The Shortest Route Is Not Always the Fastest:
Probability-Modeled Stereoscopic Eye Movement Completion Time in VR

BUDMONDE DUINKHARJAV, New York University, USA
BENJAMIN LIANG, New York University, USA
ANJUL PATNEY, NVIDIA, USA
RACHEL BROWN, NVIDIA, USA
QI SUN, New York University, USA

Fig. 1. Predicting stereoscopic eye movement completion time. Our model predicts the completion time of stereoscopic eye movement toward a target in the visual field. It provides a probability distribution of the duration between the onset of the target and the first moment we can see it in a clear, unfoveated manner, accounting for both saccadic and vergence changes necessary to do so. (a) When users focus their gaze on a specific 3D target, objects at a different depth appear doubled (due to incorrect vergence), blurred (due to peripheral vision), or both. (b) A peripheral target might trigger a gaze movement of both the left and right eyes, which initiates after a reaction time necessary for cognitive processing. (c) First, the conjugate and ballistic saccadic movement lands the target into foveal vision with high acuity; but we may still see double vision due to the slower, incomplete vergence movement. (d) Once the disconjugate vergence movement also completes, we can successfully fuse the stereoscopic imagery and see the target clearly.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
0730-0301/2023/12-ART1 $15.00
https://doi.org/10.1145/3618334

Additional Key Words and Phrases: visual performance, eye movement
Gaze movement patterns are dictated by the strengths and limitations of the visual system. Visual acuity is much higher in the central region of the retina, encouraging observers to first shift their gaze to bring targets of interest into the fovea prior to analyzing any details. Furthermore, the binocular nature of human vision dictates that both left and right eyes must move in coordination to focus at the same location. Consequently, several distinct classes of eye movement patterns have evolved in humans to fulfill various roles and are used in different situations. 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.

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. In those tasks, there is always a delay between the decision to acquire a visual target, and our two eyes successfully landing on it. We ask “how long is this delay 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.

Prior vision science studies suggest that gaze shifts move along two primary axes (Figure 2a): one in direction and the other in depth [Zee et al. 1992]. Highly rapid and consistent eye motions that quickly shift to a peripheral location, called saccades, are crucial for fast reaction to targets in different directions. In contrast, eye movements that shift the gaze in depth by rotating each eye in opposing directions, called vergence movements, are relatively slower and more inconsistent. Often, both of these movements are executed concurrently, and the performance of such combined movements exhibit a different time signature which is faster than pure vergence movements, but slower than pure saccades [Bucci et al. 2006; Lang et al. 2014; Yang and Kapoula 2004; Zee et al. 1992]. While vision science literature has extensively studied saccadic movements and provided comprehensive models for its temporal characteristics (i.e., the main sequence [Bahill et al. 1975b; van Beers 2008]), the nature of vergence and combined movements exhibit confounding theories [Chen et al. 2010; Cullen and Van Horn 2011; King 2011].

As an alternative, 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. 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 probabilistic 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. To this aim, we will release the source code and de-identified study data at www.github.com/NYU-ICL/stereo-latency. In summary, our main contributions include:

- a series of psychophysical studies and data which systematically characterize visual performance (measured by completion/offset time) across various vergence-saccade combined eye movements in VR;
- an operational model that predicts the statistical distribution of completion times;
- demonstration of the model’s accuracy and effectiveness in predicting and optimizing VR users’ target-changing performance in natural scenarios;
- model application to measure users’ visual performance discrepancies among various games, 2D and VR displays, as well as recommendations for depth designs for 3D user interfaces.

2 RELATED WORK

2.1 Eye Movement, Visual Behaviors, and Performance

Human eyes are highly dynamic, consisting of various types of movements including smooth pursuit, vestibulo-ocular, saccade, and vergence movements. Saccade and vergence are the two most frequent movements to redirect gaze in 3D spaces [Lang et al. 2014].
There has been extensive study of them in the context of computer graphics, displays, and interactions [Hadnett-Hunter et al. 2019; Yarbus 2013]. Unlike most traditional desktop displays, VR/AR platforms provide high field-of-view stereoscopic displays, which simultaneously unlock both saccade and vergence movements. Understanding the timing of these visual movements is essential in broad applications such as esports [Duijkharjav et al. 2022], driving [Salvucci and Liu 2002], and healthcare [Bertram et al. 2016].

Pure saccades are rapid and conjugate eye movements that change the direction of gaze along a circle of iso-vergence (or the geometric horopter) which is computed using the centers of the two eyes and the fixation point (Figure 2). In the scope of this work, we simplify the measurements by equalizing the optical and visual axes (cf. [Konrad et al. 2020; Krajancich et al. 2020]), leaving the study of this difference as future work. Saccades are high-speed, ballistic motions with short travel times and a probability distribution of spatial error skewing towards undershooting the target location [Lisi et al. 2019]. The scan path, speed, and spatial accuracy of a saccade are all influenced by the characteristics of the visual content [Arabadzhiiyska et al. 2017; Duijkharjav et al. 2022; Martin et al. 2022; Sitzmann et al. 2018; Specht et al. 2017; van Beers 2007], and have been extensively studied and modeled [Bahill et al. 1975b; Boghen et al. 1974; van Beers 2008]. Although those features can also be influenced by visual tasks [Hu et al. 2021a,b], studies on the main sequence [Bahill et al. 1975b] show the consistency in completion time after the ocular-motor-controlled movement starts, independent of cognitive factors.

By comparison, pure vergences are both slower and disconjugate, directing the gaze to a new location in depth and thereby defining a new geometric horopter. In stereo displays that lack accommodative cues, the displacement of the images presented to the two eyes provides an essential depth cue that drives vergence eye movements. In the context of VR/AR, the conflict between the variable vergence cues provided by stereo displacement and the static accommodation cue corresponding to the display depth commonly causes discomfort, known as vergence-accommodation conflict [Julesz 1971]. The duration of pure vergence movements is influenced by travel distance, direction, and starting depth [Templin et al. 2014b]. Measurement of vergence movements are also more challenging compared to saccades due to the relatively smaller amplitude of movements [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 a higher sensitivity to external factors such as pupil dilation [Feil et al. 2017; Jaschinski 2016; Nyström et al. 2016].

In the real 3D world, saccade and vergence movements are more commonly combined than isolated because of the 3D distribution of visual targets [Kothari et al. 2020; Lang et al. 2014]. Prior literature has demonstrated that, relative to pure vergence, these combined eye movements are accelerated by the addition of saccades [Collewijn et al. 1995; Couillard 2013; Erkelens et al. 1989; Pallus et al. 2018; Yang and Kapoula 2004]. Competing theories attempt to untangle the neurological pathways that control vergence and combined movements, and fully explain their behaviors [Mays 1984; Quinet et al. 2020; Zee et al. 1992]. However, there is no definitive and agreed-upon theory within the literature [Cullen and Van Horn 2011; King 2011], as exists for saccadic movements [Bahill et al. 1975b]. Therefore, despite the critical importance of combined eye movements, we still lack an analytical understanding of how different vergence-saccade combinations quantitatively influence visual performance. For instance, although adding a small saccade offset to a 3D target location may accelerate a slower vergence movement, would an extra long saccade provide even more acceleration, or would the benefits of the saccade be outweighed by additional travel time? If so, what size saccade is optimal for producing the fastest vergence movement? Our work attempts to answer these questions by quantifying the scale of this acceleration effect across different amplitudes of 3D gaze movements into a continuous domain probabilistic model for predicting...
gaze offset times, and side-step the need to explicitly depict the vast complexity of vergence-saccade movement behaviors.

2.2 Stereo Vision and Stereopsis-Aware Optimization

Understanding stereo vision in order to optimize computer graphics systems and user experience, especially in VR/AR environments, remains a popular research frontier [Aizenman et al. 2022; Shi et al. 2022]. Most of today’s consumer VR/AR devices are incapable of supporting accommodation; therefore, stereopsis is still the primary means by which these devices improve depth perception over conventional 2D displays.

Numerous efforts have been made to optimize stereoscopic content with gaze tracking so as to enhance the perceived realism of depth in virtual environments. Examples include grain positioning [Templin et al. 2014a], as well as optimizations considering depth [Kellnhofer et al. 2016a; Templin et al. 2014b], luminance [Wolski et al. 2022], shading material [Chapiro et al. 2015], and displays [Chapiro et al. 2014; Zhong et al. 2021]. With the surge of low-cost and low-power gaze-tracking, another emerging research line incorporates dynamic cues such as motion parallax [Kellnhofer et al. 2016b]. Depth cues may be enhanced by incorporating these various rotation and projection centers [Konrad et al. 2020; Krajancich et al. 2020]. Reduced depth acuity in peripheral vision has also been leveraged to accelerate neural rendering [Deng et al. 2022] and image reconstruction [Kaplanyan et al. 2019].

3 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 3.4) for predicting the offset timing of various eye movements.

3.1 Experimental Design

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 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 4.3 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 4.2 – 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).

The study was conducted with a Varjo Aero head-mounted VR display (HMD) with the relevant specifications detailed in Supplement A. As shown in Figure 3a, 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 3a, this stimulus included a combination of a cross and four circular flanks 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 2a. 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

\[
\alpha_v = \alpha_l - \alpha_r.
\]  

The set of gaze locations that have the same \( \alpha_v \) form an isovergence circle, visualized as the orange circles in Figure 2a. 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:

\[
\alpha_s = (\alpha_l + \alpha_r)/2.
\]

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 2a. 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, \( \alpha = (\alpha_v, \alpha_s) \), 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,

\[
\Delta \alpha = \alpha^1 - \alpha^0 = (\alpha^1_v - \alpha^0_v, \alpha^1_s - \alpha^0_s) = (\Delta \alpha_v, \Delta \alpha_s),
\]

for movement from \( \alpha^0 \) to \( \alpha^1 \).

**Conditions.** We define a condition by a pair \( (\alpha^0, \Delta \alpha) \). 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-accommodation 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 \( d_{\text{min}} \approx 0.4 \) m in front of the observer. This corresponds to a maximum vergence angle coordinate of \( \alpha_v^\text{max} = 8.4^\circ \) for an observer with an IPD of \( \omega_{\text{min}}^\text{IPD} = 59 \) mm – the lowest IPD supported by the HMD (see Supplement A). Since a larger IPD only relaxes this maximum value, we limit the maximum vergence angle to \( \alpha_v^\text{max} \leq 8.4^\circ \). See Supplement B for a more in-depth analysis.

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

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, \( \Delta \alpha \). That is, although the initial gaze position \( \alpha \) has been shown to be a relevant factor influencing offset time [Templin et al. 2014b], 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 \( \alpha_v \leq 8.4^\circ \), saccade angles \( \alpha_s \leq 15^\circ \), as well as to only consider gaze movement displacements, \( \Delta \alpha \), and to ignore initial gaze positions, \( \alpha^0 \). Within these constraints, we sample the following conditions for vergence, saccade, and combined motions respectively:

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

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. 2014b; 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. 2014b; 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., \(\frac{d}{dt} \Delta v = \frac{d}{dt} (\Delta v + \Delta \alpha)\) [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 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.

3.2 Results

Figure 3 visualizes the raw data with the identified eye movement offset time. All time values in the statistical analysis below and throughout the paper are in seconds for clarity. Additionally, Figure 4 statistically summarizes the mean of each condition.

The offset times of saccades (\(\Delta v = 0°, .37\) (mean) \(\pm .12\) (std)) are lower than offset times of vergence movements (\(\Delta \alpha v = 0°, .59 \pm .15\)). The effect applies for both divergent (\(\Delta \alpha v < 0°, .59 \pm .17\)) and convergent (\(\Delta \alpha v > 0°, .59 \pm .14\)) conditions. The average offset time of combined movements (\(48 \pm .16\)) lies in between. A repeated measures analysis of variance (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 4.

Larger vergence amplitudes (\(|\Delta \alpha v|\)) significantly prolong the offset time in combined movements. For example, the average landing time for \(|\Delta \alpha v| = 4.2°/8.6°\) 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 4b. The average landing time for pure saccade conditions, \(\Delta \alpha_s = 4°/8°12°\), are \(.38 \pm .12/.36 \pm .11/.38 \pm .13\). When \(\Delta \alpha v = -8.4°\), however, the fastest combined movement occurs for \(\Delta \alpha s = 8°\) \((.49 \pm .16)\), compared with the other two conditions \(\Delta \alpha s = 4°\) \((.55 \pm .18)\) and \(\Delta \alpha s = 12°\) \((.60 \pm .15)\). A Mann-Kendall trend test did not observe a significant monotonic trend \((r = .33, p = 1.0)\).

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 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 3.
3.3 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 4b). 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.

3.4 Generalization to Arbitrary Gaze Movements
Statistical model. The statistical analyses in Sections 3.2 and 3.3 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 3.2, 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),

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

(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, erfc$(\cdot)$ is the complementary error function. As shown in Figure 3, 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, $\Delta \alpha = (\Delta \alpha_v, \Delta \alpha_s)$, to output the predicted ExGauss random variable, $\mathcal{T}(\Delta \alpha)$, with estimated parameters

$$\hat{\mu}(\Delta \alpha) = \sum_{i} w_i^\mu \varphi(\varepsilon^\mu ||\Delta \alpha - \varepsilon_i^\mu||),$$

$$\hat{\sigma}(\Delta \alpha) = \sum_{i} w_i^\sigma \varphi(\varepsilon^\sigma ||\Delta \alpha - \varepsilon_i^\sigma||),$$

$$\hat{\tau}(\Delta \alpha) = \sum_{i} w_i^\tau \varphi(\varepsilon^\tau ||\Delta \alpha - \varepsilon_i^\tau||).$$  

(5)

$\varepsilon_i^\mu$ and $w_i^\mu$ 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 $\varepsilon_i^\mu$, $w_i^\mu$, and $\varepsilon^\mu$ 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

$$L_j = \frac{1}{N} \sum_{i} (\hat{T}_i - T_i)^2 \text{ for } j \in \{\mu, \sigma, \tau\}.$$  

(6)

The RBF parameters are regressed using batch gradient descent with the loss functions from Equation (6) and a learning rate of $10^{-2}$ for 200,000 iterations. The mean-squared losses are minimized from $137/2.3k/17k s^2$ to $230/200/120 s^2$ over the course of each regression, respectively. We report model performance metrics as well as additional evaluations in Section 4.

Discussion and applications. We compare the mean offset times predicted by our model to the means aggregated from our dataset in Figure 5. 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 4c. Additionally, we see the U-shaped behavior for divergent combined movements, as discussed in Section 3.3 and Fig. 4b. The ExGauss distribution and RBF interpolation methods are represented by parameterized differential 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:

$$\Delta \alpha^*_s = \arg \min_{\Delta \alpha_v} \mathbb{E} \left[ \mathcal{T}(\Delta \alpha = (\Delta \alpha_v, \Delta \alpha_s)) \right]$$

$$= \arg \min_{\Delta \alpha_v} \left( \hat{\mu}(\Delta \alpha_v, \Delta \alpha_s) + \hat{\tau}(\Delta \alpha_v, \Delta \alpha_s) \right).$$  

(7)

These local minima indicate the location of the lowest point in the valley of the U-shaped behavior, as visualized in Figure 5.
4 EVALUATION

We first measure the statistical accuracy and necessity of the vergence-saccade combined modeling with an ablation study in Section 4.1. We further test the model’s goodness-of-fit when generalizing to unseen users and trials in Section 4.2. Then, to evaluate its applicability in real-world scenarios and novel conditions, we perform an evaluation user study with various scenes in Section 4.3.

4.1 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.

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 4 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 3.4. See Supplement D for visualizations of the ablated model predictions.

Table 1. KL divergence of the model and ablation study.

<table>
<thead>
<tr>
<th>Condition</th>
<th>FULL</th>
<th>VER</th>
<th>SAC</th>
</tr>
</thead>
<tbody>
<tr>
<td>KL Divergence</td>
<td>.172</td>
<td>.236</td>
<td>.444</td>
</tr>
</tbody>
</table>

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 1. 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 3.3. 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.

4.2 Model Generalizability

We further evaluate generalized goodness-of-fit with unseen data partitions. We create segments of the psychophysical data from Section 3 into training-testing groups along multiple axes.

Metrics. Similar to prior work on stochastic visual behaviors [Duinkharjav et al. 2022; 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. Significance ($p < .05$) in the K.S. test indicates a rejection of the null hypothesis that two samples are drawn from the same distribution; failing to reject ($p > .05$) supports distributional matching. The $D$ value in K.S. measures the maximum distance.

Conditions. We first assess the model’s statistical goodness of fit for the full set of psychophysical data from Section 3. 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 3.4, 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 $C_i, i \in \{1, 2, \ldots, 8\}$); (2) uniformly randomly sample $1/8$ of the entire data for each condition but across all users (annotated as $C_r$). For both methods, the remaining data is used as the corresponding training set.

Results and discussion. Figure 6a shows the results for the goodness-of-fit across all conditions. Additionally in Figure 6b, 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 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.

4.3 Study: Predicting and Optimizing Visual Performance
Beyond measuring the performance of the model on data from the controlled experiment (Section 3), 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 3. All participants reported having normal or correct-to-normal vision. We utilized the same hardware and “preamble” checklist as in Section 3.1.

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 7a), (2) a rendered basketball court with a 3D ball stimulus (Figure 7b), and (3) a photographic natural outdoor scene with a virtual bird stimulus to simulate pass-through augmented reality (AR) scenarios (Figure 7c).

Tasks. We instructed participants to complete a target-changing task similar to Section 3.1. 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 3.1, we experimented with a fixation at 0.4 m and targets placed $\Delta s_{\text{p}} = 6.9^\circ$ away in depth. Using this novel vergence depth, we designed 3 conditions with various eye travel distances:

- $C_4$: pure vergence motion with the shortest distance, $\Delta s_{\text{p}} = 0^\circ$,
- $C_5$: combined motion with the medium distance $\Delta s_{\text{p}} = 7^\circ$,
- $C_6$: combined motion with the longest distance $\Delta s_{\text{p}} = 10.5^\circ$.

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 $\times$ 3 scenes $\times$ 3 conditions $\times$ 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 7 summarizes the results (see Supplement E 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 7d), participants (Figure 7e), and scenes. Specifically,

(1) Across the 3 conditions, $C_5$ exhibits the fastest average offset time (.49 $\pm$ .16), compared to $C_4$ (.58 $\pm$ .13) and $C_6$ (.52 $\pm$ .13) conditions. The trend agrees with the model’s prediction for $C_5/C_4/C_6$, as $.44 \pm .13/60.15/54.16$. The predictions for $C_4$ in Figure 7d 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 $.52 \pm .27 / .56 \pm .29 / .54 \pm .23$, 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.

5 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).

5.1 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. 2022; 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 3 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 Steam1: *Job Simulator*, *Arizona Sunshine*, *Beat Saber*, and *Pistol Whip*.

With this data, we can simulate various gaze shifts between fixations $h_{\text{fixation}}$ 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, $h_f(\alpha_v | G)$. 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, $\alpha_o$ and $\alpha_t$, the joint probability density, $h_{\text{movement}}(\Delta \alpha_v)$, is equal to

$$h_m(\Delta \alpha_v = \alpha_t - \alpha_o | G) = h_f(\alpha_t | G) \times h_f(\alpha_o | G).$$

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., $\Delta \alpha_v \in [4^\circ, 12^\circ]$).

Results and discussion. We visualize our main results in Figure 8. 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. *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.

5.2 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. Sub-optimal 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 9a), 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 m². Our model provides a quantitative metric for drivers’ target-reaching time as a consequence of varying HUD projection distances.

1https://store.steampowered.com/vr/?p=0&tab=TopSellers

https://media.mbusa.com/releases/release-9e11a76b3d4c5181484f9c1ade19bc23-meet-the-s-class-digital-my-mbux-mercedes-benz-user-experience
Therefore, driving assistant applications could leverage the predictions derived from the model to optimize the placement of UI elements. 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.

Specifically, as annotated in Figure 9b: 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 9c). 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.

Figure 10 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, we measure the average gaze offset time, \( E[T] \), 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 \( E[T] \) is visualized in Figure 10. Due to the differentiable nature of our model, we can optimize \( d_{HUD} \) to minimize \( E[T] \) 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 \( E[T] \) 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 LIMITATIONS AND FUTURE WORK

**Initial depth and eccentricity.** Our combined vergence-saccade model measures the angular displacement in 3D without considering the initial fixation depth and 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. 2014b], while off-axis movements that maintain focal depth are much more complex, and require consideration of both vertical/horizontal eccentricity and ocular-motor anatomic [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 stereocuity.** 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 stereocuity (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 paper, 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...
By leveraging our model and a variety of large-scale datasets, we continue to be useful as they identify the earliest point after initial eye movement at which the target is clearly visible. We hope that 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.

**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 et al. 1975a]. 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.

**Fig. 9.** 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 ≈ 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 ≈ 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 ≈ 7 m is optimal (342 ms at 7 m compared to 376/343 ms at 1/25 m).

**Fig. 10.** 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).

image characteristics on reaction time may add new building blocks for comprehensive measurements of visual performance.

**7 CONCLUSION**

We statistically measure and model the correlation between visual target displacement in 3D and eye movement offset time. Our data and model reveal a remarkable fact about eye movements in the 3D world: although combining a saccadic movement with a vergence movement accelerates motion towards a target in depth, the acceleration effect shows a surprisingly non-monotonic U-shape effect. Moreover, the model accurately predicts absolute temporal performance on this task without individual normalization. This is primarily because offset time for eye movements is mainly a biophysical phenomenon and not a cognitive one. We hope the research presented here inspires a new frontier exploring exciting questions about eye movements in 3D. For example, what contributes to variation in our target acquisition speeds? How do the surging virtual layers added to the physical world influence our visual attention shifts, and thus safety? And finally, how can we build future virtual environments that boost human performance in taking actions, even to outperform ourselves in the physical world?

**ACKNOWLEDGMENTS**

We would like to thank Avigael Aizenman and Agostino Gibaldi for insightful advice on processing stereo gaze data, and support in leveraging the video game gaze behavior data in their work [2022]. This project is partially supported by the National Science Foundation grants #2225861 and #2232817, and a DARPA PTG program.

**REFERENCES**


AT Bahill, D Adler, and L Stark. 1975a. Most naturally occurring human saccades have magnitudes of 15 degrees or less. *Investigative Ophthalmology* 14, 6 (1975), 468–469.


