast of 23% added to the background in **NOISE**. In the **COLOR** experiment, the color of the disk varied along the *RG* axis in *DKL* coordinates [Derrington et al., 1984b], while the luminance was kept equal to the background.

Across the experiments, we parameterize target *visibility* in detection threshold contrast units, and refer to this parameter as the *signal strength*, $s$. That is, for a target stimulus with a contrast value of $c$, its corresponding signal strength equals $s = c/c_{th}$, where $c_{th}$ is the just detectable threshold contrast (i.e., inverse of the sensitivity) of the target stimulus. In **LUM** and **NOISE**, contrast was quantified using Michelson contrast, whereas the *DKL* color contrast, as introduced in Section 3.1, is used in the **COLOR**. Even though the specific contrast values of the targets vary significantly across experiments, the generalized *signal strength* parameterization allows us to validate our hypothesis that behavioral performance during visuo-motor adaptation demonstrates similar first-order patterns, irrespective of the specific visual features being manipulated.

Since we require the contrast sensitivity thresholds calibrated for each individual in each experimental task (see further discussion in Section 5.5), prior to each experiment, the sensitivity in each task was determined via a 4-down-1-up adaptive staircase procedure. The calibration featured the same disk from its corresponding main experiment, moving either leftward or rightward at a speed of $v_t = 30$ dva/s. Participants progressed through the staircase by visually tracking the target and indicating the direction of the disk's motion via button press. Staircase steps were incremented in reciprocal contrast units with a step size = 5 for six reversals, and the average of the last three reversals were used to determine the sensitivity.

Based on the calibrated contrast threshold, the *signal strength* of the main experiments were determined. As illustrated in Figure 5.8b, across the experiments, signal strength values of $s \in \{1, 2, 4, 8\}$ were used. Overall, the target's 3 speed conditions $\times 4$ signal strength conditions were repeated 10 times per block for a total of 120 trials.

**Procedure.** Each trial began with a button press. As shown in Figure 5.8a, following a randomized stimulus onset delay of $0 - 1$ s, the target stimulus replaced the fixation crosshair and moved either leftward or rightward across the screen at a constant speed determined by the trial condition. Randomizing the stimulus onset delay and target movement direction minimized anticipatory eye movements. The target remained in motion for $1 - 1.5$ s before disappearing. At the end of each trial, participants indicated via button press whether the target object moved leftward, rightward, or if they failed to track it. Trials in which participants failed to track the target or responded incorrectly were repeated at the end of each block.

**Analysis.** As motivated in Section 2.3.4, our study focuses on the initial latency for visuo-motor adaptation, as delays in foveating and tracking a visual target can cause missed visual details and poor task performance. Visuo-motor adaptation is primarily driven by two concurrent neural mechanisms triggered at stimulus onset: pursuit eye movement control and a corrective saccadic eye movement [Nachmani et al., 2020]. Pursuit eye movement is a feedback system which is tuned to synchronize the movements of the eye and the target so that a foveated moving target remains foveated throughout the tracking process [Robinson, 1965]. However, the onset of pursuit eye movement takes at least .1 s, and even longer to fully match the target's motion following an abrupt change in the target speed [Missal and Heinen, 2017]. During such delays, the positional error between the target and eye positions accumulates and necessitates

a saccade that corrects this positional error and ultimately completes the adaptation process [Nachmani et al., 2020]. In our analysis, we leverage the onset of this corrective saccade as an indicative marker to quantify visuo-motor adaptation latency.

To detect the corrective saccade during the onset of visuo-motor adaptation, we utilize the recorded eye position traces. Eye position traces were smoothed by a Butterworth filter with a 20 Hz cut-off prior to analysis. Trials where participants failed to maintain fixation at the beginning of the trial, and failed to track the target were excluded from analysis, and constituted 2.9% of all trials.

Specifically, we implement an objective method for determining whether the eye position recordings are usable in further analysis. First, we segment the eye position traces of each trial, relative to stimulus onset time, $t$, into an initial fixation phase ($-0.2 \leq t \leq 0.1$ s) [Becker and Jürgens, 1979; De Brouwer et al., 2002], a visuo-motor adaptation onset phase ($0.1 \leq t \leq 0.6$ s) [De Brouwer et al., 2002], and a steady-state tracking phase ($0.6 \leq t \leq 0.9$ s), [De Brouwer et al., 2002; Spering et al., 2005].

Using the eye position recordings from each segment, we require that the median fixation position be within 2 dva of the central fixation positions, and the spread of eye positions, measured in standard deviations of the gaze distribution, to be less than $\sigma < 1$ dva. In the application studies of Sections 5.4.3 to 5.4.5, the fixation phase tolerances were doubled due to larger tracker errors observed due to the more complex visual stimuli presented throughout the trials. During the steady-state tracking phase, the median offset of the eye position recording relative to the target stimulus was required to be within 4 dva.

Figure 5.8c shows an example trace of target motion to the participant's gaze trajectory. Eye velocity was calculated using the central difference method and saccades were detected using a fixed velocity threshold [Gibaldi and Sabatini, 2021]. Velocity

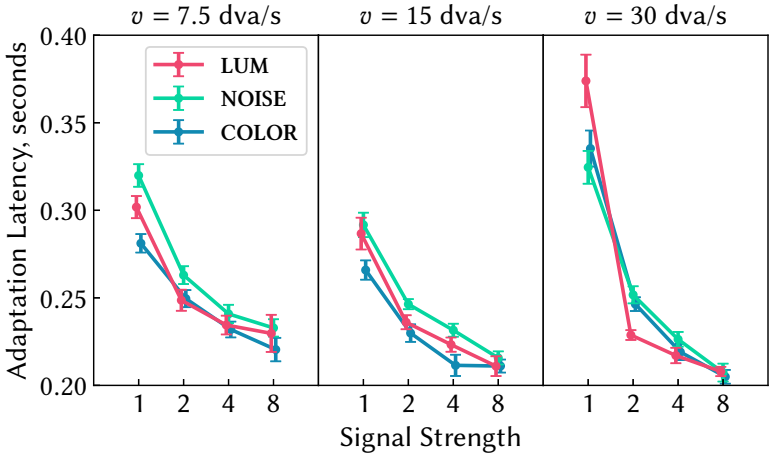



Figure 5.9: *Adaptation performance of a representative participant.* Mean adaptation latencies are plotted against signal strength, and compared across targets moving at varying speeds. Each experiment is represented by a distinct color (see Figure 5.8b for the corresponding experimental stimuli). Error bars denote SEM. See Figure 5.10 for plots of the remaining participants.

cut-off criteria of 12.5, 25, and 50 dva/s were applied for stimuli moving at 7.5, 15, and 30 dva/s, respectively. The *main* corrective saccade was identified using a positional criterion that selected the latest saccade which shifted the eye position away from the initial fixation distribution by more than two standard deviations towards the target. The initial fixation distribution was established by aggregating the eye position data within the $-0.2 < t < 0.1$ s window relative to stimulus onset. Saccades detected during $t < 0.1$ s were excluded, as such movements could not have been programmed in response to the stimulus onset [Becker and Jürgens, 1979; De Brouwer et al., 2002].

### 5.3.2 Results and Discussion

**Results.** We present our main findings for a representative participant's experimental data across the 12 conditions of each experiment in Figure 5.9. Adaptation latencies improve by up to 0.166/0.117/0.130 s in **LUM**/**NOISE**/**COLOR** respectively, based on both the target speed as well as signal strength. Increasing signal strength provides

diminishing benefits until performance nears its peak at a mean latency of .216/.218/.212 s across the experiments. Notably, at high signal strength, faster targets are adapted to faster, while at low signal strength, the opposite holds.

Two-way ANOVA tests revealed a significant main effect of target speed ($p < 0.01$) and signal strength ($p < 0.001$) on adaptation latency across all experiments, indicating that target speed and signal strength strongly influence adaptation latency. Additionally, there was a significant interaction between target speed and signal strength across all experiments ($p < 0.001$), suggesting that the effect of object speed on adaptation latency depends on the level of signal strength.

The Spearman correlation of mean adaptation latencies found between **LUM** & **NOISE**, **LUM** & **COLOR**, and **NOISE** & **COLOR** experiments were $\rho = 0.96$ ($p < 0.001$), $\rho = 0.94$ ($p < 0.001$), and $\rho = 0.99$ ($p < 0.001$) respectively, indicating a very strong positive monotonic relationship. These results suggest that the adaptation latencies measured in each experiment are highly correlated to each other with a high degree of consistency for the individual.

The remaining four participants demonstrate similar trends (Figure 5.10) and statistical significances across conditions and experiments. All the 15 experimental datasets exhibit a significant signal strength effect on latency. Among them, 12 exhibit significant target speed effects on latency, and 12 exhibit a significant interaction.

**Discussion.** The ANOVA results of our three experiments indicate that both target speed and signal strength significantly influence visuo-motor adaptation latencies. Thus, these factors should be considered independently, as evidenced by the significant interaction between them. As shown in Figure 5.9, increasing signal strength accelerates adaptation latencies. These results broadly match our results from Section 5.1.

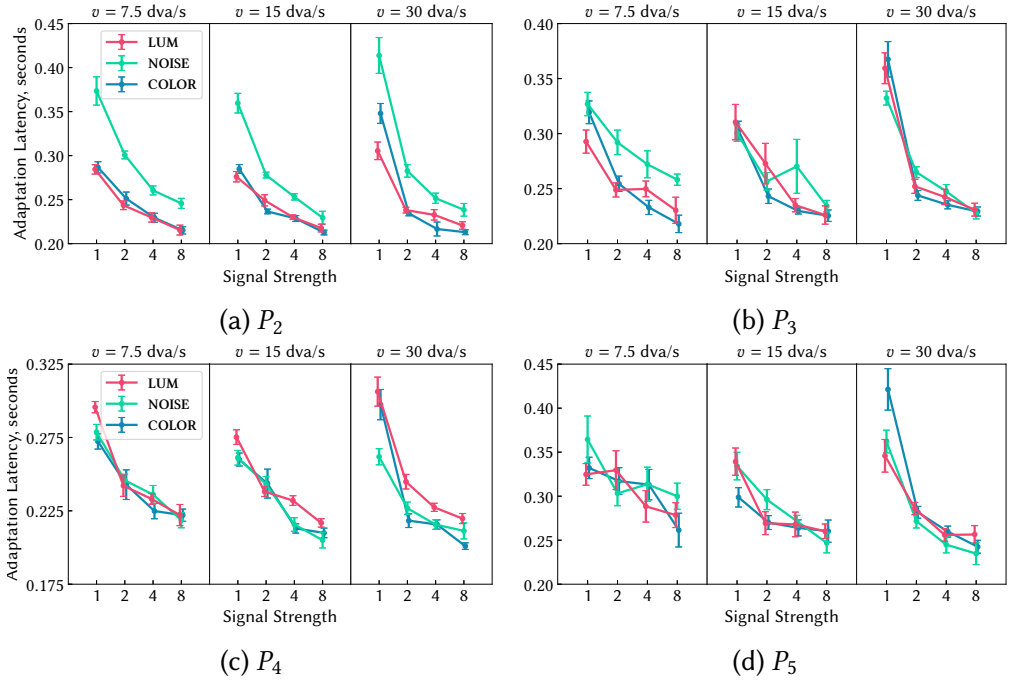



Figure 5.10: *Adaptation performance of main study participants.* Mean adaptation latencies of main study participants are visualized in the same style as in Figure 5.9.

Crucially, our results show that apart from target visibility, its speed also affects adaptation performance. Faster speeds impair adaptation latencies at low signal strength conditions, suggesting that the target tracking task becomes more challenging. We hypothesize that this decline in performance occurs because faster-moving low visibility targets exit the fovea and enter peripheral vision more quickly, where visual acuity is reduced, and target localization becomes more challenging. However, the effect of this performance deterioration might not apply when the signal strength is high and the target can be localized easily even in the periphery.

Our analysis of the correlation in adaptation latency across different visual features modulated by their signal strength suggests that the underlying neural mechanisms governing the onset of visuo-motor adaptation rely on a unified signal encoding the overall visibility of visual targets. While further investigation is needed to validate

this hypothesis, these distinctions are less critical in the context of computer grahpics applications. The presence of such first-order effects is sufficient to develop computational models for downstream applications. In the next section, we will integrate all experimental data collected in our studies to construct a human behavioral model for predicting visuo-motor adaptation latency.

## 5.4 Behavioral Model of Moving-Target Tracking Latency

### 5.4.1 Model Specification

As discussed in Section 5.3.2, increasing *signal strength* yields diminishing performance benefits, in agreement with prior work indicating that adaptation latency plateaus and performance saturates beyond a certain ceiling [Spering et al., 2005]. Just as in Section 5.2, we model the onset of the catch-up saccade during the visual-tracking task via the DDM. Notably, the trend of reaction latencies appear to be consistent across different visual feature modulations, and saturate toward a performance ceiling. Thus, we model the evidence accumulation rate, $r$, as a sigmoid function:

$$r(s) = (r_{max} - r_{min})\frac{1 - e^{-\lambda s}}{1 + e^{-\lambda s}} + r_{min}, \tag{5.7}$$

where $r_{max}$ is the peak evidence accumulation rate, $r_{min}$ is the minimum rate with $r(s = 0) = r_{min}$, and $\lambda$ determines the slope of the sigmoid.

To enable the model to predict the adaptation rate across different target speeds, $v$, we fit polynomial functions to $r_{max}$, $r_{min}$, and $\lambda$ based on our experimental data.

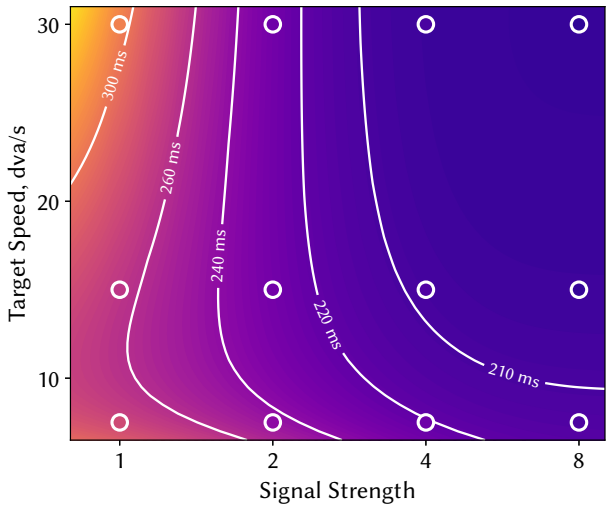



Figure 5.11: *Model visualization.* The predicted mean adaptation latencies of our model fitted to all experiment data, compared to the mean latency data it was fit on. Adaptation latencies for each target speed, and signal strength conditions are depicted in different colors, along with overlaid level sets indicating equivalent performance. Measured latencies are shown as individual points using the same color mapping.

Polynomial fits to the **LUM** data from Figure 5.9 yields the following coefficients:

$$\begin{aligned} r_{max}(v) &= -4.82 + 7.15v - 1.60v^2 \\ r_{min}(v) &= 3.62 + .763v - .113v^2 \\ \lambda(v) &= -.174 + .441v. \end{aligned} \tag{5.8}$$

### 5.4.2 Model Evaluation

To effectively evaluate our behavioral model, we aim to (1) validate the model's ability to generalize by showing that a model fitted to one type of visual feature variation (e.g., luminance in low/high external noise conditions or color) can successfully predict behavior for a different visual feature, and (2) demonstrate that the model's parameterization accurately captures the underlying patterns of behavior observed in Section 5.3.2.

**Generalizability.** We fit our model's parameters described in Equations (5.7) and (5.8) to the **LUM**, **NOISE**, and **COLOR** experimental results, separately, and measure how well each model's predictions agree with the results of the other two experiments. The $R^2$ scores for the **LUM** model were .99/.90/.93 across the **LUM**/**NOISE**/**COLOR** datasets respectively. Similarly, **NOISE** and **COLOR** model scores were .90/.97/.79 and .95/.83/.99 respectively. As expected, $R^2$ scores are highest for models evaluated against their respective trained datasets. Crucially, we observe that each model presents high fitting scores across all datasets ($R^2 > .79$), indicating that our signal strength-based model generalizes across different visual features [Chatterjee and Hadi, 2015].

**Accuracy.** We use a leave-one-out strategy to evaluate fitted models against unseen data, similar to prior work [Duinkharjav et al., 2022a] Model parameters in Equation (5.8) are fitted using data from four participants, and goodness-of-fit is assessed on the remaining participants. The $R^2$ scores for each participant (.80/.96/.92/.72/.72) indicate that our model effectively captures behavioral trends excluded from the fitting process.

### 5.4.3 Application Case Study: Performance-Aware Video Quality Assessment

Our visuo-motor adaptation performance model allows us to highlight multiple applications in computer graphics. A frequently overlooked challenge in producing highly dynamic films and animations is ensuring that rapidly moving targets remain trackable and successfully capture viewers' attention, enabling both comprehension and engagement [Smith, 2013]. Employing our model to computationally analyze animations can highlight events where the observer would miss crucial details without the need for human agents to detect them manually. In this case study, we demonstrate that our

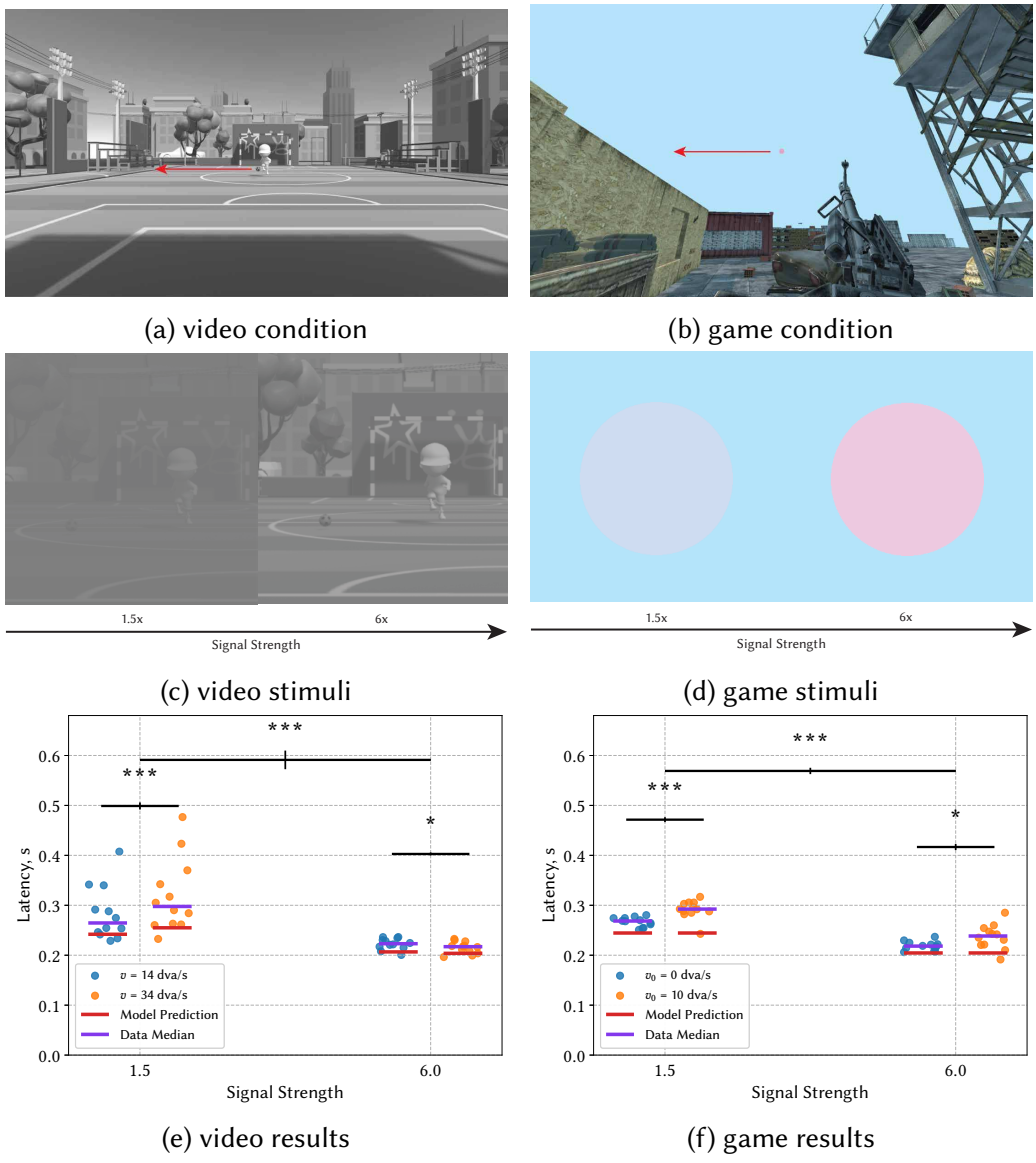
model is able to predict such events even in the presence of complex stimuli and scene backgrounds.

**Participants and setup.** Fourteen participants[2] (ages 22 - 28, 6 female) with normal or corrected-to-normal vision were recruited for a similar psychophysical experiment as in Section 5.3.1, where they were asked to complete a series of visual target tracking tasks. The overall experimental setup and protocols remain the same, but featuring major changes in the visual stimuli.

**Stimuli and design.** During each trial, as shown in Figure 5.12a, participants viewed a scene featuring a soccer goal-keeping video and were instructed to visually track a target ball (Figure 5.12c) moving either leftward or rightward. The target's motion and visibility were modulated across conditions, with two target speeds ($v \in \{14, 34\}$ dva/s), and three target visibility levels ($s \in \{1.5, 6, 8\}$).

To simulate more realistic motion, target speed was controlled by applying a force to the ball rather than maintaining a constant velocity. Target speed was calculated by averaging the ball's velocity over the first .2 s of motion. Each condition was repeated 20 times for a total of 120 trials and was completed in about 25 min during a single session. To avoid learning effects from repeated exposure to the same clips, each instance of a condition included slight variations in motion trajectories and camera angles. Before the experiment, participants completed an adaptive staircase to calibrate the global contrast of the video to determine the just-detectable threshold of successfully tracking the target.

[2]Two participants (1 female) were excluded for inability to perform the tasks (excessive blinking from dry eyes which disrupted eye-tracking accuracy, and colorblindness).




(a) video condition (b) game condition

(c) video stimuli (d) game stimuli

(e) video results (f) game results

Figure 5.12: *Scene, stimuli, and results of the application case study for videos and games.* The scenes evaluated were a soccer field (a) and an FPS game map (b). The stimuli were altered along global luminance contrast in (c) and color contrast in (d). The study results are shown in (e) and (f). The individual points reflect mean responses for each participant, color coded by target velocity. Error bars above the scatter plots indicate SEM of the difference between target velocity within each signal strength, and between different signal strengths overall. ***p < 0.001, **p < 0.01, *p < 0.05

**Analysis and results.** Adaptation latency measurement analysis was unchanged from Section 5.3.1, albeit with more relaxed gaze tracking error criteria. Following the same baseline reaction time normalization methodology as in Section 5.2, we use the $s = 8$ condition as the baseline condition. Adaptation latencies for other conditions were rescaled so that latencies in the benchmark condition were normalized across participants.

A two-way ANOVA reveals a significant main effect of both signal strength ($F = 24.12, p < 0.001$) and target speed ($F = 13.53, p < 0.01$) on adaptation latency, with a significant interaction between groups ($F = 20.54, p < 0.001$). Mean adaptation latencies were aggregated from each participant, and were compared to our model predictions in Figure 5.12e. A repeated measures ANOVA on the mean adaptation latencies reveals a significant effect of target speed on adaptation latency for both $s = 1.5$ ($F = 24.96, p < 0.0001$) and $s = 6$ ($F = 6.71, p < 0.05$) conditions.

**Discussion.** The study suggests that our model is able to effectively predict time periods during which observers are not able to visually track and thus comprehend visual details of target stimuli, even when the target and scene backgrounds are more visually complex. The model reflects the significant effects of signal strength and target speed on adaptation latency as discussed in Section 5.3.2. Since both signal strength and target speed significantly impact latency, this evaluation underscores the importance of including target speed in modeling adaptation latency.

### 5.4.4 Application Case Study: Controlling Difficulty of Visual Tasks in Gaming

In dynamic gaming environments, the complexity of visual target movement patterns can impact players' ability to react and engage with in-game events [Durst et al., 2024]. Moreover, player skill has been shown to correlate with their performance in low level gaze behaviors [Velichkovsky et al., 2019]. Consequently, the ability to predict and control users' behavioral performance by adjusting visual targets with complex target motion patterns can be a highly beneficial tool for game design.

In this section, we explore whether our model can predict human responses to complex target motion patterns towards being deployable in downstream applications in the gaming space. Specifically, we examine the performance of executing a visuo-motor adaptation from an *initial non-zero* velocity state to another velocity state. The study was completed by the same participants as Section 5.4.3, following nearly identical procedures, except for differences in environment and target stimulus appearance.

**Stimuli and conditions.** At the start of each trial, as illustrated in Figure 5.12b, a solid disk target of the same design as in the **COLOR** experiment appeared to replace the fixation crosshair, and moved at an initial speed of $v_0 \in \{0, 10\}$ dva/s. After $1 - 1.5$ s, the target's speed suddenly changed by $\Delta v = 20$ dva/s in either horizontal direction. Throughout every trial, the camera moved in a random forward direction. Target visibility conditions and number of repetitions remained unchanged as in Section 5.4.3.

**Analysis and Results.** To effectively analyze the onset of adaptation from the initial velocity condition to the latter, we required an adjustment to our adaptation onset detection protocol. To this end, we added a velocity offset to each trial's gaze recording

equal to its corresponding $v_0$ value. This velocity "nulling" process adjusts the target trajectory to be approximately in the retinal reference frame of the participant if they were tracking the target during the initial velocity tracking phase perfectly. We applied our same analysis procedure as in Section 5.4.3 on this nulled position trace to detect adaptation latencies.

A two-way ANOVA reveals a significant main effect of both signal strength ($F = 210.99, p < 0.001$) and target speed ($F = 20.84, p < 0.001$) on adaptation latency, but with no significant interaction effect. Mean adapataion latencies were aggregated as in Section 5.4.3, and compared to model predictions in Figure 5.12f. A repeated measures ANOVA on the mean adaptation latencies reveals a significant effect of target speed on adaptation latency for both $s = 1.5$ ($F = 24.96, p < 0.0001$) and $s = 6$ ($F = 6.71, p < 0.05$) conditions.

**Discussion.** We see that, despite the added complexity of the initial velocity condition, the behavioral trends remain consistent with our model predictions. Notably, a faster initial velocity significantly increase adaptation latency, as shown in Figure 5.12f. This motivates interesting future work on understanding object tracking while the eyes are already engaged in pursuit.

Overall, our model successfully captured adaptation performance trends, even under extremely challenging visual conditions. Observers faced increased target movement complexity, global optic flow from camera motion, and an off-white background adaptation—conditions not tested in our prior experiments. The consistency of the behavioral trends in these scenarios suggests that our findings can reliably extend to more complex and interactive target contexts while continuing to provide valuable predictions about user adaptation behavior.

### 5.4.5 Application Case Study: Content Optimization for Eye-Display Distance

In computer graphics applications, eye-display distance often changes due to ergonomics and display environments (e.g., VR/AR headsets), affecting perceived size, and speed of observed content. Not only does our work suggest that users' visuo-motor adaptation performance is affected by changes in content speed, but it's also affected by changes in content size as well; visual stimulus size modulates users' contrast sensitivity to the stimulus [Barten, 1999b], and by extension modulates signal strength as well. By jointly accounting for the change in contrast sensitivity, and target speed due to changes in display viewing conditions, we are able to bootstrap our behavioral model to make significantly more powerful predictions.

As a proof-of-concept visualization, in Figure 5.13 we show how the visuo-motor latency of adapting a $d = 1$ inch wide visual target with a spatial frequency of 1 cycles-per-degree moving at $v = 5$ inch/s across a 27 inch display changes as a function of both its luminance contrast, and eye-display distance. We compute the underlying signal strength (visualized as a colormap) for each contrast and eye-display distance condition by applying Barten's contrast sensitivity function on the described stimulus, and overlaid the resultant adaptation latency prediction according to our model.

## 5.5 Discussion

In our first experiment, we examined how human behavior is influenced by a few key factors that strongly impact decision-making. However, our resulting model was fairly limited in its ability to generalize to visual patterns beyond those studied in this work. As discussed in Section 5.2.3, while our predictions broadly aligned with measured

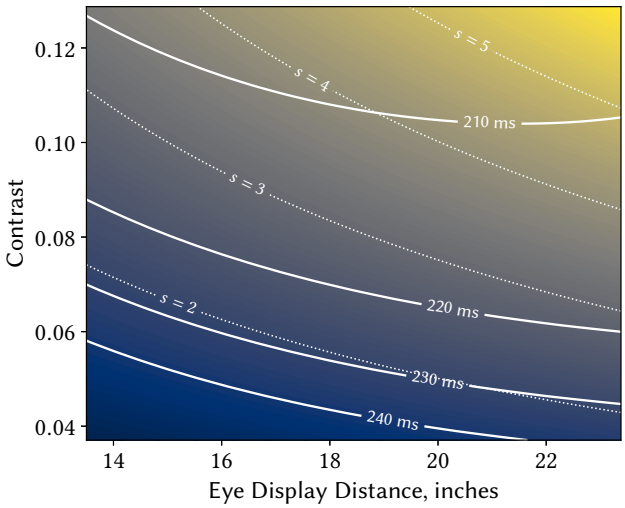



Figure 5.13: *Optimizing content appearance with eye-display distance.* Predicted visuo-motor adaptation latencies are visualized as solid contour plots overlaid on a colormap of the signal strength which was estimated using Barton's contrast sensitivity function for the stimulus described in Section 5.4.5. Signal strength contours are visualized via dotted lines.

responses, the approach of pooling band-pass-segregated visual features to predict visual performance may not be fully robust.

Additionally, this study did not account for scenarios where other factors—such as color, noise, and temporal changes—could further influence performance. Without direct measurements encompassing all possible combinations of these factors, it is challenging to apply our model beyond the specific contrast, frequency, and eccentricity measures we investigated.

In the subsequent experiment described in Section 5.3, we incorporated some of these considerations and further refined our approach to measuring and modeling human temporal eye movement behavior. A core objective of this follow-up study was to determine whether general principles of target visibility could facilitate the extension of behavioral models to a broader range of conditions. We envisioned that recent advancements in multi-dimensional target visibility models [Cai et al., 2024; Mantiuk et al., 2022] could contribute to behavioral research by complementing our

findings. Indeed, strong evidence suggests that using target *visibility* (or signal strength) as a unifying metric for visual target appearance provides a promising framework for systematically studying its effects on visual performance. If this approach holds, we could leverage existing models of visual detection and discrimination to compute target signal strength and then apply a general performance model that depends solely on this parameter.

However, our static target saccadic reaction time experiments in Section 5.1 reveal conflicting evidence: targets appearing in the mid-periphery elicit faster reactions than those in the fovea. This finding complicates the formulation of broad generalizations about the relationship between target signal strength and reaction time performance. As a next step, a systematic exploration of the relationship between visual sensitivity and behavioral performance patterns may provide a clearer understanding of how these mechanisms interact.

Beyond investigating the influence of visual target factors on the low-level behavioral performance metrics studied here, it is also crucial to explore higher-level measures of behavioral performance as well. While our work has focused exclusively on low-level performance, most real-world applications are influenced by higher-level factors—such as long-term cognitive behavior [Rosenholtz et al., 2012], visual attention [Krajancich et al., 2023; Rensink et al., 1997], and task and image salience and familiarity [Jarvenpaa, 1990; Rosenholtz, 2020]—which play a significant role in determining task difficulty. Ultimately, gaining a more comprehensive understanding of how limitations in low-level visual behavior shape higher-order decision-making would enable the design of more effective human-interactive systems.

A core aim of our work is to establish is how target visibility and motion jointly influence our adaptation efficiency. Since establishing a fine-grained and comprehensive

visibility function for signal strength itself is not our focus, we performed a calibration procedure for individual participants. We envision that recent advancements in cross-population and unified visibility models [Cai et al., 2024; Mantiuk et al., 2022] may shed light on a statistical model to bypass individual calibrations.

In Section 5.4.4, we validated the non-effect from camera motion-induced retinal optical flow during first-person shooter gameplay. In the experiments, participants were instructed to track a single moving target. However, in real-world scenarios, multiple peripheral targets may appear and move anisotropically, potentially influencing localization performance [Ludwig et al., 2014b]. A promising future research direction could be exploring the motor adaptation performance in the visual optical flow space to establish a robust and generalizable model for complex interactive applications.

Our current measurements and model are based on common desktop applications, where observers remain stationary while viewing the display. In emerging head-tracked displays, such as VR/AR headsets, vestibular cues may interact with and enhance visual localization [Fetsch et al., 2009], thereby improving oculomotor adaptation performance. While controlling observers' head motion is challenging, the emerging large-scale egocentric head-eye motion dataset [Grauman et al., 2022] may enable a characterized visuo-vestibular-motor joint modeling.

## 5.A Deriving Equation (5.6)

We are interested in deriving an expression for the probability distribution function for $T_{dual}$ as shown in Equation (5.5).

$$T_{dual} = \max(T_f, T_p).$$

We know that both $T_f$ and $T_p$ are Inverse Gaussian (IG) random variables as detailed in Equation (5.3),

$$T_f \sim \mathcal{IG}(\alpha_f, \nu_f)$$
$$T_p \sim \mathcal{IG}(\alpha_p, \nu_p).$$

The probability that $T_{dual}$ is less than some time $t$ is equivalent to the statement that both $T_f$ and $T_p$ are less than $t$. I.e.,

$$\mathbb{P}(T_{dual} \leq t) = \mathbb{P}(T_f \leq t)\mathbb{P}(T_p \leq t), \tag{5.9}$$

or,

$$H_{dual}(t) = H_f(t)H_p(t), \tag{5.10}$$

where $H_f$ denotes the cumulative density function (CDF) of the IG distribution with parameters $\alpha_f$ and $\nu_f$, and vice versa for $H_p$. The probability density function of $T_{dual}$ is therefore equal to the derivative of $H_{dual}$.

Taking the derivative from Equation (5.10) we get,

$$h_{dual}(t) = h_f(t)H_p(t) + H_f(t)h_p(t). \tag{5.11}$$

Since we have an explicit expression for the PDF of $T_{dual}$, we can finally write down

an expression for the likelihood function from Equation (5.6) as

$$\begin{aligned} L(\alpha_f, \alpha_p; t, v_f, v_p) &= h(t; \alpha_f, v_f) H(t; \alpha_p, v_p) + \\ &+ H(t; \alpha_f, v_f) h(t; \alpha_p, v_p), \end{aligned} \tag{5.12}$$

where $h$ and $H$ are the PDF, and CDF functions of the IG distribution.

## 5.B Field-of-view vs Eccentricity & Frequency

The observed image characteristics of stimuli shown on a display vary depending on how far the display is from the eye. We correlate these effects using the field-of-view that the display occupies as a measure of eye-distance. FoV is an intuitive way to measure eye-distance as it can be used regardless of the specific dimensions of a given display.

Given a display with width $w$, presented at an FoV of $\theta_{\text{fov}}$, the distance of the display equals

$$d = \frac{w/2}{\tan(\theta_{\text{fov}}/2)}. \tag{5.13}$$

If an observer is staring at the center of the display at FoV of $\theta_{\text{fov}}$ (or equivalently at a distance of $d$), an object $x$cm away from the center of the display will appear at

$$\theta = \arctan \frac{x}{d} = \arctan \left( x \frac{\tan(\theta_{\text{fov}}/2)}{w/2} \right) \tag{5.14}$$

retinal eccentricity. Hence, we notice that changing the eye-distance of a display alters the eccentricity at which stimuli appear in the retina.

Additionally, we can use this relation to derive a rate-of-change coefficient between

physical distances (in cm), and retinal eccentriticies (in degrees) by taking the derivative of Eq. (5.14),

$$\frac{d\theta}{dx} = \frac{\cos^2\theta}{d} = \cos^2\theta \frac{\tan(\theta_{\text{fov}}/2)}{w/2}. \tag{5.15}$$

This measure of "degrees-per-distance" allows us to derive the relationship between the spatial frequency of a pattern shown on the screen, $f_{\text{display}}$ (in cycles-per-centimeter), and the retinal frequency that an observer perceives, $f_{\text{retina}}$ (in cycles-per-degrees),

$$f_{\text{retina}} = f_{\text{display}} \frac{1}{\cos^2\theta} \frac{w/2}{\tan(\theta_{\text{fov}}/2)}. \tag{5.16}$$

Note that the observed frequency not only depends on the FoV, but also the eccentricity at which the stimulus is shown. For the simplest case where the stimulus is at the center of the screen, or $\theta = 0$, the relationship simplifies to

$$f_{\text{retina}} = f_{\text{display}} \frac{w/2}{\tan(\theta_{\text{fov}}/2)}. \tag{5.17}$$

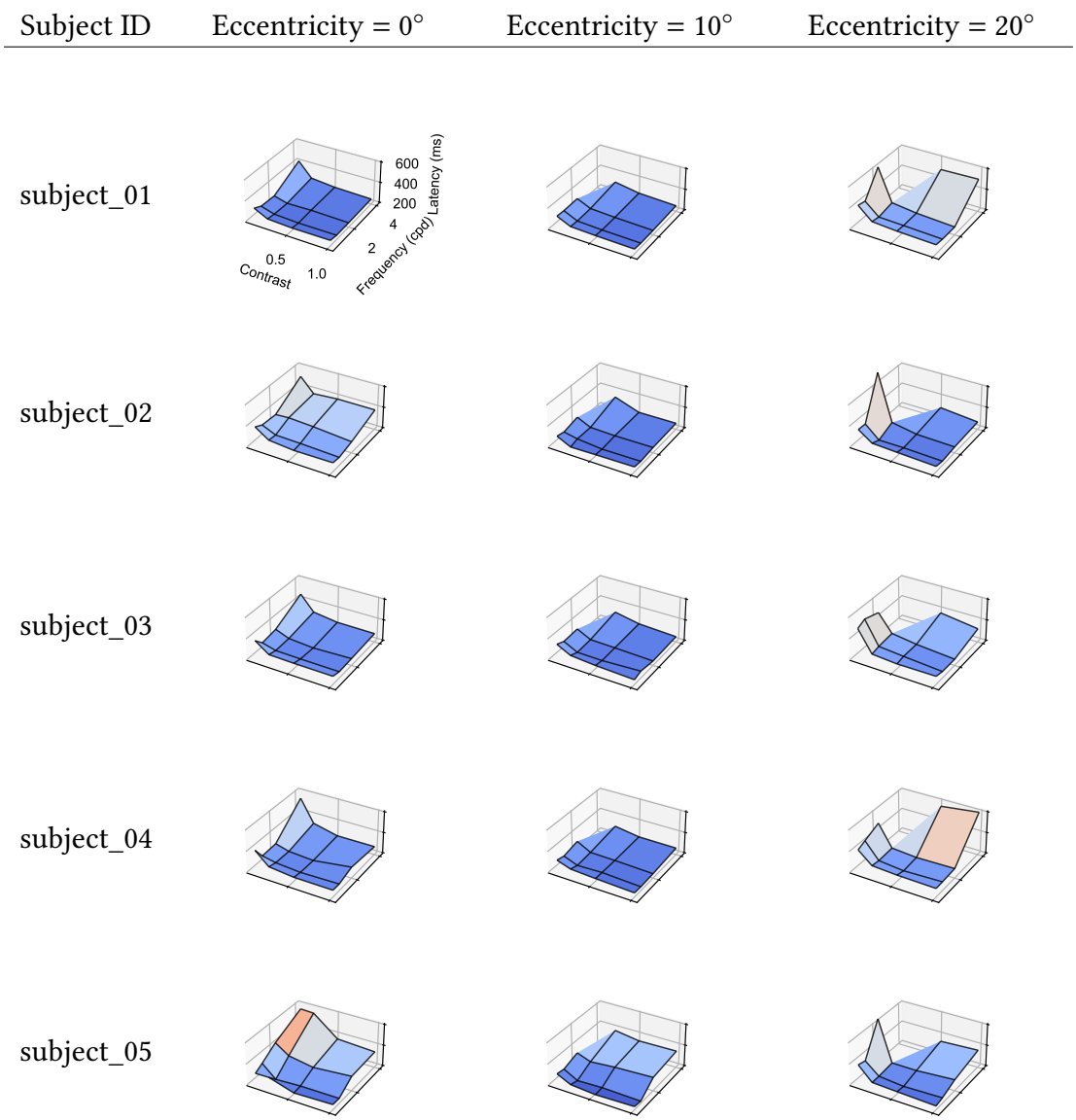



Figure 5.14: *Aggregated data of the pilot experiment.* Each subject completed 50 repetitions for each of the 45 conditions across 10 blocks of the user study. Each vertex in these surfaces represent the mean saccade latency of 50 trials with the same condition for each subject.

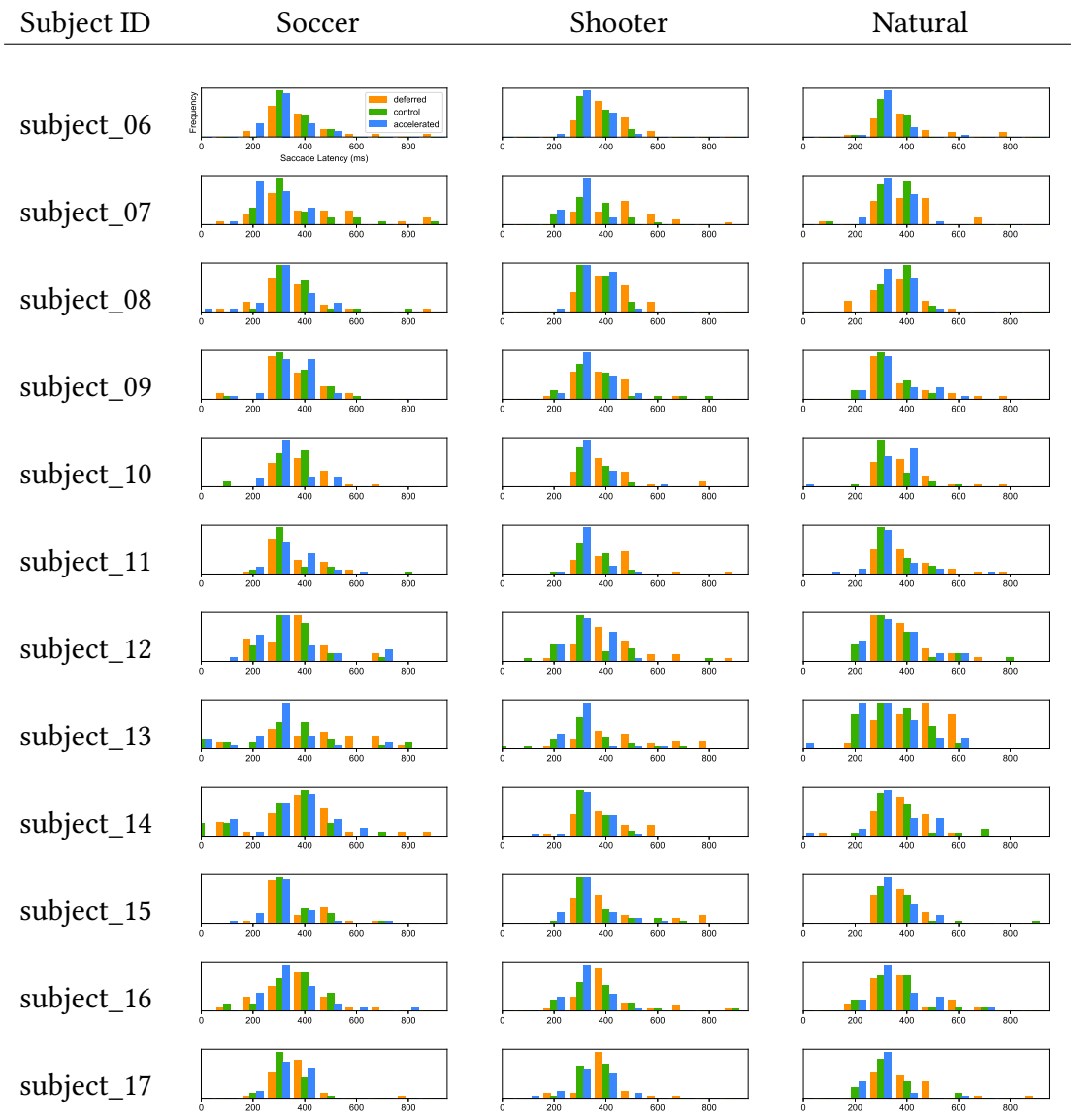



Figure 5.15: *Saccade latency histograms for Figure 5.5.* Each subject completed 51 trials for each condition, for each scene for a total of 459 trials. The latencies have been normalized to a common mean to enable quick comparisons between histograms.

# Chapter 6

# Eye Movement Motor Control Performance

Visual acuity, being much higher in the central region of the retina, encourages observers to shift their gaze to bring targets of interest into the fovea prior to analyzing any details. The speed of these movements are critical in complex tasks such as driving, where we rapidly move our eyes to acquire a plethora of information from the surroundings such as the presence of pedestrians, the approaching of vehicles, the speedometer reading, and even GPS navigation instructions.

We discussed in Section 2.3.4 that different gaze movement patterns are dictated by the strengths and limitations of the visual system. Due to the underlying neurological and mechanical limitations of eye movements, each one exhibits distinct performance characteristics; some are slow and steady, while others are ballistic and jerky. The combination of all classes of movements forms an efficient and comprehensive overall gaze behavior strategy in 3D visual environments.

In this chapter we ask "how long is the delay between beginning a gaze shift and

completing it, and how does it depend on the displacement of our gaze location?". With the emerging adoption of virtual/augmented reality (VR/AR), answering this question enables us to design 3D content that allows for an efficient target changing. We present the first operational model that predicts the required eye movement completion time necessary for shifting the gaze to new 3D targets in stereoscopic virtual environments.

We recognize the current lack of first-principle consensus on how vergence/combined eye movements are neurologically constructed (see Section 2.3.4). Additionally, we note that noise in both human behavior and eye-tracking adds difficulty to comprehensive study of complex stereoscopic movements with downstream applications. Circumventing these obstacles, we take a holistic approach to (1) focus on *when* both eyes land on a target after its onset, instead of the intermediate trajectory; and (2) form a computational model which accounts for the noise and variability to produce a *probabilitic* prediction, instead of a deterministic one.

We fit our model and validate its accuracy using our psychophysical study data, which includes more than $12,000$ individual trials to measure the temporal offsets of gaze movements in a stereo VR environment. The results evidence the model's consistent prediction accuracy, generalizability to unseen participants and trials, as well as the capability of forecasting and optimizing task performance with various real-world VR scenarios. Our model can be applied to measure the difficulty of video games in VR and how the scale of variability in depth can alter gaze movement behaviors for users. We also explore how completion time predictions can be used as a metric for evaluating the placement of 3D UI elements in VR/AR applications. Recalling the driving example, we can improve driver awareness by placing a virtual car dashboard overlay (with speedometer readings and navigation instructions etc.) in an adaptive manner to minimize completion times of objects that appear in the driver's periphery

in changing surrounding environments.

This research aims to propose an operational model for computer graphics applications for a behavioral phenomenon that is yet to be fully understood. We believe that providing a quantitative understanding of how emerging VR/AR technology influences statistical signatures of human target-changing performance during daily tasks is beneficial even without the neurological understanding of the underlying behaviors. We hope the research can serve as a novel benchmark to guide 3D interfaces and act as a metric for the user performance in various applications and mediums. Source code and data for this chapter's contents are available at `www.github.com/NYU-ICL/stereo-latency`.

## 6.1 Measuring and Predicting Stereoscopic Eye Movement Completion Time

To quantitatively understand combined stereoscopic eye movements, we first performed a psychophysical experiment with a wide field-of-view stereo VR display. The study measured how jointly varying vergence and saccade amplitudes influence the time required for an observer's eyes to reach a 3D target relative to stimulus onset; this duration is often referred to as the eye movement *offset time*. The data then serve as the foundation of our model (detailed in Section 6.1.1) for predicting the offset timing of various eye movements.

**Participants and setup.** Eight participants (ages 20-32, 6 male) with normal or corrected-to-normal vision were recruited. Due to the demanding requirements, established low-level psychophysical research commonly starts with pilot studies involving a small number of participants and leverages the collected data to develop computational

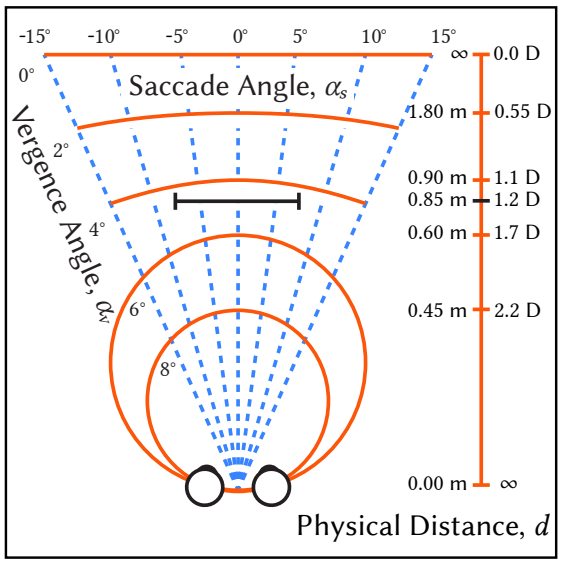



Figure 6.1: *Definition of measured angles.* We illustrate how we define and measure the angles of eye vergence movements $\alpha_v$, and saccadic movements $\alpha_s$ throughout the chapter. For further intuition, the physical distance of objects appearing at $\alpha_s = 0°$ is illustrated in units of meters, and Diopters (i.e., reciprocal of meters). Here, inter-pupillary distance (IPD) is chosen to be equal to the human average of 63 mm [Fesharaki et al., 2012]. The optical display depth of the headset is overlaid as a horizontal black bar at a depth of 0.85 m, or 1.2 D.

models (e.g., the foveated rendering literature [Krajancich et al., 2021, 2023; Patney et al., 2016; Sun et al., 2020]). These models, constructed using data from a limited set of subjects, can be evaluated for their cross-subject generalizability using a larger group of users, as we performed in Section 6.2.3 with 12 additional unseen participants. Moreover, in the context of our work, psychophysical studies examining the temporal dynamics of human behaviors require remarkably large sample sizes for a comprehensive statistical pattern to account for neural and mechanical noise [Bucci et al., 2006; Collewijn et al., 1995; Erkelens et al., 1989; van Beers, 2007; Yang and Kapoula, 2004]. Considering that variations among subjects do not exhibit a significant impact on the completion rate of low-level gaze movements like saccades [Bahill et al., 1975b] and vergence movements [Collewijn et al., 1995; Erkelens et al., 1989]—as confirmed by our cross-validation analysis in Section 6.2.2—and given that these are objective psychophysical behaviors

not reliant on subjective reporting, we chose to enlist a small number of participants while acquiring an extensive sample size (1,500+ trials) per participant. To this aim, we split the study across multiple days for every participant (see *Conditions* paragraph for details).

Table 6.1: *Varjo Aero specifications.*

| Resolution | Frequency | Peak Luminance |
|---|---|---|
| $2880 \times 2720$ | 90 Hz | 150 cd/m$^2$ |
| Focal Distance | FoV | Supported IPD |
| 0.85 m | 134° (diagonal) | $59 - 71$ mm |
| Eye Tracker | Frequency | Accuracy |
| | 200 Hz | $< 1°$ |

The study was conducted with a Varjo Aero head-mounted VR display (HMD) with the relevant specifications detailed in Table 6.1. As shown in Figure 6.2a, throughout the study, participants wearing the HMD remained seated and performed the visual-target-changing task as detailed in the *Task and Stimuli* paragraph. Before the experiment, participants underwent a "preamble" checklist to ensure proper task completion and accuracy, including:

1. Measure and calibrate the HMD's inter-pupillary distance (IPD).
2. Complete a five-point calibration for accurate binocular gaze tracking (repeat whenever the HMD is re-mounted after breaks).
3. Adjust a fixation point between the nearest and furthest depths at which experimental stimuli appeared to ensure the success of fusing the stereoscopic visual stimuli (i.e., no double-vision).

**Task and stimuli.** Participants' task was to shift their gaze to land on targets appearing in 3D space. At the beginning of each trial, they were instructed to observe the

fixation stimulus at the center of the screen. As illustrated in Figure 6.2a, this stimulus included a combination of a cross and four circular flankers to assist fixation [Thaler et al., 2013]. Once successful fixation was detected, this stimulus disappeared and was immediately replaced by a target stimulus, to which participants were instructed to move their gaze to as naturally as possible with a single gaze motion. The target stimulus was a Gaussian blob with $\sigma = 0.25^\circ$ and peak luminance of 150 cd/m$^2$ — a similar design as in Lisi et al. [2019].

To ensure stable tracking, a trial only began if the participant's eyes were within $1.2^\circ$ to the center of the fixation point for a consecutive 0.4 s. If the participant failed to hold their gaze at the fixation point for sufficient duration more than three consecutive times, the eye-tracker was re-calibrated. Additionally, to ensure correct task completion, we rejected and repeated a trial if it was completed in less than 0.1 s or more than 1.3 s. To avoid fatigue, participants were shown a darkened screen between trials as a cue to blink or close their eyes, if they: (1) successfully completed a trial, (2) failed to hold their gaze on the starting fixation point, or (3) failed a trial.

**Definitions and annotations.** Offset times are known to vary depending on the spatial location of the stimuli, mostly due to the varying contributions of either saccadic or vergence movements, often superimposed on each other [Zee et al., 1992]. In order to study how the spatial placement of the stimuli influences what type of eye movements arise, we parameterize spatial locations using two parameters: the vergence angle, $\alpha_v$, and the saccade angle, $\alpha_s$, as illustrated in Figure 6.1. All locations in the transverse plane containing the participants' eyes, and the stimuli can be encoded using the two degrees of freedom provided by $\alpha_v$ and $\alpha_s$.

Specifically, following vision science practice, we define the vergence angle as the

angle formed by the intersection of the gaze rays. That is, if we denote the signed angles of the left and right eyes, with respect to the forward "$z$" direction (i.e. the intersection between the transverse and median planes) as $\alpha_l$ and $\alpha_r$, the vergence angle is equal to

$$\alpha_v = \alpha_l - \alpha_r. \tag{6.1}$$

The set of gaze locations that have the same $\alpha_v$ form an *isovergence circle*, visualized as the orange circles in Figure 6.1. Pure vergence movements maintain the direction of gaze and move the gaze point from one isovergence circle to another.

On the other hand, the saccade angle, $\alpha_s$, is defined as the mean of the angles of the left and right eyes:

$$\alpha_s = (\alpha_l + \alpha_r)/2. \tag{6.2}$$

The set of gaze locations that have the same $\alpha_s$ form a ray representing the direction of gaze, visualized as the blue lines in Figure 6.1. Pure saccadic movements remain on the same isovergence circle while rotating the direction of gaze across the transverse plane.

Therefore, a vergence and saccade angle pair, $\boldsymbol{\alpha} = (\alpha_v, \alpha_s)$, uniquely defines a point on the transverse plane via the intersection of the isovergence circle which corresponds to $\alpha_v$, and the direction of gaze which corresponds to $\alpha_s$. An arbitrary gaze movement in this coordinate system can be represented as a displacement vector,

$$\Delta\boldsymbol{\alpha} = \boldsymbol{\alpha}^t - \boldsymbol{\alpha}^o = (\alpha_v^t - \alpha_v^o, \alpha_s^t - \alpha_s^o) = (\Delta\alpha_v, \Delta\alpha_s), \tag{6.3}$$

for movement from $\boldsymbol{\alpha}^{o(rigin)} = (\alpha_v^o, \alpha_s^o)$ to $\boldsymbol{\alpha}^{t(arget)} = (\alpha_v^t, \alpha_s^t)$.

**Conditions.** We define a condition by a pair $\{\boldsymbol{\alpha}^o, \Delta\boldsymbol{\alpha}\}$. We sought to create a grid of experimental conditions which cover a wide set of possible gaze movements. Today's VR devices limit the breadth of applicable eye movements. Here we discuss these limitations as well as the solutions we implemented to ensure study accuracy.

First, we observed that participants could not fuse a stereo stimulus when it was placed too close, causing double (yet in-focus) vision. This restricted the range of possible vergence movements we could study in VR. We believe this effect is due to the lack of support for variable accommodation in VR displays, and thus distorted depth cues due to the *vergence-accomodation conflict* [Aizenman et al., 2022; Hoffman et al., 2008; March et al., 2022]. To establish a conservative *minimum* depth with successful stereo stimulus fusion, we performed a pre-study test with 4 participants with various inter pupil distances (IPDs) (64 − 71 mm). Through this experiment, we established that this depth is approximately $d_{min} = 0.4$ m in front of the observer. This corresponds to a *maximum* vergence angle coordinate of $\alpha_v^{max} = 8.4°$ for an observer with an IPD of $w_{IPD}^{min} = 59$ mm — the lowest IPD supported by the HMD (see Table 6.1). Since a larger IPD only relaxes this maximum value, we limit the maximum vergence angle to $\alpha_v^{max} \leq 8.4°$. See Section 6.A for a more in-depth analysis.

Second, we found that the accuracy of the HMD eye tracker deteriorates significantly further in the periphery for $\alpha_s \geq 15°$. We recognize that the majority of saccades naturally performed by humans have amplitudes $\alpha_s \leq 15°$ [Bahill, 1975], due to a preference to move the head otherwise. Therefore, we limit the maximum saccade angle to $\alpha_s^{max} \leq 15°$.

Lastly, due to the inconsistent nature of temporal human behavior, our study requires many repeats for each condition in order to reveal statistical trends. It is therefore infeasible to include a large number of conditions in our study. We address this by only

sampling gaze movement displacements, $\Delta\boldsymbol{\alpha}$. That is, although the initial gaze position $\boldsymbol{\alpha}$ has been shown to be a relevant factor influencing offset time [Templin et al., 2014], we chose not to consider it in our analysis and modeling for the current study. We leave characterizing the effects of "starting pose" as future work.

To summarize, our study design is constrained to vergence angles $\alpha_v \leq 8.4°$, saccade angles $\alpha_s < 15°$, as well as to only consider gaze movement displacements, $\Delta\boldsymbol{\alpha}$, and to ignore initial gaze positions, $\boldsymbol{\alpha}^o$. Within these constraints, we sample the following conditions for vergence, saccade, and combined motions respectively:

- 2 vergence conditions with amplitudes ($|\Delta\alpha_v| \in \{4.2°, 8.4°\}$) conducted for both divergent (−) and convergent (+) movements,
- 3 saccade conditions with amplitudes ($\Delta\alpha_s \in \{4°, 8°, 12°\}$) conducted at near and far depths,
- $2 \times 3$ combined movements for every combination of the above conditions for both convergent and divergent movements,

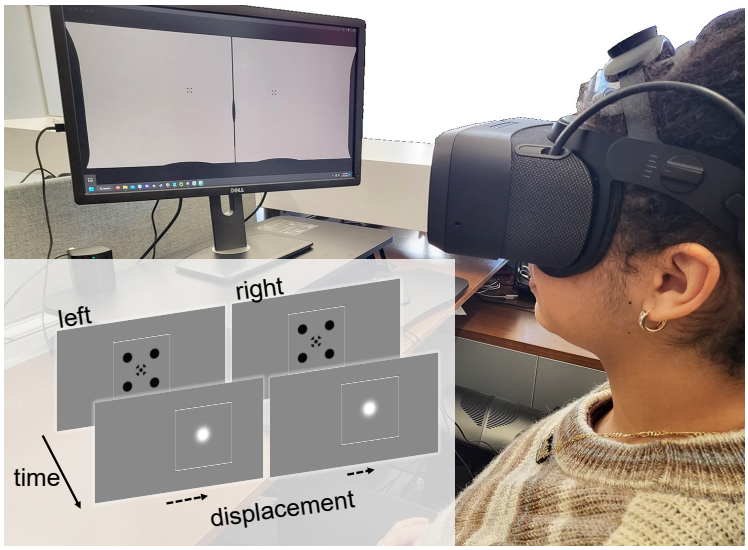
totaling in $(2 + 3 + 2 \times 3) \times 2 = 22$ conditions, as in Figures 6.2b and 6.2c. We treated leftward and rightward saccades as symmetric; therefore, while we randomized stimulus location to appear on the left or right side, in data processing, we remove the distinction by taking the absolute value of the saccade amplitudes. Implementation of the conditions is detailed in Section 6.A.

To account for human sensory and behavioral noise [van Beers, 2007], we repeated each condition 6 times within one experimental block (totaling in $6 \times 22 = 132$ trials per block), and instructed participants to complete a total of 12 blocks. Each block took $10 - 15$ minutes to complete, with a $2 - 3$ minute break between blocks. The experiment was split into sessions across 3 days to avoid fatigue, with each session

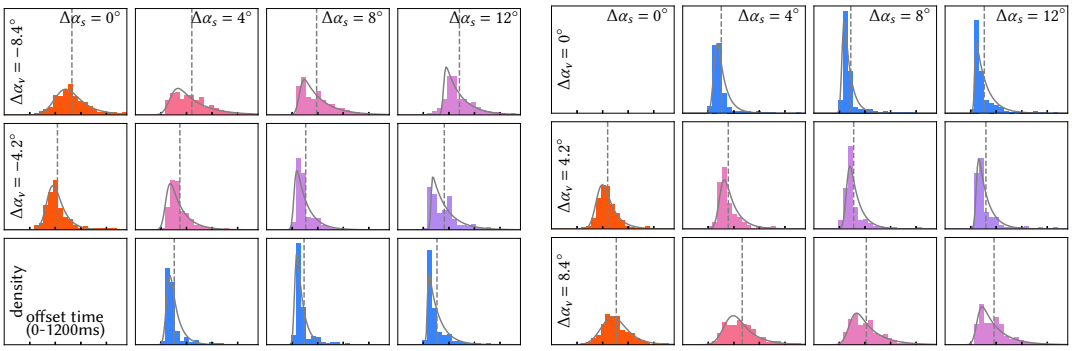
scheduled at approximately the same time for consistent performance. Before each session, participants also performed a short warm-up session of 24 trials to familiarize themselves with the task and target positions and eliminate potential variance in reaction time. Overall, each experimental condition was repeated a total of 72 times, and the entire experiment took about 3 hours for each participant, including intermediate breaks. Running the experiment across 8 participants, we collected a total of $8 \times 72 \times 22 = 12,672$ trials.

**Data analysis.** Each experimental trial yields a time-series of eye directions recorded during the trial, sampled at 200 Hz. Similar to Templin et al. [2014]; Yang et al. [2002, 2010], we performed post-hoc processing and analysis on the raw data to more precisely identify gaze movement offset times. To address tracker noise from high sampling frequency [van Beers, 2007], we first applied a 25 Hz smoothing filter [Butterworth et al., 1930], similar to Templin et al. [2014]; Yang et al. [2010].

We compute the angular velocity over time across each trial from the smoothed eye direction data and apply a constant velocity threshold to detect offset timestamps of gaze movement. Specifically, for a reliable offset time measurement, we require two conditions to be met: (1) individual speeds of the left and right eyes to be below a threshold of 5°/sec, as well as (2) each eye to be directed within 1° relative to the target. While some prior work suggests that vergence offset times can be detected by the angular velocity in the vergence dimension, i.e., $\frac{d}{dt}\alpha_v = \frac{d}{dt}(\alpha_l - \alpha_r)$ [Yang and Kapoula, 2004], we found that our strategy is more fitting in our use case due to the additional challenges in eye tracker precision, accuracy, and frequency posed by consumer VR devices. For consistency and fairness across all conditions, we applied this detection approach for all the conditions, including vergence-only, saccade-only, and combined

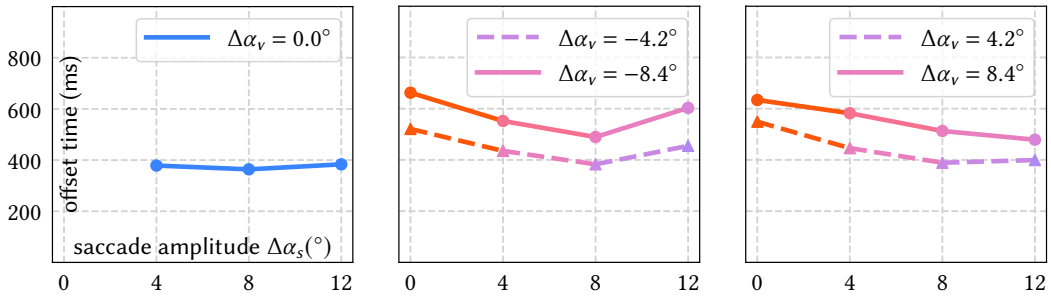



(a) setting and stimuli




(b) divergent (c) convergent

Figure 6.2: *Study setup and results.* (a) visualizes the setup and temporal stimuli (zoomed-in for illustration) of an example condition. (b)/(c) shows the histogram of the collected offset times, with divergent/convergent movement. Each sub-figure block indicates an individual condition. Higher vertical/horizontal locations imply higher vergence ($\Delta\alpha_v$)/saccade($\Delta\alpha_s$) amplitudes. In each block, the X-axis denotes the observed offset time ($0 - 1200$ ms range; 250 ms for each tick) and Y-axis denotes the corresponding distribution density. The dashed lines indicate the mean offset time of each histogram. For each histogram an Exponentially modified Gaussian (*ExGauss*) distribution is fitted via Maximum Likelihood Estimation (MLE); refer to Section 6.1.1 for details on the fitting procedure.




(a) saccade (b) divergent (c) convergent

Figure 6.3: *Aggregated mean offset time of studied conditions across all participants.* (a) shows the mean offset time of pure saccade conditions. X- and Y-axes indicate saccade amplitudes, $\Delta\alpha_s$, and mean offset time, respectively (offset time std shown in Figure 6.11). Note the consistency across varied amplitudes. (b)/(c) show the mean offset times with pure vergence ($\Delta\alpha_s = 0$) and combined movement ($\Delta\alpha_s \neq 0$) conditions. Note the non-monotonic/u-shaped effect of $\Delta\alpha_s$ on the offset time.

movement trails. A small percentage of trials (6.4%) were rejected from analysis and training due to the gaze offset position falling outside the allowable range. Manual inspection of these trials indicates that the users' eye movements only satisfied the second condition (2) above, but not the first (1). These cases could not be identified during experiment run-time due to the inability to reliably perform post-processing filters to the raw data on the fly.

**Results.** Figure 6.2 visualizes the raw data with the identified eye movement offset time. All time values in the statistical analysis below and throughout the chapter are in *seconds* for clarity. Additionally, Figure 6.3 statistically summarizes the mean of each condition.

The offset times of saccades ($\Delta\alpha_v = 0°$, .37 (mean) ± .12 (std)) are lower than offset times of vergence movements ($\Delta\alpha_s = 0°$, .59 ± .15). The effect applies for both divergent ($\Delta\alpha_v < 0°$, .59 ± .17) and convergent ($\Delta\alpha_v > 0°$, .59 ± .14) conditions. The average offset time of combined movements (.48 ± .16) lies in between. A repeated measures ANOVA

indicated that the type of eye movement (saccade/vergence/combined) had a significant effect on the offset time ($F_{2,14} = 339.3, p < .001$). Additionally, the range (max-min) of mean offset times across saccade conditions (.02) is significantly narrower than across vergence conditions (.14). The effect can be visualized by comparing the span of values on the $y$-axis of Figure 6.3.

Larger vergence amplitudes ($|\Delta\alpha_v|$) significantly prolong the offset time in combined movements. For example, the average landing time for $|\Delta\alpha_v| = 4.2^\circ/8.4^\circ$ is $.53 \pm .12/.65 \pm .16$. A repeated measures ANOVA indicated that the $|\Delta\alpha_v|$ had a statistically significant effect on the offset time ($F_{2,14} = 384.7, p < .001$).

For combined offset times, we did not observe a monotonic effect of saccade amplitude ($\Delta\alpha_s$). In fact, with a given vergence amplitude, the effect of saccade amplitude on the combined movement time is inconsistent and commonly non-monotonic, as visualized with the "U-shape" in Figure 6.3b. The average landing time for pure saccade conditions, $\Delta\alpha_s = 4^\circ/8^\circ/12^\circ$, are $.38 \pm .12/.36 \pm .11/.38 \pm .13$. When $\Delta\alpha_v = -8.4^\circ$, however, the fastest combined movement occurs for $\Delta\alpha_s = 8^\circ$ ($.49 \pm .16$), compared with the other two conditions $\Delta\alpha_s = 4^\circ$ ($.55 \pm .18$) and $\Delta\alpha_s = 12^\circ$ ($.60 \pm .15$). A Mann-Kendall trend test did not observe a significant monotonic trend ($\tau = .33, p = 1.0$).

The distribution of offset times across all conditions exhibits positive skewness ($\gamma_1 = 1.94 \pm .89$). Among the conditions, skewness varied by condition with pure vergence movements is the smallest (1.4), combined movements in the middle (1.8), and pure saccadic movements is the highest (3.1). This indicates that different gaze movements change the shape of the distribution of offset times, which can also be visualized from the histograms in Figure 6.2.

**Discussion.** The visualization and analysis draw us to several conclusions. First, the offset times of singular saccade movements are significantly shorter and more consistent than those of vergence movements. Second, statistical analysis of our data evidenced that slow vergence movements are "accelerated" if combined with faster saccades. Third, the acceleration effect varies depending on how they are combined. Saccade acceleration exhibits a "U-shape" for divergent combined movements (Figure 6.3b). The optimality (i.e., the amplitude of the saccade that accelerates vergence the most, thus the fastest combined movement) depends on the corresponding vergence amplitude. Lastly, human performance on changing 3D visual targets is inconsistent across trials, even within the same participant. Moreover, the scale of the inconsistency varies across different eye movements. These observations inspire us to develop a computational model that 1) depicts quantitatively how saccades accelerate vergence, and 2) predicts the probability distribution of target landing offset time with combined vergence-saccade movements.

### 6.1.1 Generalization to Arbitrary Gaze Movements

**Statistical model.** The statistical analyses in Section 6.1 motivate us to develop a model for predicting the target landing offset times for arbitrary gaze movements not present within our dataset. As reported in Section 6.1, the distributions observed in our dataset are positively skewed, and vary across different conditions; so an Exponentially modified Gaussian (*ExGauss*), which features fine control over skewness via its parameters, is a viable choice of statistical model for these distributions [Marmolejo-Ramos et al., 2023]. Specifically, offset time, $\mathcal{T}$, represented as an *ExGauss* random variable has

a probability density function (PDF),

$$f_{\mathcal{T}}(t;\mu,\sigma^2,\tau) = \frac{1}{2\tau} e^{2\mu+\frac{\sigma^2}{\tau}-2t} \operatorname{erfc}\left(\frac{\mu+\frac{\sigma^2}{\tau}-t}{\sqrt{2}\sigma}\right), \tag{6.4}$$

parameterized by $\mu$, $\sigma$, and $\tau$, to depict the location, spread, and asymmetry of the resulting distribution, respectively. All parameters are in units of *seconds*. Here, $\operatorname{erfc}(\cdot)$ is the complementary error function. As shown in Figure 6.2, we estimate the *ExGauss* parameters for each condition separately via Maximum Likelihood Estimation (MLE) to collect a total of $N = 19$ sets of parameters (not double counting the saccade conditions).

In this work, offset times are modeled as *ExGauss* random variables, but note that modeling with a different random variable may also be valid. We leave the analysis and comparisons among model choices as future work since the specific presentation is beyond our focus, and other parameterizations are adaptable to our framework.

**Parameter interpolation.** Our focus, instead, is on how the parameters of a given model should be interpolated to provide predictions of gaze offset times for arbitrary gaze movements. To this end, we leverage the *ExGauss* parameter estimations of each condition and smoothly interpolate each parameter via Radial Basis Function (RBF) interpolation. Concretely, each RBF takes, as input, the amplitude of the gaze movement, $\Delta\boldsymbol{\alpha} = (\Delta\alpha_v, \Delta\alpha_s)$, to output the predicted *ExGauss* random variable, $\mathcal{T}(\Delta\boldsymbol{\alpha})$,

with estimated parameters

$$\begin{aligned}\hat{\mu}(\Delta\boldsymbol{\alpha}) &\coloneqq \sum_{i}^{M} \boldsymbol{\lambda}_i^{\mu} \varphi(\varepsilon^{\mu} ||\Delta\boldsymbol{\alpha} - \mathbf{c}_i^{\mu}||),\\ \hat{\sigma}(\Delta\boldsymbol{\alpha}) &\coloneqq \sum_{i}^{M} \boldsymbol{\lambda}_i^{\sigma} \varphi(\varepsilon^{\sigma} ||\Delta\boldsymbol{\alpha} - \mathbf{c}_i^{\sigma}||),\\ \hat{\tau}(\Delta\boldsymbol{\alpha}) &\coloneqq \sum_{i}^{M} \boldsymbol{\lambda}_i^{\tau} \varphi(\varepsilon^{\tau} ||\Delta\boldsymbol{\alpha} - \mathbf{c}_i^{\tau}||).\end{aligned} \tag{6.5}$$

$\mathbf{c}_i^{\mu}$ and $\boldsymbol{\lambda}_i^{\mu}$ represent the location and weight of each of the $M = 4$ radial bases, $\varphi$ is the radial function, and $\varepsilon^{\mu}$ is a tuning shape parameter for the radial function. In our implementation, we used the Gaussian kernel, $\varphi(r) = \exp(-r^2)$. Overall, the learnable parameters in this regression are $\mathbf{c}_i^j$, $\boldsymbol{\lambda}_i^j$, and $\varepsilon^j$ for $i \in [1 \ldots M]$, totalling in $4 + 4 + 1 = 9$ variables for each *ExGauss* parameter $j \in \{\mu, \sigma, \tau\}$.

**Regression.** We optimize the adjustable variables via gradient descent to minimize the mean-squared error between the MLE-estimated *ExGauss* parameters for each condition, and the RBF-interpolated parameters, with the loss

$$L_j = \frac{1}{N} \sum^{N} \left(j - \hat{j}\right)^2 \text{ for } j \in \{\mu, \sigma, \tau\}. \tag{6.6}$$

The RBF parameters are regressed using batch gradient descent with the loss functions from Equation (6.6) and a learning rate of $10^{-2}$ for $200,000$ iterations. The mean-squared losses are minimized from $137k/2.3k/17k$ $\text{s}^2$ to $230/200/120$ $\text{s}^2$ over the course of each regression, respectively. We report model performance metrics as well as additional evaluations in Section 6.2.

**Discussion and applications.** We compare the mean offset times predicted by our model to the means aggregated from our dataset in Figure 6.4. This visualization demonstrates how the offset times differ between convergent and divergent gaze movements. For convergent combined movement, we observe the same monotonic decrease in offset time as a function of saccade amplitude as reported in Figure 6.3c. Additionally, we see the U-shaped behavior for divergent combined movements, as discussed in Section 6.1 and Fig. 6.3b.

The *ExGauss* distribution and RBF interpolation methods are represented by parameterized differentiable functions. This allows us to compose these components to construct an end-to-end differentiable model for predicting the probability distribution of arbitrary gaze movements. This formulation can be leveraged in various ways for practical applications. For example, the "optimal" saccade amplitude, $\Delta\alpha_s^*$, which minimizes the offset time at various vergence amplitudes, $\Delta\alpha_v$ can be computed analytically:

$$\begin{aligned} \Delta\alpha_s^* &= \arg\min_{\Delta\alpha_s} \mathbb{E}\left[\mathcal{T}\left(\Delta\boldsymbol{\alpha} = (\Delta\alpha_v, \Delta\alpha_s)\right)\right] \\ &= \arg\min_{\Delta\alpha_s} \left(\hat{\mu}\left(\Delta\alpha_v, \Delta\alpha_s\right) + \hat{\tau}\left(\Delta\alpha_v, \Delta\alpha_s\right)\right). \end{aligned} \tag{6.7}$$

These local minima indicate the location of the lowest point in the valley of the U-shaped behavior, as visualized in Figure 6.4.

## 6.2 Model Evaluation

We first measure the statistical accuracy and necessity of the vergence-saccade combined modeling with an ablation study in Section 6.2.1. We further test the model's goodness-

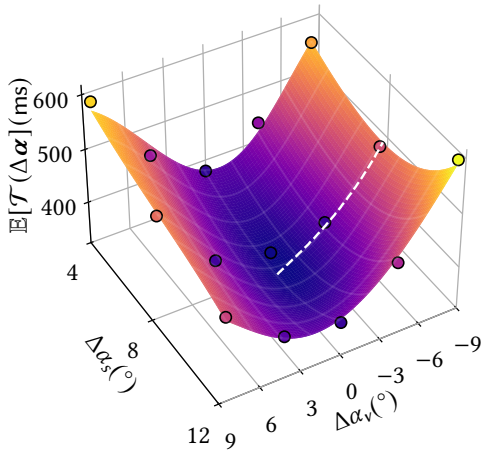



Figure 6.4: *Visualization of the interpolated model.* The sparsely sampled data visualized in Figure 6.3 is smoothly interpolated via RBF interpolation. The surface heatmap shows the mean offset times across all interpolated conditions, and the measured data is overlaid as a scatter plot for comparison. The "optimal" combined gaze movements at various vergence amplitude settings are computed using Equation (6.7) and visualized as a dashed white line on the surface of the model prediction.

of-fit when generalizing to unseen users and trials in Section 6.2.2. Then, to evaluate its applicability in real-world scenarios and novel conditions, we perform an evaluation user study with various scenes in Section 6.2.3.

### 6.2.1 Model Accuracy and Ablation Study

**Metrics.** We utilize the Kullback–Leibler divergence (KLdiv) as a continuous domain metric for measuring the similarity between model-predicted probability densities and the histograms obtained from the psychophysical data. A model with *lower* KLdiv relative to a ground truth histogram indicates a *better* prediction.

Table 6.2: *KL divergence of the model and ablation study.*

| Condition | **FULL** | **VER** | **SAC** |
|---|---|---|---|
| KL Divergence | .172 | .236 | .444 |

**Conditions.** We conduct an ablation study and utilize the KLdiv to validate the necessity of modeling combined movements. Specifically, we consider the model's prediction accuracy if not supplying it with information on either saccade or vergence movement. For this purpose, we re-aggregate our psychophysical data into groups separated only by saccade amplitude (**SAC**), or only by vergence amplitude (**VER**) conditions. That is, we pool together the histograms in Figure 6.2 across the columns, or rows respectively. The re-aggregation is then utilized to regenerate an ablated model following the same steps as described in Section 6.1.1. See Figure 6.12 for visualizations of the ablated model predictions.

While the probability distribution predicted by our model is continuous, the psychophysical study dataset only provides a finite sample of the theoretical ground truth distribution of offset times. Therefore, we apply the discrete version of KLdiv onto histograms of the ground truth data for each condition with $n = 50$ bins ($\Delta t = 24$ ms).

**Results and discussion.** The resulting average KLdivs for the two ablated models are compared to the full model (**FULL**) in Table 6.2. We observe that the **FULL** model exhibits significantly lower KLdiv than **VER** and **SAC**. While the number of bins does have an effect on the divergence values, we extensively tested and confirmed that the relative relationship across the three conditions was not influenced by this factor. These results demonstrate that combined eye movements exhibit remarkably distinct temporal patterns that depend both on saccade and vergence movement amplitudes, agreeing with our observations in Section 6.1. Quantitatively, the combined model predicts participants' behaviors significantly more accurately, and thus proves the necessity and effectiveness of considering amplitudes of both components of movement.

### 6.2.2 Model Generalizability

We further evaluate generalized goodness-of-fit with unseen data partitions. We create segments of the psychophysical data from Section 6.1 into training-testing groups along multiple axes.

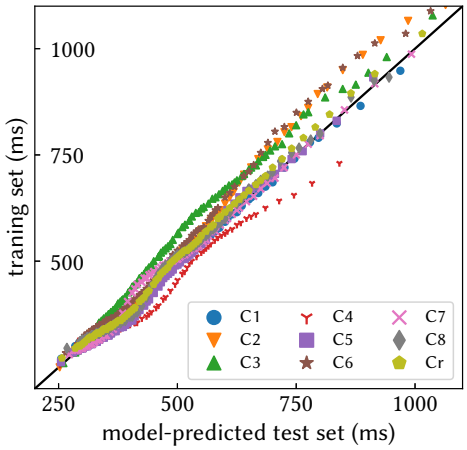
**Metrics.** Similar to prior art on stochastic visual behaviors [Duinkharjav et al., 2022b; Le Meur et al., 2017], we utilize the Kolmogorov-Smirnov (K.S.) goodness-of-fit test [Massey Jr, 1951] between the test set and the corresponding model prediction, using ten quantiles for the offset time.

**Conditions.** We first assess the model's statistical goodness of fit for the full set of psychophysical data from Section 6.1. Then we analyze the model's generalizability based on its capability to successfully fit the statistical distribution with unseen trials or subjects. To this end, the collected dataset is split into two fully separated training and testing sets without overlap. The training set is leveraged to re-train a new model as in Section 6.1.1, which tests the fitness on the corresponding unseen test set. We experiment with two methods of partitions: (1) reserve each one of the eight participants' data as the test set (annotated as $\mathbf{C}_i, i \in \{1, 2, \ldots, 8\}$; (2) uniformly randomly sample 1/8 of the entire data for each condition but across all users (annotated as $\mathbf{C}_r$). For both methods, the remaining data is used as the corresponding training set.

**Results and discussion.** Figure 6.5a shows the results for the goodness-of-fit across all conditions. Additionally in Figure 6.5b, we provide a quantile-quantile (Q-Q) visualization between the training set and the model prediction on the test set: samples closer to the diagonal line indicate better distribution agreement. As a baseline reference, the K.S. test between the model and all collected data shows $D = .1, p = 1$. For all

| | D | p |
|---|---|---|
| $C_1$ | .1 | 1 |
| $C_2$ | .1 | 1 |
| $C_3$ | .2 | .99 |
| $C_4$ | .2 | .99 |
| $C_5$ | .1 | 1 |
| $C_6$ | .1 | 1 |
| $C_7$ | .1 | 1 |
| $C_8$ | .1 | 1 |
| $C_r$ | .1 | 1 |

(a) K.S. statistics




(b) Q-Q visualization

Figure 6.5: *Results of the model generalization evaluation with various partition conditions.* (a) shows the K.S. analysis. The color indicates the corresponding partition condition. (b) shows the Q-Q plot for all conditions, comparing the distributions between the model-prediction on test set vs. training set.

experimented partitioning conditions, the K.S. tests exhibit $p > .99$, failing to reject the null hypothesis that the model prediction acquired from the training set and the unseen test data are drawn from the same distribution. The goodness-of-fit analyses above reveal that our probabilistic model can be generalized to unseen users and trials, implying that it can predict user behavior without observing it in advance.

### 6.2.3 Study: Predicting and Optimizing Visual Performance.

Beyond measuring the performance of the model on data from the controlled experiment (Section 6.1), we further design and conduct a second study with more complex stimuli. We aim to gauge the model's capability to predict and optimize visual performance with realistic VR/AR scenarios, novel conditions, and unseen participants.

**Participants and setup.** We recruited 12 participants (ages 20 – 33, 3 female). To validate the generalizability of the model, we ensured no overlap of participants with the

study from Section 6.1. All participants reported having normal or correct-to-normal vision. We utilized the same hardware and "preamble" checklist as in Section 6.1.

**Scenes and stimuli.** To validate how our model performs for varied scenarios and content, we designed 3 distinct environments: (1) a rendered archery range with a 2D bullseye stimulus (Figure 6.6a), (2) a rendered basketball court with a 3D ball stimulus (Figure 6.6b), and (3) a photographic natural outdoor scene with a virtual bird stimulus to simulate pass-through augmented reality (AR) scenarios (Figure 6.6c).

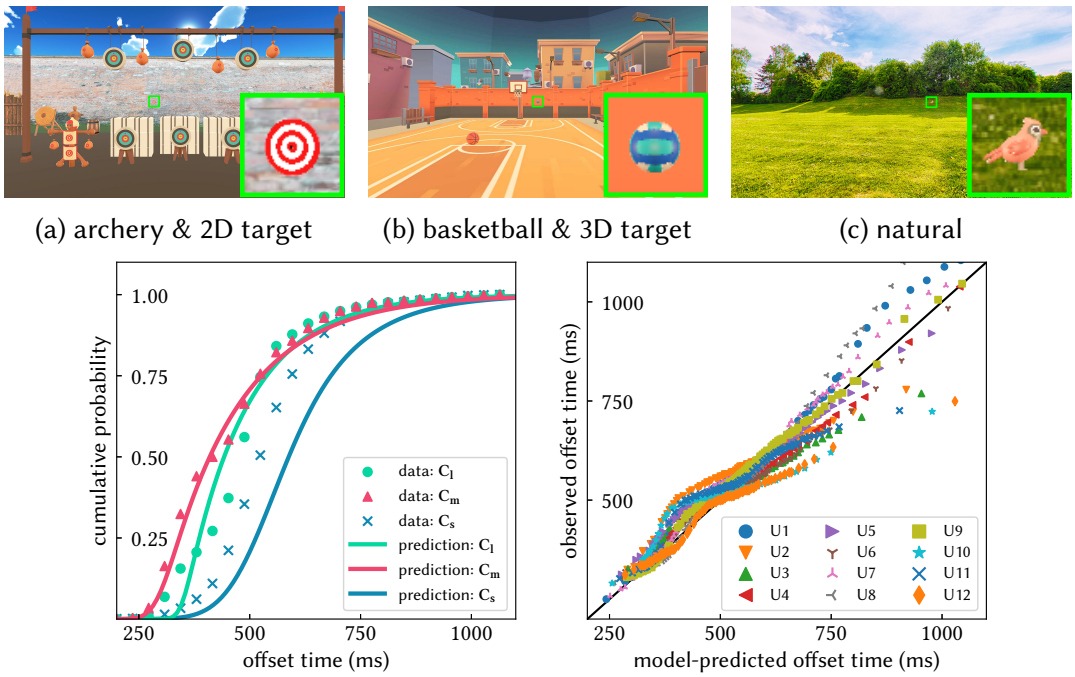
**Tasks.** We instructed participants to complete a target-changing task similar to Section 6.1. During each trial, participants were first instructed to fixate on a cross at the center of the screen. After successfully fixating for 0.4 s, the cross was immediately replaced by one of the three scenes, containing the corresponding target at a new location. The participant then made an eye movement to direct their gaze at the target stimulus. To reduce the influence of progressive learning effects on reaction time, as well as to familiarize the participants with the environment and task, participants performed 36 warm-up trials for each of the scenes, followed by a short break.

**Conditions.** We aim to validate our realistic scenarios with unseen conditions during the model training. Given the hardware limitations in Section 6.1, we experimented with a fixation at 0.4 m and targets placed $\Delta\alpha_v = 6.9°$ away in depth. Using this novel vergence depth, we designed 3 conditions with various eye travel distances:

$\mathbf{C_s}$: pure vergence motion with the **shortest** distance, $\Delta\alpha_s = 0°$,

$\mathbf{C_m}$: combined motion with the **medium** distance $\Delta\alpha_s = 7°$,

$\mathbf{C_l}$: combined motion with the **longest** distance $\Delta\alpha_s = 10.5°$.




(a) archery & 2D target (b) basketball & 3D target (c) natural

(d) by conditions (e) by participants

Figure 6.6: *Evaluation user study scenes and results.* The first row shows the 3 scenes leveraged for the study. The target stimuli are zoomed-in with insets. The second row visualizes the comparisons across various dimensions. (d) compares the model vs. data for the 3 conditions, aggregating all users and scenes. The X-axis/Y-axis indicates offset time/cumulative probability. Note the discrepancy between eye travel distance ($C_s < C_m < C_l$) and landing time ($C_m < C_l < C_s$). Predictions for $C_s$ appear higher than measured data, but are statistically similar (Section 6.2.3). (e) visualizes the model vs. data for each of the participants with a Q-Q plot, aggregating all conditions and scenes. Samples closer to the diagonal line indicate better fitting.

We used the same conditions across all three tested scenes to statistically compare inter-scene generalizability, as detailed in the *results* paragraph below. To acquire enough data for robust statistical distributions, we included 72 repeats per condition on each scene, with fully randomized order. Therefore, the experiment generated 12 participants ×3 scenes ×3 conditions ×72 repeats = 7776 trials in total. We avoided participant fatigue by partitioning the study into 6 blocks, with each block containing trials from only one scene. Additionally, the scene order was fully counterbalanced with a Latin square to avoid carry-on effects.

**Results.** The second row of Figure 6.6 summarizes the results (see Figure 6.13 for the full visualization). To measure the model's applicability and generalizability, we compare its predictions with the obtained human data along multiple axes, including unseen conditions (Figure 6.6d), participants (Figure 6.6e), and scenes. Specifically,

1. Across the 3 conditions, $\mathbf{C_m}$ exhibits the fastest average offset time (.49 ± .16), compared to $\mathbf{C_s}$ (.58 ± .13) and $\mathbf{C_l}$ (.52 ± .13) conditions. The trend agrees with the model's prediction for $\mathbf{C_m}/\mathbf{C_s}/\mathbf{C_l}$, as .44 ± .13/.60 ± .15/.54 ± .16. The predictions for $\mathbf{C_s}$ in Figure 6.6d appear to be slightly higher than measured data, however, K.S. tests failed to reject the null hypothesis that the model prediction and the user-exhibited data are drawn from the same distribution ($p > .99$ for each condition). A repeated measures ANOVA indicated that the condition had a significant effect on the offset time ($F_{2,22} = 21.75, p < .001$).

2. Across the 12 participants, K.S. tests failed to reject the null hypothesis that the model prediction and the user-exhibited data are drawn from the same distribution ($p > .79$ for each).

3. Across the 3 scenes, K.S. tests failed to reject the null hypothesis that the model prediction and the user-exhibited data are drawn from the same distribution ($p > .99$ for each scene). A repeated measures ANOVA did not observe that the scene had a significant effect on the offset time ($F_{2,22} = 1.93, p = .17$). We further calculated the KLdivs between observed data and model predictions for each scene to investigate whether the choice of scene affects model alignment. The KLdiv for archery/basketball/natural is .52 ± .27/.56 ± .29/.54 ± .23, respectively. A repeated measures ANOVA did not observe that scene had a significant effect on the KLdiv ($F_{2,22} = .51, p = .61$).

**Discussion.** The statistical analysis demonstrates the model's consistent capability of predicting and thus optimizing users' task performance during 3D visual target changes. In addition to averaged offset times, the model also accurately predicts probability distributions with statistical accuracy, considering individual differences and sensory/behavioral randomness. Our predictions are consistent with unseen conditions and participants, without being affected by novel and realistic scenes. We also re-observe the remarkable fact that offset time performance is not positively correlated to the travel distance, again evidenced by a significant "U-shape" effect.

## 6.3 Application Case Studies

We apply our model to two applications considering 3D gaze movements. First, we explore how gaze movement variability between VR games can influence video game difficulty experienced by players. Second, we make recommendations for scene-aware design and placement of 3D UI elements to minimize the cost of users' target changing in scenarios such as automotive head-up displays (HUD).

### 6.3.1 Gaze Movement Performance in Games for VR vs. 2D

The relationship between human performance in video games and target placement has been studied in traditional 2D displays [Duinkharjav et al., 2022b; Kim et al., 2022]. In this case study, we consider whether the game-dependent content depth has an effect on this performance. Since gaming in 2D does not involve vergence movements, our evidence in Section 6.1 suggests that gaze movements would be faster than in 3D environments. To measure the scale of this difference across display environments as well as individual games, we conduct a numerical simulation using our model.

**Setup.** We experiment with a large-scale VR player behavior dataset established by Aizenman et al. [2022]. The dataset investigates how often users fixate at various depths during gameplay. It contains games which mimic four top-rated games on Steam[1]: *Job Simulator®*, *Arizona Sunshine®*, *Beat Saber®*, and *Pistol Whip®*. With this data, we can simulate various gaze shifts between fixations $h_{f(ixation)}$ that occur during real gameplay and use our model to predict the corresponding average offset time. Concretely, the distribution of gaze fixation depth is described via a probability density function, $h_f(\alpha_v \mid G)$. The PDF value at some vergence angle, $\alpha_v$, represents the proportion of total time spent fixating at that depth when a user plays a given game $G$.

We model each gaze movement during play as originating and targeting two fixation points sampled from the same distribution $h_f$. Given an origin and target vergence angles, $\alpha_v^o$ and $\alpha_v^t$, the joint probability density, $h_{m(ovement)}(\Delta\alpha_v)$, is equal to

$$h_m(\Delta\alpha_v = \alpha_v^t - \alpha_v^o \mid G) = h_f(\alpha_v^t \mid G) \times h_f(\alpha_v^o \mid G). \tag{6.8}$$

Using this distribution of vergence movement amplitudes, $h_m$, as a weight factor, we compute the mean gaze movement offset times at all saccade amplitudes our model supports (i.e., $\Delta\alpha_s \in [4^\circ, 12^\circ]$).

**Results and discussion.** We visualize our main results in Figure 6.7. Across all gaze depths reported by Aizenman et al. [2022], 98.7% of the duration was fixated at vergence angles $\alpha_v \leq 8.4^\circ$ — the maximum supported by our model. In analysis, we excluded the remaining 1.3% data. The baseline 2D condition without vergence movements between fixations (i.e., $\Delta\alpha_v = 0$) exhibits the fastest offset times of 354 ms. The mean offset times for the four games are, on average, 10 ms slower compared to the baseline 2D condition.

[1] `https://store.steampowered.com/vr/#p=0&tab=TopSellers`

*Job Simulator*® and *Arizona Sunshine*® present a mean gaze offset time of around 20 ms more than baseline, while *Beat Saber*®, and *Pistol Whip*® present a mean gaze offset time of around 5 ms.

The additional time and effort resulting from stereoscopic eye movements in different games will likely translate to increased difficulty. Notably, the performance regression varies across games and depends on the scale of players' gaze depth variance. These results suggest that gaming in VR comes with a "performance overhead" when compared to playing in 2D. Games that feature more salient objects at shallow depths such as *Job Simulator*® and *Arizona Sunshine*® result in up to 20 ms longer gaze offset times compared to the other two games where very little performance is lost. Further investigations to characterize the relationship between gaze offset times and player-experienced difficulties are interesting future work but beyond the scope of this research.

### 6.3.2 Scene-Aware Optimization for 3D User Interface

The surging automotive head-up displays (HUD) and wearable AR devices raise new demands in user-centric 3D interface design. Suboptimal designs may slow users' reactions and cause dangers [Sabelman and Lam, 2015]. When it comes to HUD interface, a desirable design target is the "optimal" virtual projection distance that preserves or even accelerates drivers' reaction to road conditions (see Figure 6.8a), in addition to factors such as focal depths. However, the optimization still remains debated and thus confounds designs. For example, while some literature suggests the distance to be $2.5 - 4$ m [Betancur, 2011], some manufacturers instead designed it as 10 m[2]. Our model provides a quantitative metric for drivers' target-reaching time as a consequence of

[2] https://media.mbusa.com/releases/release-9e110a76b364c518148b9c1ade19bc23-meet-the-s-class-digital-my-mbux-mercedes-benz-user-experience

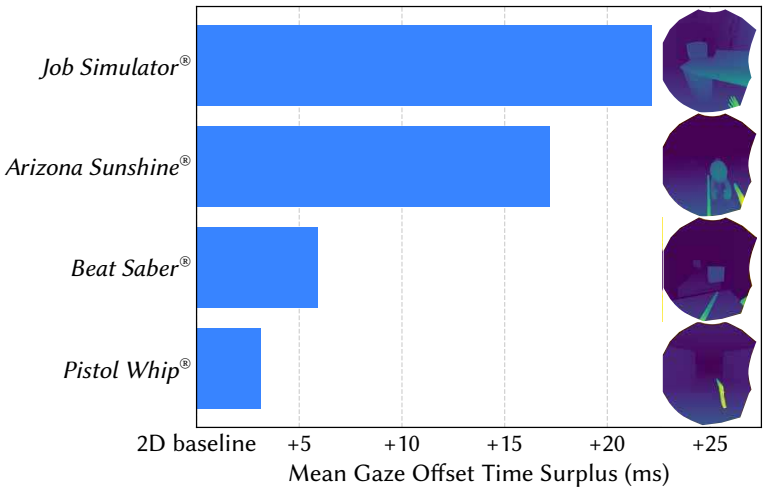
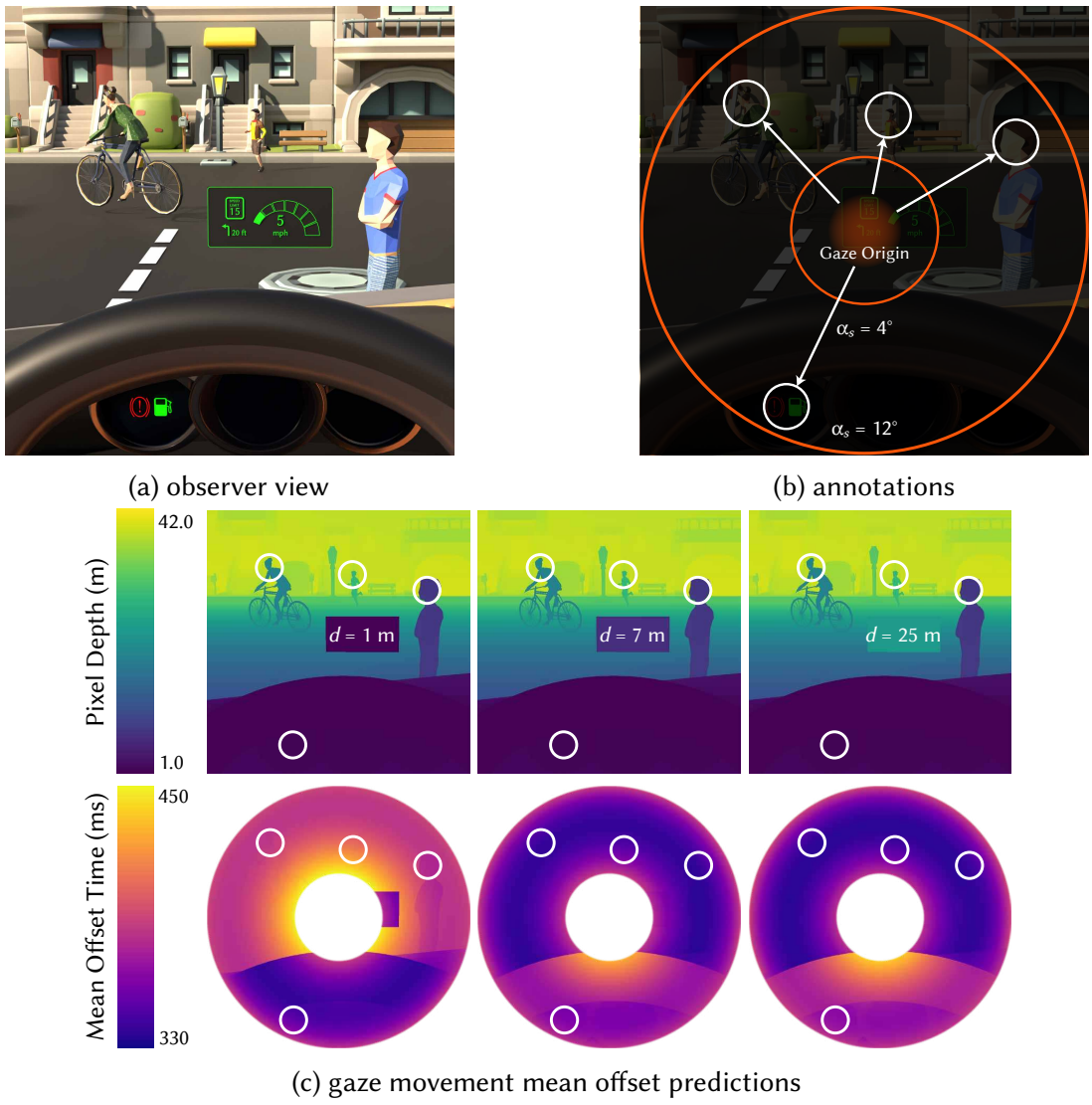



Figure 6.7: *Measuring target-shifting offset times in VR games.* Variability in the depth of salient regions in VR games induces longer gaze movement offset times due to combined vergence-saccade gaze movements. Representative depth-buffer frames from each image are shown as insets for each game. Games with higher variation in depth (*Job Simulator*® and *Arizona Sunshine*®) exhibit longer offset times as predicted by our model. Traditional 2D video games do not involve depth changes during gaze movements, and therefore have a faster average offset time of 354 ms, shown here as a "baseline" for comparison.

varying HUD projection distances.

Specifically, as annotated in Figure 6.8b: if the driver were to initiate a gaze movement from looking at the HUD image, depending on the depth of the UI element as well as the target location, the gaze offset times would vary anywhere between $330 - 450$ ms (Figure 6.8c). Therefore, driving assistant applications could leverage the predictions in gaze offset to adjust the placement of UI elements, or to provide timely intervention/alerts in case of emergencies. While the specific optimization goal for object placement will vary depending on the application, we conducted an example optimization using our model without loss of generality. Specifically, we leverage large-scale datasets to collect the depth distribution of various scenes and suggest the ideal placement of a "HUD overlay image" which would minimize the average gaze offset time from the display element to arbitrary points of focus within the scene.




(a) observer view (b) annotations

(c) gaze movement mean offset predictions

Figure 6.8: *Predicted gaze movement offset times with vehicle HUD projected at various depths.* The offset time varies when a driver shifts their gaze from the green HUD virtual dashboard (a) to different peripheral targets (b), depending on the depth discrepancy between the source and target depths. (c) If the gaze origin is placed at the same depth as the car interior ($d \approx 1$ m), gaze movements towards these locations are faster (346 ms at 1 m compared to 359/365 ms at 7/25 m). In other words, as the depth of the gaze origin moves further ($d \approx 25$ m), the gaze offset towards the car interior begins to increase. However, for the goal of minimizing the offset time required to change gaze to the pedestrian on the right, a medium depth of $d \approx 7$ m is optimal (342 ms at 7 m compared to 376/343 ms at 1/25 m).

Figure 6.9 shows our experimental results with two datasets containing depth maps of natural outdoor environments; DIODE [Vasiljevic et al., 2019] (18, 206 frames), KITTI [Geiger et al., 2012] (12, 919 frames). The average distances of objects are visualized in the top row of the histograms. Assuming a starting gaze centered on a HUD overlay image, positioned at some depth, $d_{HUD}$, we measure the average gaze offset time, $\mathbb{E}[\mathcal{T}]$, for saccade amplitudes uniformly sampled from $\Delta\alpha_s \in [4^\circ, 12^\circ]$, and depth targets sampled from the dataset depth histograms. The resulting relationship between $d_{HUD}$ and $\mathbb{E}[\mathcal{T}]$ is visualized in Figure 6.9. Due to the differentiable nature of our model, we can optimize $d_{HUD}$ to minimize $\mathbb{E}[\mathcal{T}]$ via gradient descent. As a result, the optimal image placements, $d^*_{HUD}$, are 1.8 m and 2.5 m for the outdoor DIODE and KITTI datasets. Beyond HUD in outdoor environments, we may also leverage the model for AR devices in indoor scenarios. Therefore, we further leveraged the indoor portion from DIODE (9, 652 frames), and NYUv2 [Silberman et al., 2012] (407, 024 indoor frames). Intuitively, the depths that minimize $\mathbb{E}[\mathcal{T}]$ are smaller for indoor datasets because more objects are closer in the distance. Indeed, we found 1.3 m to be the optimal projection depths for both the indoor-DIODE and NYUv2 datasets.

Our model helps design HUD displays in various applications, as the optimized image placements clearly vary significantly with scenes, e.g. indoor or outdoor ones. They can also be further optimized by using distributions of saccade amplitudes that are more representative of each application.

## 6.4 Limitations

**Initial depth and eccentricity.** Our combined vergence-saccade model measures the angular displacement in 3D without considering the initial fixation depth and

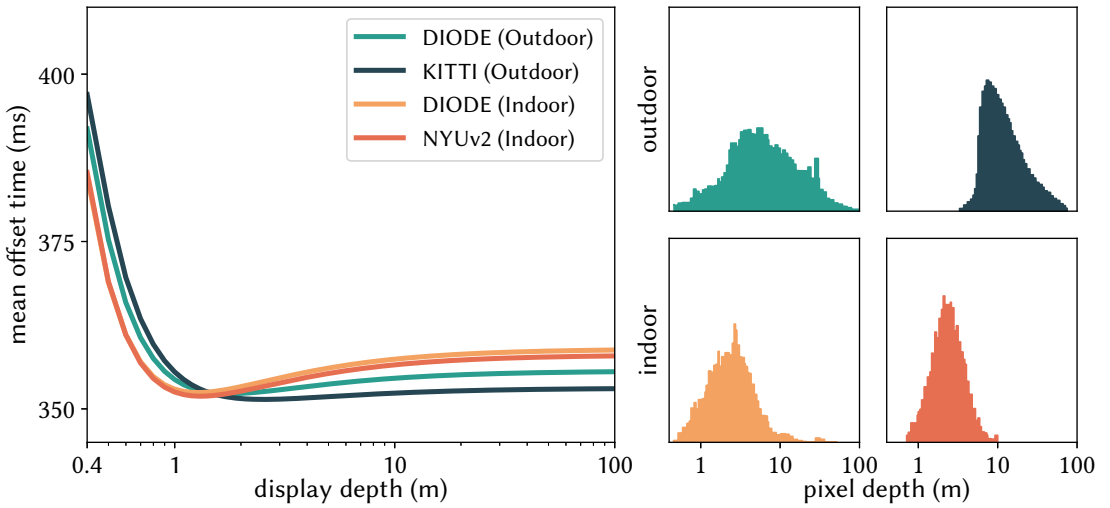



Figure 6.9: *Approximating offset times for VR/AR displays in natural scenes.* (left): By leveraging our model and a variety of large-scale datasets, we measure the average gaze movement offset time ($y$-axis) originating from a HUD or AR display at various projection distances ($x$-axis) towards random locations in a natural 3D environment. We use publicly available datasets containing depth information in indoor and outdoor scenes. (right): shows the statistical density (Y-axis) of each dataset's per-pixel depths (X-axis).

eccentricity, even though both of these factors do influence eye movement offset time. Specifically, prior literature suggests that convergence/divergence-only movements show a linear correlation for offset times [Templin et al., 2014], while off-axis movements that maintain focal depth are much more complex, and require consideration of both vertical/horizontal eccentricity and ocular-motor anatomics [van Beers, 2007]. In order to develop a model that predicts gaze offset times between arbitrary points in 3D space, we would need to individually measure and account for all these factors as a high-dimensional grid of conditions. Our main focus of this research is to demonstrate the importance and possibility of modeling gaze offset times for computer graphics applications; therefore, we plan to investigate all the factors above in future work.

**Influence of accommodation and peripheral stereoacuity.** Vergence accommodation conflict may, in addition to discomfort, also cause incorrect visual fidelity

[March et al., 2022] and depth acuity [Sun et al., 2020], thus potentially degrading target localization accuracy. Similarly, the inherent mismatch between the geometric and empirical horopters may result in poor stereoacuity (and therefore localization) for targets at farther eccentricities along the iso-vergence circle [Ogle, 1952]. Additionally, accommodation speeds have been shown to be slower than vergence speeds [Heron et al., 2001]; hence, while our methods have comprehensive predictive capability in VR and pass-through AR devices (such as the Oculus Quest, and Apple Vision Pro), future investigations are necessary to fully model the latency of accommodation in *see-through* AR devices. Our stimuli cover a conservative range of vergence depths and eccentricities, with targets placed close to where the geometric and empirical horopters meet, and having little to no VAC. While this range is appropriate for the contemporary (vergence-only) VR/AR displays [Aizenman et al., 2022], however, future work on understanding and optimizing for the influence of accommodation on 3D temporal visual behaviors may shed light on new performance-aware metrics to guide 3D display optics design.

**Reaction time and image-space features.** Throughout this work, we eliminated, as much as possible, any image-dependent variance in reaction time. Therefore, our measured offset time is primarily influenced by biomechanical responses to the spatial distribution of the stimuli, and not influenced by task difficulties or image characteristics such as contrast and spatial frequency [Devillez et al., 2020; Lisi et al., 2019]. Exploring the combined effect of cognitive load or image characteristics on reaction time may add new building blocks for comprehensive measurements of visual performance.

**Eye-head coordination.** During free-viewing, head movements often accompany eye movements and we tend to rotate our heads toward visual targets, especially for

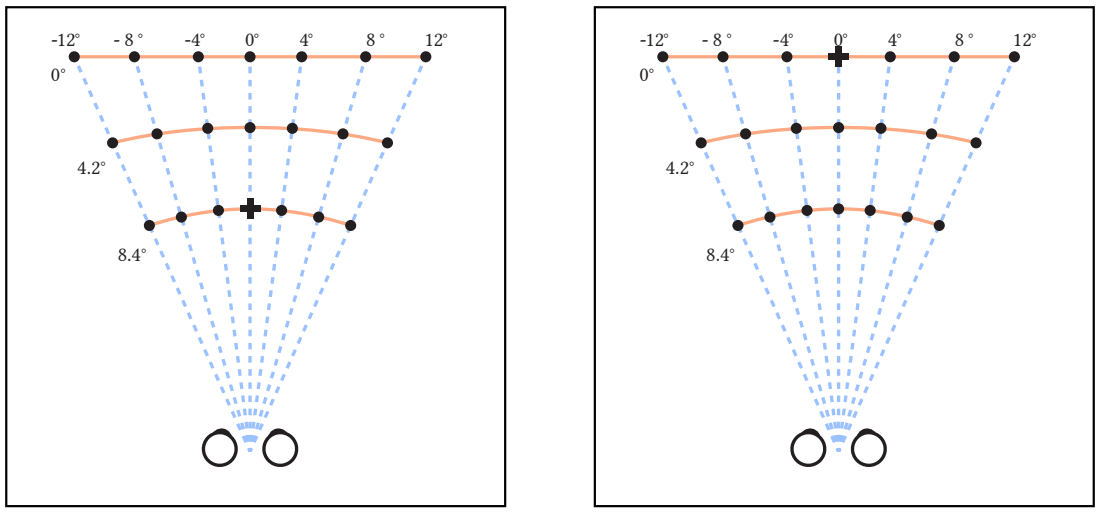
large eccentricities beyond 15° [Bahill, 1975]. Our model does not predict the duration or impact of this concurrent head movement. However, even though moving the head to center the target is a slower movement that typically completes after initial eye movement [Sağlam et al., 2011], our retinal image during the re-centering phase is stabilized, similar to Vestibular Ocular Reflex. Hence, our model's predictions are likely to continue to be useful as they identify the earliest point after initial eye movement at which the target is clearly visible. We hope that future work in eye-head movement validates this expectation.

## 6.A Psychophysical Study Conditions

**Calibration of maximum vergence amplitudes.** The closest depth at which majority of user study participants could fuse a stereo image in VR was approximately $d_{min} = 0.4$ m. Depth, $d$, and vergence angle coordinates, $\alpha_v$, have an inversely proportional relationship,

$$\alpha_v = \arctan\left(\frac{w_{IPD}}{2d}\right), \tag{6.9}$$

which varies from person to person depending on their IPD, $w_{IPD}$. This relationship, and the fact that there are no negative vergence angle coordinates, effectively limits the range of vergence gaze movement amplitudes, $\Delta\alpha_v$, a user study participant can make. Crucially, since the IPD, $w_{IPD}$, of participants varied, and we couldn't foresee the IPDs of all future user study participants, we could not determine the maximum vergence angle coordinate, $\alpha_v^{max}$, by applying Equation (6.9) naively. Therefore, to ensure consistency across different participants, we selected the most conservative value of maximum vergence angle coordinates by minimizing Equation (6.9) under




(a) divergent (b) convergent

Figure 6.10: *Study conditions.* All visualized conditions originate at a + sign (near for divergent, far for convergent conditions), and target · signs. Leftward and rightward saccades are treated as equivalent in data analysis, but there are equal number of leftward and rightward conditions implemented.

the constraints of $d > d_{min} = 0.4$ m, and $w_{IPD} > w_{IPD}^{min} = 59$ mm — the minimum IPD supported by the HMD. Then, applying these edge conditions to Equation (6.9), we get our maximum vergence angle coordinate of $\alpha_v^{max} = 8.4°$.

**Implementation of Study Conditions.** We construct three isovergence circles for each $\alpha_v^{init} + \Delta\alpha_v$, starting with the smallest. As established earlier, this circle must be at least $d_{min}$ away from the observer. Therefore we pick the first isovergence circle to be $d^{(0)} = d_{min}$ away, which corresponds to a vergence angle coordinate equal to

$$\alpha_v^{(0)} = \arctan\left(\frac{w_{IPD}}{2d^{(0)}}\right). \tag{6.10}$$

The following circles are constructed by adding the $\Delta\alpha_v$ to $\alpha_v^{(0)}$:

$$\alpha_v^{(i)} = \alpha_v^{(0)} + \Delta\alpha_v^{(i-1)}, \text{ for } i \in \{1, 2\}, \tag{6.11}$$

where $\Delta\alpha_v^{(i-1)}$ is the $i-1$th condition among vergence conditions.

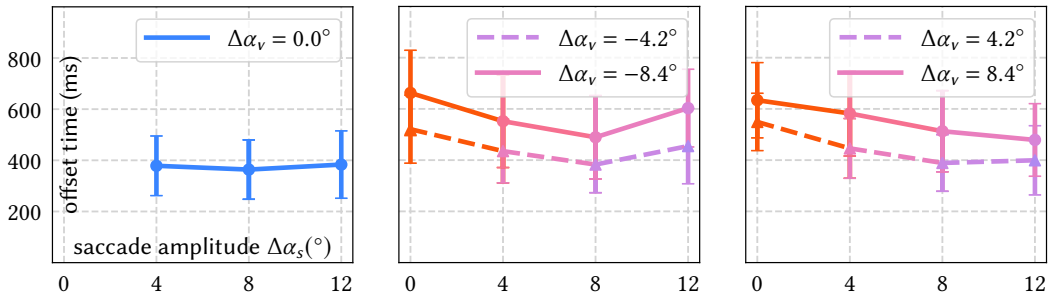
Equipped with the isovergence circles with angles $\{\alpha_v^{(i)}\}$ for $i \in \{0, 1, 2\}$, we can select the initial fixation point for all divergent and convergent gaze motions to be at coordinates

$$\begin{aligned}(\alpha_v^{init,\ div}, \alpha_s^{init,\ div}) &= (\alpha_v^{(0)}, 0^\circ) \\ (\alpha_v^{init,\ conv}, \alpha_s^{init,\ conv}) &= (\alpha_v^{(2)}, 0^\circ),\end{aligned} \tag{6.12}$$

respectively. Originating from a given fixation point, the rest of the condition locations are found as

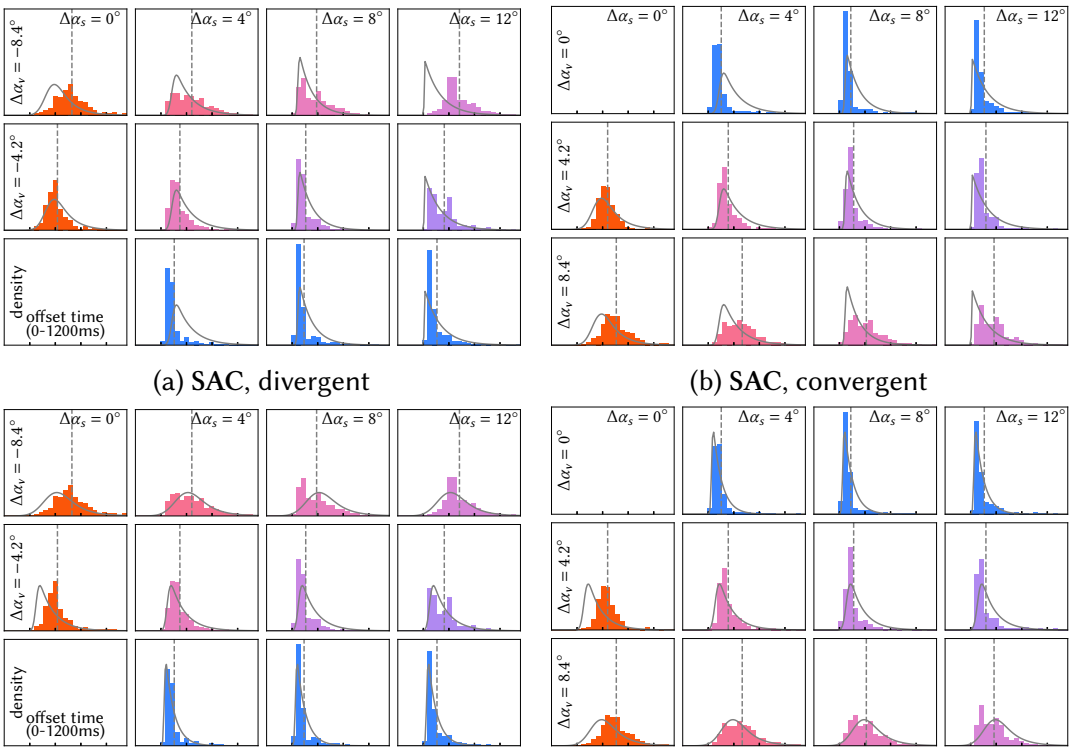
$$(\alpha_v, \alpha_s) = (\alpha_v^{init} + \Delta\alpha_v, \alpha_s^{init} + \Delta\alpha_s), \tag{6.13}$$

where $\Delta\alpha_v$ and $\Delta\alpha_s$ correspond to the specific experimental condition of interest. The resulting grid of conditions are visualized in Figure 6.10.




(a) saccade (b) divergent (c) convergent

Figure 6.11: *Aggregated mean offset time of studied conditions across all participants with error bars.* This is a version of Figure 6.3 with std error bars as a more detailed visualization. See Figure 6.3 for further details.




(a) **SAC**, divergent (b) **SAC**, convergent

(c) **VER**, divergent (d) **VER**, convergent

Figure 6.12: *Histograms vs. predicted distributions of ablation models.* Predicted distributions by the ablation models are compared to measured data from psychophysical study. Ablation model **SAC** was trained using only saccade amplitude information from the study data, while **VER** only used vergence amplitude information. Since either model does not have full information that distinguishes individual conditions within a single column and row respectively, the models make the same predictions across multiple conditions within this histogram visualization. Thus, in (a)/(b) the model makes the same prediction within the same columns, while in (c)/(d) the model makes the same predictions within the same rows.

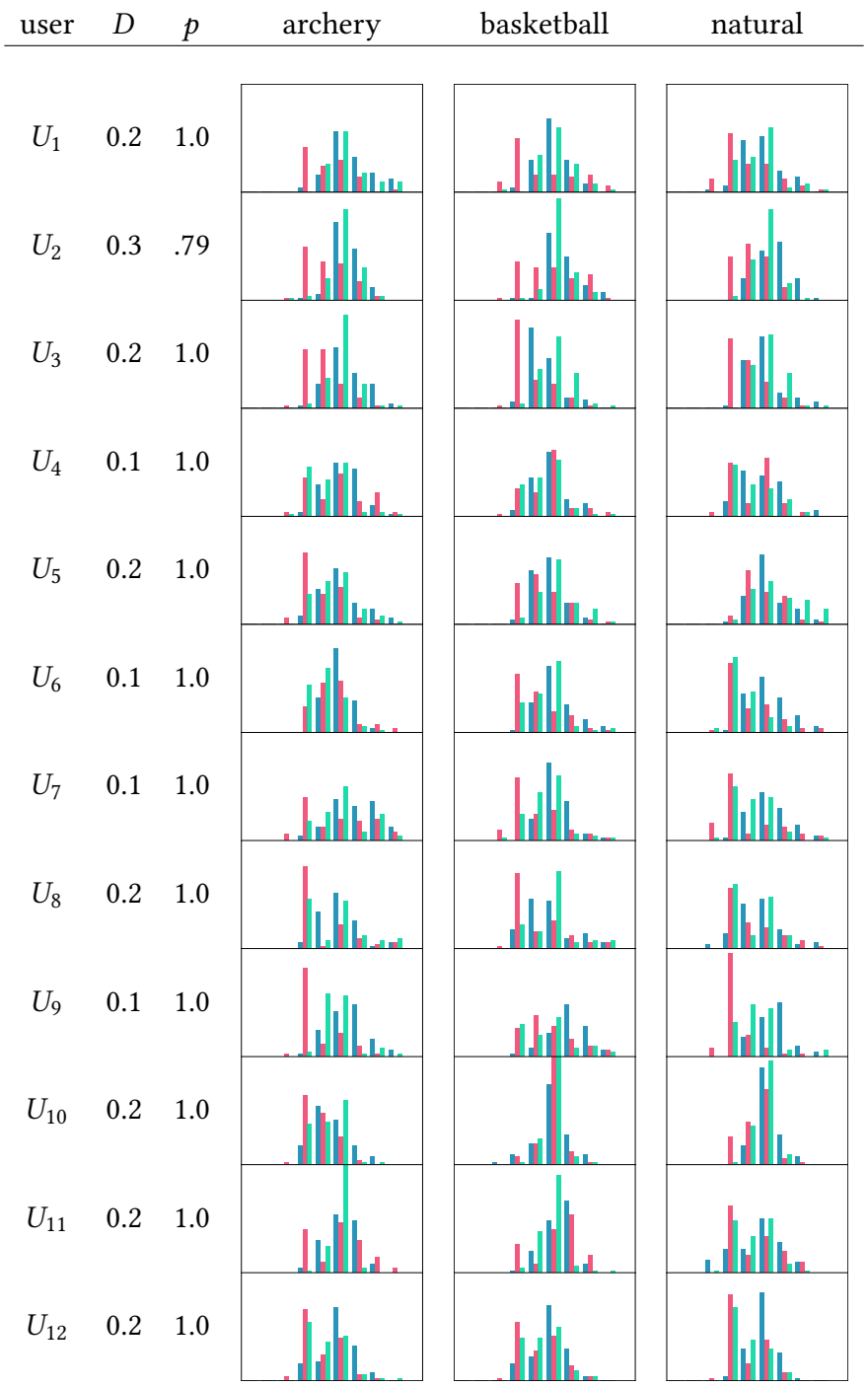



Figure 6.13: *Responses across all participants, conditions and scenes of Section 6.2.3.* Plots visualize the histogram of gaze offset times (0-1000 ms). Blue/red/green bars represent $\mathbf{C_s}$ : $\Delta\alpha_v = 0°$ / $\mathbf{C_m}$ : $\Delta\alpha_v = 7°$ / $\mathbf{C_l}$ : $\Delta\alpha_v = 10.5°$. K.S. test results are shown in the $D$, and $p$ columns.

# Chapter 7

# Conclusion

Understanding the interaction between humans and computer systems has long been a fundamental aspect of computer graphics. Effectively quantifying human perception across diverse contexts is essential for addressing these challenges. In this dissertation, we have demonstrated numerous instances where psychophysical techniques were applied to model human perception and behavior. By incorporating perceptual and behavioral models of human color sensitivity, motion estimation mechanisms, decision-making, and eye control timing, we systematically designed computer systems that align with human strengths and limitations. Throughout our work, we proposed optimization strategies such as adjusting the luminance and chromaticity characteristics of visual content, refining the position and velocity of visual targets, and determining optimal display system configurations and eye-display alignments. In turn, we have shown that these strategies can enhance reaction times and accuracy in human performance while also improving power efficiency and bandwidth utilization in computer systems. We hope this work serves as a reference for contemporary research in perceptual graphics and as a foundation for future studies aimed at advancing human-aware system design.

# Bibliography

Aizenman, A. M., Koulieris, G. A., Gibaldi, A., Sehgal, V., Levi, D. M., and Banks, M. S. (2022). The statistics of eye movements and binocular disparities during vr gaming: Implications for headset design. *ACM Transactions on Graphics (TOG)*, 42(1).

Albert, R., Patney, A., Luebke, D., and Kim, J. (2017). Latency requirements for foveated rendering in virtual reality. *ACM Transactions on Applied Perception*, 14(4).

Arabadzhiyska, E., Tursun, O. T., Myszkowski, K., Seidel, H.-P., and Didyk, P. (2017). Saccade landing position prediction for gaze-contingent rendering. *ACM Transactions on Graphics (TOG)*, 36(4):1–12.

Ashraf, M., Mantiuk, R. K., Chapiro, A., and Wuerger, S. (2024). castlecsf—a contrast sensitivity function of color, area, spatiotemporal frequency, luminance and eccentricity. *Journal of Vision*, 24(4):5–5.

Bahill, A., Clark, M. R., and Stark, L. (1975a). The main sequence, a tool for studying human eye movements. *Mathematical Biosciences*, 24(3):191–204.

Bahill, A. T. (1975). Most naturally occurring human saccades have magnitudes of 15 deg or less. *Invest. Ophthalmol*, 14:468–469.

Bahill, A. T., Clark, M. R., and Stark, L. (1975b). The main sequence, a tool for studying human eye movements. *Mathematical biosciences*, 24(3-4):191–204.

Barten, P. G. (1999a). *Contrast sensitivity of the human eye and its effects on image quality*. SPIE press.

Barten, P. G. (1999b). *Contrast sensitivity of the human eye and its effects on image quality*. SPIE press.

Becker, W. and Fuchs, A. (1969). Further properties of the human saccadic system: Eye movements and correction saccades with and without visual fixation points. *Vision Research*, 9(10):1247–1258.

Becker, W. and Jürgens, R. (1979). An analysis of the saccadic system by means of double step stimuli. *Vision research*, 19(9):967–983.

Bell, A., Meredith, M., Van Opstal, A., and Munoz, D. (2006). Stimulus intensity modifies saccadic reaction time and visual response latency in the superior colliculus. *Experimental Brain Research*, 174(1):53–59.

Bernal-Berdun, E., Vallejo, M., Sun, Q., Serrano, A., and Gutierrez, D. (2024). Modeling the impact of head-body rotations on audio-visual spatial perception for virtual reality applications. *IEEE Transactions on Visualization and Computer Graphics*.

Betancur, J. A. (2011). Physical variable analysis involved in head-up display systems applied to automobiles. *Augmented Reality-Some Emerging Application Areas*, 13:244–266.

Blackwell, H. R. (1946). Contrast thresholds of the human eye. *Journal of the Optical Society of America*, 36(11):624–643.

Blake, R. and Shiffrar, M. (2007). Perception of human motion. *Annu. Rev. Psychol.*, 58:47–73.

Blohm, G. and Lefèvre, P. (2010). Visuomotor velocity transformations for smooth pursuit eye movements. *Journal of neurophysiology*, 104(4):2103–2115.

Bohr, M. (2007). A 30 year retrospective on dennard's mosfet scaling paper. *IEEE Solid-State Circuits Society Newsletter*, 12(1):11–13.

Boroson, M., Ludwicki, J., and Murdoch, M. (US Patent 7,586,497, Sep. 8, 2009). Oled display with improved power performance.

Bowmaker, J. K. and Dartnall, H. (1980). Visual pigments of rods and cones in a human retina. *The Journal of physiology*, 298(1):501–511.

Braddick, O. J., O'Brien, J. M., Wattam-Bell, J., Atkinson, J., Hartley, T., and Turner, R. (2001). Brain areas sensitive to coherent visual motion. *Perception*, 30(1):61–72.

Braun, D. I., Schütz, A. C., and Gegenfurtner, K. R. (2017). Visual sensitivity for luminance and chromatic stimuli during the execution of smooth pursuit and saccadic eye movements. *Vision Research*, 136:57–69.

Bruce, V., Georgeson, M. A., and Green, P. R. (2014). *Visual perception: Physiology, psychology and ecology*. Psychology Press.

Brühlmann, F., Petralito, S., Aeschbach, L. F., and Opwis, K. (2020). The quality of data collected online: An investigation of careless responding in a crowdsourced sample. *Methods in Psychology*, 2:100022.

Bucci, M. P., Yang, Q., Brémond-Gignac, D., et al. (2006). Latency of saccades, vergence, and combined movements in children with early onset convergent or divergent strabismus. *Vision Research*, 46(8-9):1384–1392.

Burlingham, C. S. and Heeger, D. J. (2020). Heading perception depends on time-varying evolution of optic flow. *Proceedings of the National Academy of Sciences*, 117(52):33161–33169.

Burr, D., Morrone, M., and Ross, J. (1994). Selective suppression of the magnocellular visual pathway during saccadic eye movements. *Nature*, 371(6497):511–513.

Butterworth, S. et al. (1930). On the theory of filter amplifiers. *Wireless Engineer*, 7(6):536–541.

Cai, Y., Bozorgian, A., Ashraf, M., Wanat, R., and Mantiuk, K. R. (2024). elatcsf: A temporal contrast sensitivity function for flicker detection and modeling variable refresh rate flicker. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11.

Cajar, A., Engbert, R., and Laubrock, J. (2016). Spatial frequency processing in the central and peripheral visual field during scene viewing. *Vision Research*, 127:186–197.

Campbell, F. W. and Robson, J. G. (1968). Application of fourier analysis to the visibility of gratings. *The Journal of physiology*, 197(3):551.

Caroux, L., Le Bigot, L., and Vibert, N. (2013). Impact of the motion and visual complexity of the background on players' performance in video game-like displays. *Ergonomics*, 56(12):1863–1876.

Carpenter, R. (2004). Contrast, probability, and saccadic latency: evidence for independence of detection and decision. *Current Biology*, 14(17):1576–1580.

Carpenter, R. H. and Williams, M. (1995). Neural computation of log likelihood in control of saccadic eye movements. *Nature*, 377(6544):59–62.

Chatterjee, S. and Hadi, A. S. (2015). *Regression analysis by example*. John Wiley & Sons.

Chen, H., Mori, Y., and Matsuba, I. (2014). Solving the balance problem of massively multiplayer online role-playing games using coevolutionary programming. *Applied Soft Computing*, 18:1–11.

Chen, H.-W., Zhu, R.-D., He, J., Duan, W., Hu, W., Lu, Y.-Q., Li, M.-C., Lee, S.-L., Dong, Y.-J., and Wu, S.-T. (2017). Going beyond the limit of an lcd's color gamut. *Light: Science & Applications*, 6(9):e17043–e17043.

Chen, S., Duinkharjav, B., Sun, X., Wei, L.-Y., Petrangeli, S., Echevarria, J., Silva, C., and Sun, Q. (2022). Instant reality: Gaze-contingent perceptual optimization for 3d virtual reality streaming. *IEEE Transactions on Visualization and Computer Graphics*, 28(5):2157–2167.

Chinazzo, G., Chamilothori, K., Wienold, J., and Andersen, M. (2021). Temperature–color interaction: subjective indoor environmental perception and physiological responses in virtual reality. *Human Factors*, 63(3):474–502.

Claypool, M., Claypool, K., and Damaa, F. (2006). The effects of frame rate and resolution on users playing first person shooter games. In *Multimedia computing and networking 2006*, volume 6071, page 607101. SPIE.

Cohen, M. A., Botch, T. L., and Robertson, C. E. (2020). The limits of color awareness during active, real-world vision. *Proceedings of the National Academy of Sciences*, 117(24):13821–13827.

Collewijn, H., Erkelens, C. J., and Steinman, R. M. (1995). Voluntary binocular gaze-shifts in the plane of regard: dynamics of version and vergence. *Vision research*, 35(23-24):3335–3358.

Conway, B. R., Eskew Jr, R. T., Martin, P. R., and Stockman, A. (2018). A tour of contemporary color vision research. *Vision research*, 151:2–6.

Cornilleau-Pérès, V. and Gielen, C. (1996). Interactions between self-motion and depth perception in the processing of optic flow. *Trends in Neurosciences*, 19(5):196–202.

Cotti, J., Panouilleres, M., Munoz, D. P., Vercher, J.-L., Pélisson, D., and Guillaume, A. (2009). Adaptation of reactive and voluntary saccades: different patterns of adaptation revealed in the antisaccade task. *The Journal of Physiology*, 587(1):127–138.

Coubard, O. A. (2013). Saccade and vergence eye movements: a review of motor and premotor commands. *European journal of neuroscience*, 38(10):3384–3397.

Cullen, K. E. and Van Horn, M. R. (2011). The neural control of fast vs. slow vergence eye movements. *European Journal of Neuroscience*, 33(11):2147–2154.

Daly, S. J. (1992). Visible differences predictor: an algorithm for the assessment of image fidelity. In *Human Vision, Visual Processing, and Digital Display III*, volume 1666, pages 2–15. International Society for Optics and Photonics.

Daniel, P. and Whitteridge, D. (1961). The representation of the visual field on the cerebral cortex in monkeys. *The Journal of physiology*, 159(2):203.

Danz, A. D., Angelaki, D. E., and DeAngelis, G. C. (2020). The effects of depth cues and vestibular translation signals on the rotation tolerance of heading tuning in macaque area mstd. *Eneuro*, 7(6).

Dartnall, H. J., Bowmaker, J. K., and Mollon, J. D. (1983). Human visual pigments: microspectrophotometric results from the eyes of seven persons. *Proceedings of the Royal society of London. Series B. Biological sciences*, 220(1218):115–130.

Dash, P. and Hu, Y. C. (2021). How much battery does dark mode save? an accurate oled display power profiler for modern smartphones. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*, MobiSys '21, page 323–335, New York, NY, USA. Association for Computing Machinery.

De Brouwer, S., Yuksel, D., Blohm, G., Missal, M., and Lefèvre, P. (2002). What triggers catch-up saccades during visual tracking? *Journal of neurophysiology*, 87(3):1646–1650.

De Valois, R. L., Abramov, I., and Jacobs, G. H. (1966). Analysis of response patterns of lgn cells. *JOSA*, 56(7):966–977.

DeAngelis, G. C. and Angelaki, D. E. (2012). Visual–vestibular integration for self-motion perception. *The Neural Bases of Multisensory Processes*.

Debattista, K., Bugeja, K., Spina, S., Bashford-Rogers, T., and Hulusic, V. (2018). Frame rate vs resolution: A subjective evaluation of spatiotemporal perceived quality under varying computational budgets. In *Computer Graphics Forum*, volume 37, pages 363–374. Wiley Online Library.

Denes, G., Jindal, A., Mikhailiuk, A., and Mantiuk, R. K. (2020). A perceptual model of motion quality for rendering with adaptive refresh-rate and resolution. *ACM Transactions on Graphics (TOG)*, 39(4):133–1.

Deng, N., He, Z., Ye, J., Duinkharjav, B., Chakravarthula, P., Yang, X., and Sun, Q. (2022). Fov-nerf: Foveated neural radiance fields for virtual reality. *IEEE Transactions on Visualization and Computer Graphics*, 28(11):3854–3864.

Derrington, A. M., Krauskopf, J., and Lennie, P. (1984a). Chromatic mechanisms in lateral geniculate nucleus of macaque. *The Journal of physiology*, 357(1):241–265.

Derrington, A. M., Krauskopf, J., and Lennie, P. (1984b). Chromatic mechanisms in lateral geniculate nucleus of macaque. *The Journal of physiology*, 357(1):241–265.

Deubel, H., Wolf, W., and Hauske, G. (1982). Corrective saccades: Effect of shifting the saccade goal. *Vision Research*, 22(3):353–364.

Devillez, H., Guyader, N., Curran, T., and O'Reilly, R. C. (2020). The bimodality of saccade duration during the exploration of visual scenes. *Visual Cognition*, 28(9):484–512.

Diamond, M. R., Ross, J., and Morrone, M. C. (2000). Extraretinal control of saccadic suppression. *Journal of Neuroscience*, 20(9):3449–3455.

Didyk, P., Ritschel, T., Eisemann, E., Myszkowski, K., and Seidel, H.-P. (2011). A perceptual model for disparity. *ACM Transactions on Graphics (TOG)*, 30(4):1–10.

Dobrow, R. P. (2016). *Introduction to stochastic processes with R*. John Wiley & Sons.

Dokka, K., Park, H., Jansen, M., DeAngelis, G. C., and Angelaki, D. E. (2019). Causal inference accounts for heading perception in the presence of object motion. *Proceedings of the National Academy of Sciences*, 116(18):9060–9065.

Dong, M., Choi, Y.-S. K., and Zhong, L. (2009). Power modeling of graphical user interfaces on oled displays. In *2009 46th ACM/IEEE Design Automation Conference*, pages 652–657. IEEE.

Dong, M. and Zhong, L. (2011). Chameleon: A color-adaptive web browser for mobile oled displays. In *Proceedings of the 9th international conference on Mobile systems, applications, and services*, pages 85–98.

Duchowski, A. T., House, D. H., Gestring, J., Wang, R. I., Krejtz, K., Krejtz, I., Mantiuk, R., and Bazyluk, B. (2014). Reducing visual discomfort of 3d stereoscopic displays with gaze-contingent depth-of-field. In *Proceedings of the ACM Symposium on Applied Perception*, SAP '14, page 39–46, New York, NY, USA. Association for Computing Machinery.

Duinkharjav, B., Chakravarthula, P., Brown, R., Patney, A., and Sun, Q. (2022a). Image features influence reaction time: A learned probabilistic perceptual model for saccade latency. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 41(4):144:1–144:15.

Duinkharjav, B., Chakravarthula, P., Brown, R., Patney, A., and Sun, Q. (2022b). Image features influence reaction time: A learned probabilistic perceptual model for saccade latency. *ACM Trans. Graph.*, 41(4).

Dunn, D., Tursun, O., Yu, H., Didyk, P., Myszkowski, K., and Fuchs, H. (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)*, pages 90–100. IEEE.

Durst, D., Xie, F., Sarukkai, V., Shacklett, B., Frosio, I., Tessler, C., Kim, J., Taylor, C., Bernstein, G., Choudhury, S., et al. (2024). Learning to move like professional counter-strike players. In *Computer Graphics Forum*, volume 43, page e15173. Wiley Online Library.

Engbert, R. and Mergenthaler, K. (2006). Microsaccades are triggered by low retinal image slip. *Proceedings of the National Academy of Sciences*, 103(18):7192–7197.

Erkelens, C., Steinman, R., and Collewijn, H. (1989). Ocular vergence under natural conditions. ii. gaze shifts between real targets differing in distance and direction. *Proceedings of the Royal Society of London. B. Biological Sciences*, 236(1285):441–465.

Fabius, J. H., Fracasso, A., Nijboer, T. C., and Van der Stigchel, S. (2019). Time course of spatiotopic updating across saccades. *Proceedings of the National Academy of Sciences*, 116(6):2027–2032.

Fairchild, M. D. (2013). *Color appearance models*. John Wiley & Sons.

Fairchild, M. D. and Reniff, L. (1995). Time course of chromatic adaptation for color-appearance judgments. *JOSA A*, 12(5):824–833.

Fairman, H. S., Brill, M. H., and Hemmendinger, H. (1997). How the cie 1931 color-matching functions were derived from wright-guild data. *Color Research & Application*, 22(1):11–23.

Feil, M., Moser, B., and Abegg, M. (2017). The interaction of pupil response with the vergence system. *Graefe's archive for clinical and experimental ophthalmology*, 255:2247–2253.

Fesharaki, H., Rezaei, L., Farrahi, F., Banihashem, T., and Jahanbakhshi, A. (2012). Normal interpupillary distance values in an iranian population. *Journal of ophthalmic & vision research*, 7(3):231.

Fetsch, C. R., Turner, A. H., DeAngelis, G. C., and Angelaki, D. E. (2009). Dynamic reweighting of visual and vestibular cues during self-motion perception. *Journal of Neuroscience*, 29(49):15601–15612.

Folks, J. L. and Chhikara, R. S. (1978). The inverse gaussian distribution and its statistical application—a review. *Journal of the Royal Statistical Society: Series B (Methodological)*, 40(3):263–275.

Franke, L., Fink, L., Martschinke, J., Selgrad, K., and Stamminger, M. (2021). Time-warped foveated rendering for virtual reality headsets. *Computer Graphics Forum*, 40(1):110–123.

Freeman, J. and Simoncelli, E. P. (2011). Metamers of the ventral stream. *Nature neuroscience*, 14(9):1195–1201.

Fudenberg, D., Newey, W., Strack, P., and Strzalecki, T. (2020). Testing the drift-diffusion model. *Proceedings of the National Academy of Sciences*, 117(52):33141–33148.

Furht, B., Greenberg, J., and Westwater, R. (2012). *Motion estimation algorithms for video compression*, volume 379. Springer Science & Business Media.

Geiger, A., Lenz, P., and Urtasun, R. (2012). Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE.

Gibaldi, A. and Banks, M. S. (2019). Binocular eye movements are adapted to the natural environment. *Journal of Neuroscience*, 39(15):2877–2888.

Gibaldi, A. and Sabatini, S. P. (2021). The saccade main sequence revised: A fast and repeatable tool for oculomotor analysis. *Behavior Research Methods*, 53(1):167–187.

Gnanadesikan, R. and Wilk, M. B. (1968). Probability plotting methods for the analysis of data. *Biometrika*, 55(1):1–17.

Goodale, M. A. and Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in neurosciences*, 15(1):20–25.

Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al. (2022). Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012.

Groen, I. I., Silson, E. H., and Baker, C. I. (2017). Contributions of low-and high-level properties to neural processing of visual scenes in the human brain. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1714):20160102.

Guan, P., Mercier, O., Shvartsman, M., and Lanman, D. (2022). Perceptual requirements for eye-tracked distortion correction in vr. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–8.

Guan, P., Penner, E., Hegland, J., Letham, B., and Lanman, D. (2023). Perceptual requirements for world-locked rendering in ar and vr. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–10.

Guenter, B., Finch, M., Drucker, S., Tan, D., and Snyder, J. (2012). Foveated 3d graphics. *ACM Transactions on Graphics (TOG)*, 31(6):1–10.

Guild, J. (1931). The colorimetric properties of the spectrum. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 230(681-693):149–187.

Gupta, A., Bansal, R., Alashwal, H., Kacar, A. S., Balci, F., and Moustafa, A. A. (2022). Neural substrates of the drift-diffusion model in brain disorders. *Frontiers in computational neuroscience*, 15:678232.

Hainich, R. R. and Bimber, O. (2016). *Displays: fundamentals & applications*. AK Peters/CRC Press.

Halpern, M., Zhu, Y., and Reddi, V. J. (2016). Mobile cpu's rise to power: Quantifying the impact of generational mobile cpu design trends on performance, energy, and user satisfaction. In *2016 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, pages 64–76. IEEE.

Hansen, T., Giesel, M., and Gegenfurtner, K. R. (2008). Chromatic discrimination of natural objects. *Journal of Vision*, 8(1):2–2.

Hansen, T., Pracejus, L., and Gegenfurtner, K. R. (2009). Color perception in the intermediate periphery of the visual field. *Journal of vision*, 9(4):26–26.

Hartmann, E., Lachenmayr, B., and Brettel, H. (1979). The peripheral critical flicker frequency. *Vision Research*, 19(9):1019–1023.

Hassani, N. and Murdoch, M. J. (2016). Color appearance modeling in augmented reality. In *Proceedings of the ACM Symposium on Applied Perception*, pages 132–132.

Hatada, T., Sakata, H., and Kusaka, H. (1980). Psychophysical analysis of the "sensation of reality" induced by a visual wide-field display. *Smpte Journal*, 89(8):560–569.

Hautus, M. J., Macmillan, N. A., and Creelman, C. D. (2021). *Detection theory: A user's guide*. Routledge.

Henderson, M. M., Tarr, M. J., and Wehbe, L. (2023). A texture statistics encoding model reveals hierarchical feature selectivity across human visual cortex. *Journal of Neuroscience*, 43(22):4144–4161.

Hermann, K. L., Singh, S. R., Rosenthal, I. A., Pantazis, D., and Conway, B. R. (2021). Temporal dynamics of the neural representation of hue and luminance contrast. *BioRxiv*, pages 2020–06.

Heron, G., Charman, W., and Schor, C. (2001). Dynamics of the accommodation response to abrupt changes in target vergence as a function of age. *Vision research*, 41(4):507–519.

Hillaire, S., Lecuyer, A., Cozot, R., and Casiez, G. (2008). Using an eye-tracking system to improve camera motions and depth-of-field blur effects in virtual environments. In *2008 IEEE Virtual Reality Conference*, pages 47–50.

Hoffman, D. M., Girshick, A. R., Akeley, K., and Banks, M. S. (2008). Vergence–accommodation conflicts hinder visual performance and cause visual fatigue. *Journal of vision*, 8(3):33–33.

Hore, A. and Ziou, D. (2010). Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE.

Howard, I. P. and Howard, A. (1994). Vection: the contributions of absolute and relative visual motion. *Perception*, 23(7):745–751.

Hsu, W.-H., Zhang, Y., and Ma, K.-L. (2013). A multi-criteria approach to camera motion design for volume data animation. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2792–2801.

Hu, P., Sun, Q., Didyk, P., Wei, L.-Y., and Kaufman, A. E. (2019). Reducing simulator sickness with perceptual camera control. *ACM Transactions on Graphics (TOG)*, 38(6):1–12.

Huang, Y., Hsiang, E.-L., Deng, M.-Y., and Wu, S.-T. (2020). Mini-led, micro-led and oled displays: Present status and future perspectives. *Light: Science & Applications*, 9(1):1–16.

Huang, Y., Palaniappan, K., Zhuang, X., and Cavanaugh, J. E. (1995). Optic flow field segmentation and motion estimation using a robust genetic partitioning algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(12):1177–1190.

Hummel, N., Cuturi, L. F., MacNeilage, P. R., and Flanagin, V. L. (2016). The effect of supine body position on human heading perception. *Journal of vision*, 16(3):19–19.

Ibbotson, M. R. and Cloherty, S. L. (2009). Visual perception: Saccadic omission—suppression or temporal masking? *Current Biology*, 19(12):R493–R496.

In, J. (2017). Introduction of a pilot study. *Korean journal of anesthesiology*, 70(6):601–605.

Jain, R. (1983). Direct computation of the focus of expansion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-5(1):58–64.

Jain, R. (1984). Complex logarithmic mapping and the focus of expansion. *ACM SIGGRAPH Computer Graphics*, 18(1):24–24.

Jameson, D. and Hurvich, L. M. (1955). Some quantitative aspects of an opponent-colors theory. i. chromatic responses and spectral saturation. *JOSA*, 45(7):546–552.

Jarvenpaa, S. L. (1990). Graphic displays in decision making—the visual salience effect. *Journal of Behavioral Decision Making*, 3(4):247–262.

Jaschinski, W. (2016). Pupil size affects measures of eye position in video eye tracking: implications for recording vergence accuracy. *Journal of Eye Movement Research*, 9(4).

Jiang, H., Christie, M., Wang, X., Liu, L., Wang, B., and Chen, B. (2021). Camera keyframing with style and control. *ACM Transactions on Graphics (TOG)*, 40(6):1–13.

Jiménez Navarro, D., Peng, X., Zhang, Y., Myszkowski, K., Seidel, H.-P., Sun, Q., and Serrano, A. (2024). Accelerating saccadic response through spatial and temporal cross-modal misalignments. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–12.

Jindal, A., Wolski, K., Myszkowski, K., and Mantiuk, R. K. (2021). Perceptual model for adaptive local shading and refresh rate. *ACM Transactions on Graphics (TOG)*, 40(6):1–18.

Julesz, B. (1971). *Foundations of cyclopean perception.* U. Chicago Press.

Kalesnykas, R. and Hallett, P. (1994). Retinal eccentricity and the latency of eye saccades. *Vision research*, 34(4):517–531.

Kang, K. and Cho, S. (2019). Interactive and automatic navigation for 360 video playback. *ACM Transactions on Graphics (TOG)*, 38(4):1–11.

Kaplanyan, A. S., Sochenov, A., Leimkühler, T., Okunev, M., Goodall, T., and Rufo, G. (2019). Deepfovea: neural reconstruction for foveated rendering and video compression using learned statistics of natural videos. *ACM Transactions on Graphics (TOG)*, 38(6):1–13.

Kellnhofer, P., Ritschel, T., Myszkowski, K., and Seidel, H.-P. (2013). Optimizing disparity for motion in depth. In *Computer Graphics Forum*, volume 32, pages 143–152. Wiley Online Library.

Kelly, D. H. (1979). Motion and vision. ii. stabilized spatio-temporal threshold surface. *Josa*, 69(10):1340–1349.

Kersten, D., Mamassian, P., and Knill, D. C. (1997). Moving cast shadows induce apparent motion in depth. *Perception*, 26(2):171–192.

Kim, I., Hong, S. W., Shevell, S. K., and Shim, W. M. (2020). Neural representations of perceptual color experience in the human ventral visual pathway. *Proceedings of the National Academy of Sciences*, 117(23):13145–13150.

Kim, J., Madhusudan, A., Watson, B., Boudaoud, B., Tarrazo, R., and Spjut, J. (2022). Display size and targeting performance: Small hurts, large may help. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–8.

Kim, J., Spjut, J., McGuire, M., Majercik, A., Boudaoud, B., Albert, R., and Luebke, D. (2019). Esports arms race: Latency and refresh rate for competitive gaming tasks. *Journal of Vision*, 19(10):218c–218c.

King, W. (2011). Binocular coordination of eye movements–hering's law of equal innervation or uniocular control? *European Journal of Neuroscience*, 33(11):2139–2146.

Konrad, R., Angelopoulos, A., and Wetzstein, G. (2020). Gaze-contingent ocular parallax rendering for virtual reality. *ACM Trans. Graph.*, 39.

Koposov, D., Semenova, M., Somov, A., Lange, A., Stepanov, A., and Burnaev, E. (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)*, pages 1560–1565. IEEE.

Koskela, M., Lotvonen, A., Mäkitalo, M., Kivi, P., Viitanen, T., and Jääskeläinen, P. (2019). Foveated real-time path tracing in visual-polar space. In *Eurographics Symposium on Rendering*. The Eurographics Association.

Koskela, M., Viitanen, T., Jääskeläinen, P., and Takala, J. (2016). Foveated path tracing. In *International Symposium on Visual Computing*, pages 723–732. Springer.

Kothari, R., Yang, Z., Kanan, C., Bailey, R., Pelz, J. B., and Diaz, G. J. (2020). Gaze-in-wild: A dataset for studying eye and head coordination in everyday activities. *Scientific reports*, 10(1):1–18.

Kowler, E. (2011). Eye movements: The past 25years. *Vision Research*, 51(13):1457–1483. Vision Research 50th Anniversary Issue: Part 2.

Krajancich, B., Kellnhofer, P., and Wetzstein, G. (2020). Optimizing depth perception in virtual and augmented reality through gaze-contingent stereo rendering. *ACM Transactions on Graphics (TOG)*, 39(6):1–10.

Krajancich, B., Kellnhofer, P., and Wetzstein, G. (2021). A perceptual model for eccentricity-dependent spatio-temporal flicker fusion and its applications to foveated graphics. *ACM Trans. Graph.*, 40.

Krajancich, B., Kellnhofer, P., and Wetzstein, G. (2023). Towards attention–aware foveated rendering. *ACM Transactions on Graphics (TOG)*, 42(4):1–10.

Krauskopf, J. and Karl, G. (1992). Color discrimination and adaptation. *Vision research*, 32(11):2165–2175.

Kwak, Y., Penner, E., Wang, X., Saeedpour-Parizi, M. R., Mercier, O., Wu, X., Murdison, S., and Guan, P. (2024). Saccade-contingent rendering. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–9.

Lang, A., Gaertner, C., Ghassemi, E., Yang, Q., Orssaud, C., and Kapoula, Z. (2014). Saccade-vergence properties remain more stable over short-time repetition under overlap than under gap task: a preliminary study. *Frontiers in Human Neuroscience*, 8:372.

Lappe, M., Bremmer, F., and van den Berg, A. V. (1999). Perception of self-motion from visual flow. *Trends in cognitive sciences*, 3(9):329–336.

Larimer, J., Cicerone, C. M., et al. (1974). Opponent-process additivity—i: Red/green equilibria. *Vision Research*, 14(11):1127–1140.

Larimer, J., Krantz, D. H., and Cicerone, C. M. (1975). Opponent process additivity—ii. yellow/blue equilibria and nonlinear models. *Vision research*, 15(6):723–731.

Layton, O. W. and Fajen, B. R. (2016). The temporal dynamics of heading perception in the presence of moving objects. *Journal of neurophysiology*, 115(1):286–300.

Le Meur, O., Coutrot, A., Liu, Z., Rämä, P., Le Roch, A., and Helo, A. (2017). Visual attention saccadic models learn to emulate gaze patterns from childhood to adulthood. *IEEE Transactions on Image Processing*, 26(10):4777–4789.

Legge, G. E. and Foley, J. M. (1980). Contrast masking in human vision. *Journal of the optical Society of America*, 70(12):1458–1471.

Leigh, R. J. and Zee, D. S. (2015). *The neurology of eye movements*. Oxford university press.

Leng, Y., Chen, C.-C., Sun, Q., Huang, J., and Zhu, Y. (2019). Energy-efficient video processing for virtual reality. In *Proceedings of the 46th International Symposium on Computer Architecture*, pages 91–103.

Li, L., Chen, J., and Peng, X. (2009). Influence of visual path information on human heading perception during rotation. *Journal of Vision*, 9(3):29–29.

Li, L., Ni, L., Lappe, M., Niehorster, D. C., and Sun, Q. (2018). No special treatment of independent object motion for heading perception. *Journal of Vision*, 18(4):19–19.

Lin, W., Feng, Y., and Zhu, Y. (2025). Metasapiens: Real-time neural rendering with efficiency-aware pruning and accelerated foveated rendering. In *Proceedings of the 30th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 1*, pages 669–682.

Lino, C. and Christie, M. (2015). Intuitive and efficient camera control with the toric space. *ACM Transactions on Graphics (TOG)*, 34(4):1–12.

Lisberger, S. G., Morris, E. J., and Tychsen, L. (1987). Visual motion processing and sensory-motor integration for smooth pursuit eye movements. *Annual review of neuroscience*, 10:97–129.

Lisi, M., Solomon, J. A., and Morgan, M. J. (2019). Gain control of saccadic eye movements is probabilistic. *Proceedings of the National Academy of Sciences*, 116(32):16137–16142.

Luce, R., Bush, R. R., and Galanter, E. E. (1963). *Handbook of mathematical psychology: I.* John Wiley.

Ludwig, C. J., Davies, J. R., and Eckstein, M. P. (2014a). Foveal analysis and peripheral selection during active visual sampling. *Proceedings of the National Academy of Sciences*, 111(2):E291–E299.

Ludwig, C. J., Davies, J. R., and Eckstein, M. P. (2014b). Foveal analysis and peripheral selection during active visual sampling. *Proceedings of the National Academy of Sciences*, 111(2):E291–E299.

Lutwak, H., Bonnen, K., and Simoncelli, E. (2022). Detecting object motion during self motion. *Journal of Vision*, 22(14):3235–3235.

Lutwak, H., Murdison, T. S., and Rio, K. W. (2023). User self-motion modulates the perceptibility of jitter for world-locked objects in augmented reality. In *2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 346–355. IEEE.

MacAdam, D. L. (1942). Visual sensitivities to color differences in daylight. *Josa*, 32(5):247–274.

Mahadevan, M. S., Bedell, H. E., and Stevenson, S. B. (2018). The influence of endogenous attention on contrast perception, contrast discrimination, and saccadic reaction time. *Vision research*, 143:89–102.

Mantiuk, R., Krawczyk, G., Myszkowski, K., and Seidel, H.-P. (2004). Perception-motivated high dynamic range video encoding. *ACM Transactions on Graphics (TOG)*, 23(3):733–741.

Mantiuk, R. K., Ashraf, M., and Chapiro, A. (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):1–16.

Mantiuk, R. K., Denes, G., Chapiro, A., Kaplanyan, A., Rufo, G., Bachy, R., Lian, T., and Patney, A. (2021). Fovvideovdp: A visible difference predictor for wide field-of-view video. *ACM Transactions on Graphics (TOG)*, 40(4):1–19.

Mantiuk, R. K., Hanji, P., Ashraf, M., Asano, Y., and Chapiro, A. (2024). Colorvideovdp: A visual difference predictor for image, video and display distortions. *arXiv preprint arXiv:2401.11485*.

March, J., Krishnan, A., Watt, S., Wernikowski, M., Gao, H., Yöntem, A. Ö., and Mantiuk, R. (2022). Impact of correct and simulated focus cues on perceived realism. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9.

Marmolejo-Ramos, F., Barrera-Causil, C., Kuang, S., Fazlali, Z., Wegener, D., Kneib, T., De Bastiani, F., and Martinez-Florez, G. (2023). Generalised exponential-gaussian distribution: a method for neural reaction time analysis. *Cognitive Neurodynamics*, 17(1):221–237.

Massey Jr, F. J. (1951). The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78.

Matin, E. (1975). Saccadic suppression: A review and an analysis. *Psychological bulletin*, 81:899–917.

Mauderer, M., Conte, S., Nacenta, M. A., and Vishwanath, D. (2014). Depth perception with gaze-contingent depth of field. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, page 217–226, New York, NY, USA. Association for Computing Machinery.

Mays, L. E. (1984). Neural control of vergence eye movements: convergence and divergence neurons in midbrain. *Journal of Neurophysiology*, 51(5):1091–1108.

Mazurek, M. E., Roitman, J. D., Ditterich, J., and Shadlen, M. N. (2003). A role for neural integrators in perceptual decision making. *Cerebral cortex*, 13(11):1257–1269.

McKee, S. P., Klein, S. A., and Teller, D. Y. (1985). Statistical properties of forced-choice psychometric functions: Implications of probit analysis. *Perception & psychophysics*, 37(4):286–298.

McKee, S. P. and Nakayama, K. (1984). The detection of motion in the peripheral visual field. *Vision research*, 24(1):25–32.

Meng, X., Du, R., Zwicker, M., and Varshney, A. (2018). Kernel foveated rendering. *Proceedings of ACM Computer Graphics and Interactive Techniques*, 1(1).

Metha, A. B., Vingrys, A. J., and Badcock, D. R. (1994). Detection and discrimination of moving stimuli: the effects of color, luminance, and eccentricity. *JOSA A*, 11(6):1697–1709.

Miller, M., Cok, R., Arnold, A., and Murdoch, M. (US Patent 7,230,594, Jun. 12, 2007). Color oled display with improved power efficiency.

Miller, M., Murdoch, M., Cok, R., and Arnold, A. (US Patent 7,333,080, Feb. 19, 2008). Color oled display with improved power efficiency.

Miller, M. E., Murdoch, M. J., Ludwicki, J. E., and Arnold, A. D. (2006). P-73: Determining power consumption for emissive displays. In *SID Symposium Digest of Technical Papers*, volume 37, pages 482–485. Wiley Online Library.

Min, K. and Corso, J. J. (2019). Tased-net: Temporally-aggregating spatial encoder-decoder network for video saliency detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2394–2403.

Missal, M. and Heinen, S. J. (2017). Stopping smooth pursuit. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1718):20160200.

Mulckhuyse, M. and Theeuwes, J. (2010). Unconscious cueing effects in saccadic eye movements–facilitation and inhibition in temporal and nasal hemifield. *Vision Research*, 50(6):606–613.

Murdoch, M. J., Stokkermans, M. G., and Lambooij, M. (2015). Towards perceptual accuracy in 3d visualizations of illuminated indoor environments. *Journal of Solid State Lighting*, 2(1):1–19.

Murray, J. (1994). Some perspectives on visual depth perception. *ACM SIGGRAPH Computer Graphics*, 28(2):155–157.

Myers, C. E., Interian, A., and Moustafa, A. A. (2022). A practical introduction to using the drift diffusion model of decision-making in cognitive psychology, neuroscience, and health sciences. *Frontiers in Psychology*, 13:1039172.

Nachmani, O., Coutinho, J., Khan, A. Z., Lefèvre, P., and Blohm, G. (2020). Predicted position error triggers catch-up saccades during sustained smooth pursuit. *eneuro*, 7(1).

Neumann, B. (1984). Optical flow. *ACM SIGGRAPH Computer Graphics*, 18(1):17–19.

Noorlander, C., Koenderink, J. J., Olden, R. J. D., and Edens, B. W. (1983). Sensitivity to spatiotemporal colour contrast in the peripheral visual field. *Vision Research*, 23(1):1–11.

Norren, D. V. and Vos, J. J. (1974). Spectral transmission of the human ocular media. *Vision research*, 14(11):1237–1244.

Nyström, M., Hooge, I., and Andersson, R. (2016). Pupil size influences the eye-tracker signal during saccades. *Vision research*, 121:95–103.

Ogle, K. N. (1952). On the limits of stereoscopic vision. *Journal of experimental psychology*, 44(4):253.

Ozili, P. K. (2023). The acceptable r-square in empirical modelling for social science research. In *Social research methodology and publishing results: A guide to non-native English speakers*, pages 134–143. IGI global.

Pallus, A. C., Walton, M. M., and Mustari, M. J. (2018). Response of supraoculomotor area neurons during combined saccade-vergence movements. *Journal of Neurophysiology*, 119(2):585–596.

Palmer, E. M., Horowitz, T. S., Torralba, A., and Wolfe, J. M. (2011). What are the shapes of response time distributions in visual search? *Journal of experimental psychology: human perception and performance*, 37(1):58.

Palmer, J., Huk, A. C., and Shadlen, M. N. (2005). The effect of stimulus strength on the speed and accuracy of a perceptual decision. *Journal of vision*, 5(5):1–1.

Park, M. H., Yun, K., and Kim, G. J. (2022). Focused area of movement as an effective rest frame for reducing vr sickness. In *2022 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pages 712–713. IEEE.

Patney, A., Salvi, M., Kim, J., Kaplanyan, A., Wyman, C., Benty, N., Luebke, D., and Lefohn, A. (2016). Towards foveated rendering for gaze-tracked virtual reality. *ACM Trans. Graph.*, 35(6).

Peng, R., Cao, M., Zhai, Q., and Luo, M. R. (2021). White appearance and chromatic adaptation on a display under different ambient lighting conditions. *Color Research & Application*, 46(5):1034–1045.

Peng, X., Zhang, Y., Jiménez-Navarro, D., Serrano, A., Myszkowski, K., and Sun, Q. (2024). Measuring and predicting multisensory reaction latency: A probabilistic model for visual-auditory integration. *IEEE Transactions on Visualization and Computer Graphics*.

Polychronakis, A., Koulieris, G. A., and Mania, K. (2021). Emulating foveated path tracing. In *Proceedings of the 14th ACM SIGGRAPH Conference on Motion, Interaction and Games*, MIG '21, New York, NY, USA. Association for Computing Machinery.

Pöppel, E. and Harvey, L. O. (1973). Light-difference threshold and subjective brightness in the periphery of the visual field. *Psychologische Forschung*, 36(2):145–161.

Purves, D., Cabeza, R., Huettel, S. A., LaBar, K. S., Platt, M. L., Woldorff, M. G., and Brannon, E. M. (2008). *Cognitive neuroscience*. Sunderland: Sinauer Associates, Inc.

Quinet, J., Schultz, K., May, P. J., and Gamlin, P. D. (2020). Neural control of rapid binocular eye movements: Saccade-vergence burst neurons. *Proceedings of the National Academy of Sciences*, 117(46):29123–29132.

Ranganathan, P., Geelhoed, E., Manahan, M., and Nicholas, K. (2006). Energy-aware user interfaces and energy-adaptive displays. *Computer*, 39(3):31–38.

Rangarajan, K. and Shah, M. (1992). Interpretation of motion trajectories using focus of expansion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(12):1205–1210.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological review*, 85(2):59.

Reddi, B. A., Asrress, K. N., and Carpenter, R. H. (2003). Accuracy, information, and response time in a saccadic decision task. *Journal of neurophysiology*, 90(5):3538–3546.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788.

Rensink, R. A., O'regan, J. K., and Clark, J. J. (1997). To see or not to see: The need for attention to perceive changes in scenes. *Psychological science*, 8(5):368–373.

Rimac-Drıje, S., Vranješ, M., and Žagar, D. (2010). Foveated mean squared error–a novel video quality metric. *Multimedia Tools and Applications*, 49(3):425–445.

Rimac-Drlje, S., Martinović, G., and Zovko-Cihlar, B. (2011). Foveation-based content adaptive structural similarity index. In *2011 18th International Conference on Systems, Signals and Image Processing*, pages 1–4.

Robinson, D. A. (1965). The mechanics of human smooth pursuit eye movement. *The Journal of Physiology*, 180(3):569.

Roorda, A. and Williams, D. R. (1999). The arrangement of the three cone classes in the living human eye. *Nature*, 397(6719):520–522.

Rosenholtz, R. (2020). Demystifying visual awareness: Peripheral encoding plus limited decision complexity resolve the paradox of rich visual experience and curious perceptual failures. *Attention, Perception, & Psychophysics*, 82(3):901–925.

Rosenholtz, R., Huang, J., Raj, A., Balas, B. J., and Ilie, L. (2012). A summary statistic representation in peripheral vision explains visual search. *Journal of vision*, 12(4):14–14.

Roufs, J. (1978). Light as a true visual quantity: principles of measurement. *CIE publication*, 41.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252.

Sabelman, E. E. and Lam, R. (2015). The real-life dangers of augmented reality. *IEEE Spectrum*, 52(7):48–53.

Sağlam, M., Lehnen, N., and Glasauer, S. (2011). Optimal control of natural eye-head movements minimizes the impact of noise. *Journal of Neuroscience*, 31(45):16185–16193.

Sasaki, R., Angelaki, D. E., and DeAngelis, G. C. (2017). Dissociation of self-motion and object motion by linear population decoding that approximates marginalization. *Journal of Neuroscience*, 37(46):11204–11219.

Schiller, P. H. and Logothetis, N. K. (1990). The color-opponent and broad-band channels of the primate visual system. *Trends in neurosciences*, 13(10):392–398.

Schlachter, F. (2013). No moore's law for batteries. *Proceedings of the National Academy of Sciences*, 110(14):5273–5273.

Schwartz, S., Maquet, P., and Frith, C. (2002). Neural correlates of perceptual learning: a functional mri study of visual texture discrimination. *Proceedings of the National Academy of Sciences*, 99(26):17137–17142.

Semmlow, J. L., Yaramothu, C., and Alvarez, T. L. (2019). Dynamics of the disparity vergence slow (fusion sustaining) component. *Journal of eye movement research*, 12(4).

Serrano, A., Sitzmann, V., Ruiz-Borau, J., Wetzstein, G., Gutierrez, D., and Masia, B. (2017). Movie editing and cognitive event segmentation in virtual reality video. *ACM Transactions on Graphics (TOG)*, 36(4):1–12.

Shi, P., Billeter, M., and Eisemann, E. (2022). Stereo-consistent screen-space ambient occlusion. *Proc. ACM Comput. Graph. Interact. Tech.*, 5(1):2–1.

Shin, D., Kim, Y., Chang, N., and Pedram, M. (2013). Dynamic driver supply voltage scaling for organic light emitting diode displays. *IEEE Transactions on Computer-Aided Design of integrated circuits and systems*, 32(7):1017–1030.

Shye, A., Scholbrock, B., and Memik, G. (2009). Into the wild: studying real user activity patterns to guide power optimizations for mobile architectures. In *Proceedings of the 42nd annual IEEE/ACM international symposium on microarchitecture*, pages 168–178.

Silberman, N., Hoiem, D., Kohli, P., and Fergus, R. (2012). Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pages 746–760. Springer.

Sincich, L. C., Zhang, Y., Tiruveedhula, P., Horton, J. C., and Roorda, A. (2009). Resolving single cone inputs to visual receptive fields. *Nature neuroscience*, 12(8):967–969.

Sitzmann, V., Serrano, A., Pavel, A., Agrawala, M., Gutierrez, D., Masia, B., and Wetzstein, G. (2018). Saliency in vr: How do people explore virtual environments? *IEEE transactions on visualization and computer graphics*, 24(4):1633–1642.

Skirzewski, M., Molotchnikoff, S., Hernandez, L. F., and Maya-Vetencourt, J. F. (2022). Multisensory integration: is medial prefrontal cortex signaling relevant for the treatment of higher-order visual dysfunctions? *Frontiers in molecular neuroscience*, 14:806376.

Smith, T. and Guild, J. (1931). The cie colorimetric standards and their use. *Transactions of the optical society*, 33(3):73.

Smith, T. J. (2013). 165watching you watch movies: Using eye tracking to inform cognitive film theory. In *Psychocinematics: Exploring Cognition at the Movies*. Oxford University Press.

Smith, V. C. and Pokorny, J. (1975). Spectral sensitivity of the foveal cone photopigments between 400 and 500 nm. *Vision research*, 15(2):161–171.

Solomon, S. G. (2021). Retinal ganglion cells and the magnocellular, parvocellular, and koniocellular subcortical visual pathways from the eye to the brain. In *Handbook of clinical neurology*, volume 178, pages 31–50. Elsevier.

Song, H., Chui, T. Y. P., Zhong, Z., Elsner, A. E., and Burns, S. A. (2011). Variation of cone photoreceptor packing density with retinal eccentricity and age. *Investigative ophthalmology & visual science*, 52(10):7376–7384.

Spering, M. and Carrasco, M. (2015). Acting without seeing: eye movements reveal visual processing without awareness. *Trends in neurosciences*, 38(4):247–258.

Spering, M., Kerzel, D., Braun, D. I., Hawken, M. J., and Gegenfurtner, K. R. (2005). Effects of contrast on smooth pursuit eye movements. *Journal of vision*, 5(5):6–6.

Stockman, A. and Sharpe, L. T. (2000). The spectral sensitivities of the middle-and long-wavelength-sensitive cones derived from measurements in observers of known genotype. *Vision research*, 40(13):1711–1737.

Sun, Q., Huang, F.-C., Kim, J., Wei, L.-Y., Luebke, D., and Kaufman, A. (2017). Perceptually-guided foveation for light field displays. *ACM Trans. Graph.*, 36(6).

Sun, Q., Huang, F.-C., Wei, L.-Y., Luebke, D., Kaufman, A., and Kim, J. (2020). Eccentricity effects on blur and depth perception. *Optics express*, 28(5):6734–6739.

Sun, Q., Patney, A., Wei, L.-Y., Shapira, O., Lu, J., Asente, P., Zhu, S., Mcguire, M., Luebke, D., and Kaufman, A. (2018). Towards virtual reality infinite walking: Dynamic saccadic redirection. *ACM Trans. Graph.*, 37(4).

Talukder, A. and Matthies, L. (2004). Real-time detection of moving objects from moving vehicles using dense stereo and optical flow. In *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)(IEEE Cat. No. 04CH37566)*, volume 4, pages 3718–3725. IEEE.

Taylor, M., Creelman, C. D., et al. (1967). Pest: Efficient estimates on probability functions. *Journal of the acoustical society of America*, 41(4):782–787.

Templin, K., Didyk, P., Myszkowski, K., Hefeeda, M. M., Seidel, H.-P., and Matusik, W. (2014). Modeling and optimizing eye vergence response to stereoscopic cuts. *ACM Transactions on Graphics (TOG)*, 33(4):1–8.

Thaler, L., Schütz, A. C., Goodale, M. A., and Gegenfurtner, K. R. (2013). What is the best fixation target? the effect of target shape on stability of fixational eye movements. *Vision research*, 76:31–42.

Thibos, L., Walsh, D., and Cheney, F. (1987a). Vision beyond the resolution limit: Aliasing in the periphery. *Vision Research*, 27(12):2193–2197.

Thibos, L. N., Cheney, F. E., and Walsh, D. J. (1987b). Retinal limits to the detection and resolution of gratings. *Journal of the Optical Society of America A*, 4(8):1524–1529.

Tovar, D., Wilmott, J., Wu, X., Martin, D., Proulx, M., Lindberg, D., Zhao, Y., Mercier, O., and Guan, P. (2024). Identifying behavioral correlates to visual discomfort. *ACM Transactions on Graphics (TOG)*, 43(6):1–10.

Tovée, M. J. (2008). *An introduction to the visual system*. Cambridge University Press.

Tsujimura, T. (2017). *OLED display fundamentals and applications*. John Wiley & Sons.

Tursun, O. T., Arabadzhiyska-Koleva, E., Wernikowski, M., Mantiuk, R., Seidel, H.-P., Myszkowski, K., and Didyk, P. (2019). Luminance-contrast-aware foveated rendering. *ACM Transactions on Graphics (TOG)*, 38(4):1–14.

Tyler, C. W. (1987). Analysis of visual modulation sensitivity. iii. meridional variations in peripheral flicker sensitivity. *JOSA A*, 4(8):1612–1619.

Ujjainkar, N., Shahan, E., Chen, K., Duinkharjav, B., Sun, Q., and Zhu, Y. (2024). Exploiting human color discrimination for memory-and energy-efficient image encoding in virtual reality. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 1*, pages 166–180.

van Beers, R. J. (2007). The sources of variability in saccadic eye movements. *Journal of Neuroscience*, 27(33):8757–8770.

Van de Walle, G. A., Rubenstein, J. S., and Spelke, E. S. (1998). Infant sensitivity to shadow motions. *Cognitive Development*, 13(4):387–419.

Van den Berg, A. and Brenner, E. (1994a). Humans combine the optic flow with static depth cues for robust perception of heading. *Vision research*, 34(16):2153–2167.

Van den Berg, A. and Brenner, E. (1994b). Why two eyes are better than one for judgements of heading. *Nature*, 371(6499):700–702.

Vasiljevic, I., Kolkin, N., Zhang, S., Luo, R., Wang, H., Dai, F. Z., Daniele, A. F., Mostajabi, M., Basart, S., Walter, M. R., and Shakhnarovich, G. (2019). DIODE: A Dense Indoor and Outdoor DEpth Dataset. *CoRR*, abs/1908.00463.

Velichkovsky, B. B., Khromov, N., Korotin, A., Burnaev, E., and Somov, A. (2019). Visual fixations duration as an indicator of skill level in esports. In *IFIP Conference on Human-Computer Interaction*, pages 397–405. Springer.

Vingrys, A. J. and Mahon, L. E. (1998). Color and luminance detection and discrimination asymmetries and interactions. *Vision research*, 38(8):1085–1095.

Von Der Heydt, R., Zhou, H., and Friedman, H. S. (2000). Representation of stereoscopic edges in monkey visual cortex. *Vision research*, 40(15):1955–1967.

Walton, D. R., Dos Anjos, R. K., Friston, S., Swapp, D., Akşit, K., Steed, A., and Ritschel, T. (2021). Beyond blur: Real-time ventral metamers for foveated rendering. *ACM Transactions on Graphics*, 40(4):1–14.

Wang, R., Yu, B., Marco, J., Hu, T., Gutierrez, D., and Bao, H. (2016). Real-time rendering on a power budget. *ACM Transactions on Graphics (TOG)*, 35(4):1–11.

Wang, S. and Zhao, J. (2022). New prospectives on light adaptation of visual system research with the emerging knowledge on non-image-forming effect. *Frontiers in Built Environment*, 8:1019460.

Wang, Z., Bovik, A. C., Lu, L., and Kouloheris, J. L. (2001). Foveated wavelet image quality index. In Tescher, A. G., editor, *Applications of Digital Image Processing XXIV*, volume 4472, pages 42 – 52. International Society for Optics and Photonics, SPIE.

Warren, W. H., Morris, M. W., and Kalish, M. (1988). Perception of translational heading from optical flow. *Journal of Experimental Psychology: Human Perception and Performance*, 14(4):646.

Warren Jr, W. H. and Hannon, D. J. (1988). Direction of self-motion is perceived from optical flow. *Nature*, 336(6195):162–163.

Watson, A. B. (2014). A formula for human retinal ganglion cell receptive field density as a function of visual field location. *Journal of vision*, 14(7):15–15.

Watson, A. B. and Pelli, D. G. (1983). Quest: A bayesian adaptive psychometric method. *Perception & psychophysics*, 33(2):113–120.

Watson, A. B. and Solomon, J. A. (1997). Model of visual contrast gain control and pattern masking. *Journal of the optical society of America A*, 14(9):2379–2391.

Weier, M., Roth, T., Kruijff, E., Hinkenjann, A., Pérard-Gayot, A., Slusallek, P., and Li, Y. (2016). Foveated real-time ray tracing for head-mounted displays. In *Computer Graphics Forum*, volume 35, pages 289–298. Wiley Online Library.

Welchman, A. E., Lam, J. M., and Bülthoff, H. H. (2008). Bayesian motion estimation accounts for a surprising bias in 3d vision. *Proceedings of the National Academy of Sciences*, 105(33):12087–12092.

Wetherill, G. and Levitt, H. (1965). Sequential estimation of points on a psychometric function. *British Journal of Mathematical and Statistical Psychology*, 18(1):1–10.

Wichmann, F. A. and Hill, N. J. (2001). The psychometric function: I. fitting, sampling, and goodness of fit. *Perception & psychophysics*, 63(8):1293–1313.

Williamson, S. J. and Cummins, H. Z. (1983). *Light and color in nature and art*, volume 1. Wiley New York.

Woodworth, R. S. and Schlosberg, H. (1954). *Experimental psychology*. Oxford and IBH Publishing.

Wright, W. D. (1929). A re-determination of the trichromatic coefficients of the spectral colours. *Transactions of the Optical Society*, 30(4):141.

Xie, H., Farnand, S. P., and Murdoch, M. J. (2020a). Observer metamerism in commercial displays. *JOSA A*, 37(4):A61–A69.

Xie, M., Niehorster, D. C., Lappe, M., and Li, L. (2020b). Roles of visual and non-visual information in the perception of scene-relative object motion during walking. *Journal of Vision*, 20(10):15–15.

Xing, X. and Saunders, J. A. (2022). Perception of object motion during self-motion: Correlated biases in judgments of heading direction and object motion. *Journal of Vision*, 22(11):8–8.

Yamagishi, S. and Furukawa, S. (2020). Factors influencing saccadic reaction time: Effect of task modality, stimulus saliency, spatial congruency of stimuli, and pupil size. *Frontiers in Human Neuroscience*, page 513.

Yan, Z., Song, C., Lin, F., and Xu, W. (2018). Exploring eye adaptation in head-mounted display for energy efficient smartphone virtual reality. In *Proceedings of the 19th International Workshop on Mobile Computing Systems & Applications*, HotMobile '18, page 13–18, New York, NY, USA. Association for Computing Machinery.

Yang, Q., Bucci, M., and Kapoula, Z. (2002). The latency of saccades, vergence, and combined eye movements in children and in adults. *Investigative Ophthalmology & Visual Science*, 43(9):2939–2949.

Yang, Q. and Kapoula, Z. (2004). Saccade–vergence dynamics and interaction in children and in adults. *Experimental Brain Research*, 156:212–223.

Yang, Q., Vernet, M., Orssaud, C., Bonfils, P., Londero, A., and Kapoula, Z. (2010). Central crosstalk for somatic tinnitus: abnormal vergence eye movements. *PLoS One*, 5(7):e11845.

Zee, D. S., Fitzgibbon, E. J., and Optican, L. M. (1992). Saccade-vergence interactions in humans. *Journal of Neurophysiology*, 68(5):1624–1641.

Zhang, L., Murdoch, M. J., and Bachy, R. (2021a). Color appearance shift in augmented reality metameric matching. *JOSA A*, 38(5):701–710.

Zhang, Y., Ortin, M., Arellano, V., Wang, R., Gutierrez, D., and Bao, H. (2018). On-the-fly power-aware rendering. In *Computer Graphics Forum*, volume 37, pages 155–166. Wiley Online Library.

Zhang, Y., Wang, R., Huo, Y., Hua, W., and Bao, H. (2021b). Powernet: Learning-based real-time power-budget rendering. *IEEE Transactions on Visualization and Computer Graphics*.