What Draws Your Attention First? An Attention Prediction Model Based on Spatial Features in Virtual Reality

Matthew S. Castellana, *Student Member, IEEE,*, Ping Hu, Doris Gutierrez, Arie E. Kaufman, *Fellow, IEEE*

***Abstract*—Understanding visual attention is key to designing efficient human-computer interaction, especially for virtual reality (VR) and augmented reality (AR) applications. However, the relationship between 3D spatial attributes of visual stimuli and visual attention is still underexplored. To this end, we design an experiment to collect a gaze dataset in VR, and use it to quantitatively model the probability of first attention between two stimuli. First, we construct the dataset by presenting subjects with a synthetic VR scene containing varying spatial configurations of two spheres. Second, we formulate their selective attention based on a probability model that takes as input two view-specific stimuli attributes: their eccentricities in the field of view and their sizes as visual angles. Third, we train two models using our gaze dataset to predict the probability distribution of user's preferences of visual stimuli within the scene. We evaluate our method by comparing model performance across two challenging synthetic scenes in VR. Our application case study demonstrates that VR designers can utilize our models for attention prediction in two-foreground-object scenarios, which are common when designing 3D content for storytelling or scene guidance. We make the dataset and the source code to visualize it available alongside this work.**

***Index Terms*—Gaze, visual perception, attention analysis, interaction, virtual reality, augmented reality.**

Gaze-contingent techniques, such as foveated rendering [1], [2] and gaze-based interaction interfaces [3], can significantly reduce computation costs and data transportation workloads of extended reality (XR) applications, because of the gaze's capability of reflecting visual attention. To study user's visual attention in an XR system, researchers have explored incorporating higher-level features from gaze data, or other features taken from the users or the environments. For example, analysis of saccades (i.e., the quick movements of the eye between points of fixation) is often used to detect current attention within a view [4]–[6].

In a 3D scene, researchers often analyze the visual impact of foreground objects as cohesive entities when studying attention; the interaction between foreground and background objects has also garnered interest [7]–[10]. Most research in this area focuses on images and videos — meaning direct depth information is absent. There is still very limited research on how different spatial attributes of foreground objects within a foreground object group affect visual attention, which is the central issue of our study.

Castellana, Hu, Gutierrez, and Kaufman are with the Department of Computer Science, Stony Brook University, NY 11794, USA. {matcastellan, pihu, dogutierrez, ari}@cs.stonybrook.edu.

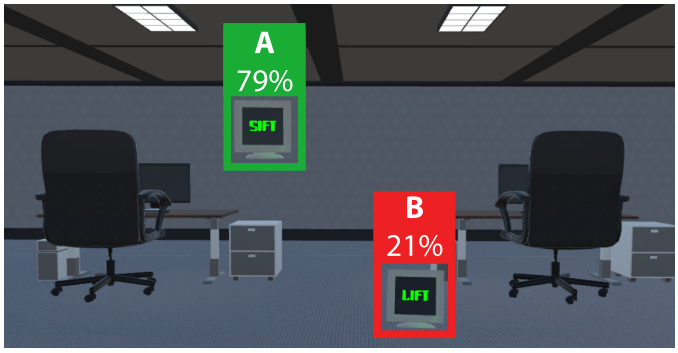


Fig. 1. Given two visual targets (e.g., target A and target B) simultaneously presented in 3D space, our model predicts the probability for each target that an observer's attention is drawn to it first. In the depicted scenario, A and B are equivalent in terms of visual angle size to the user, but since A is positioned closer to the center gaze line, our model predicts it is more likely to draw the user's attention.

We believe that when there is more than one foreground object, or when the foreground objects occupy a large area of the visual field, it is key to analyze how they influence the observer's attention. To clarify from a semantic perspective, a foreground object group can be decomposed into N foreground objects ($N \geq 2$), each consisting of a single entity (e.g., a flower) or multiple entities (e.g., a bouquet). To simplify this model, we treat the foreground object group as a two-body problem. As such, this research serves as a foundational, initial model for understanding how foreground objects affect visual attention, without implying that all real-world scenarios must necessarily be categorized into two bodies. Future work may refine this into a multi-body model in the future.

Effective multi-object visual attention prediction relies on understanding the complex interactions within the human visual cortex [11]. Stereoscopic VR provides a unique opportunity for studying these interactions, especially with regards to how various stimuli properties affect visual perception and attention. Its software can quickly and accurately report detailed spatial information of all scene content, providing advantages over other displays and real-world testing. This raises the question: how do spatial attributes of objects in VR influence a user's perceptional patterns, and in turn, their attention? To answer this, we investigate the relationship between these variables, seeking to understand users' initial selective attention in the perceptual

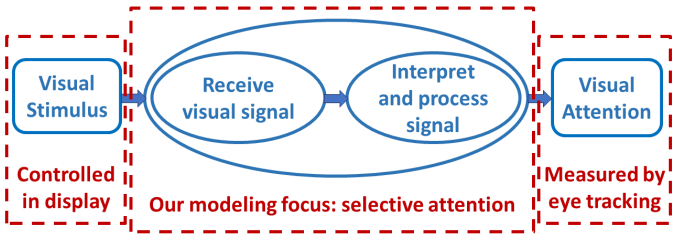


Fig. 2. A high-level overview of the logistics of our visual attention prediction model. We aim to investigate how visual attention is influenced by visual perception. We collect gaze data during an experiment with varying spatial attributes of visual stimuli under controlled conditions. Each participant's visual attention is captured by an eye-tracking HMD. Our modeling goal is to predict the attention selection between two foreground objects.

choice process (i.e., the binary choice problem between two exclusive actions [12]) when presented with two distinctive visual stimuli with a similar perceptual workload as the foreground objects.

Towards this end, we construct a gaze dataset, captured in a eye-tracking head-mounted display (HMD), and present a visual attention prediction model that considers two foreground objects with spatial features. Our innovation is to model the selective attention in a two-object foreground condition [11], an overview of which is depicted in Fig. 2.

We accomplish this through a controlled perceptual experiment conducted in VR using the HTC Vive Pro Eye, in which subject gaze data is recorded over time and compiled into a dataset. For each experiment trial, we also capture three key spatial characteristics for two visual stimuli within the scene, each of which can have two or three distinct levels:

- *Eccentricity* — the angle of deviation of the stimuli from the center of the user's view — (5°, 15°, 25°).
- *Depth* — the distance along the viewing direction of the stimuli from the gaze origin — (2m, 5m, 8m).
- *Size* — the *visual angle*, i.e., the angle subtended at the eye by the stimulus [13], [14] — (4°, 8°).

These spatial characteristics of each foreground object do not change within an experiment trial.

We use discrete levels for each characteristic to test all subjects across all combinations, ensuring maximal, balanced coverage. By providing all users with the same conditions, we can effectively measure our dependent variable: given the properties of an object and that of another competing for attention, the percentage of users that pay first attention to the former over the latter.

We consider our work as a foundational approach, creating an initial model that explores eccentricity, depth, size and showing that there are simple yet effective methods for approximating and quantifying user attention. It is important to note that horizontal and vertical eccentricity are reported to have an unequivalent effect on attention in the paracentra view [15], [16]. While separating horizontal and vertical eccentricities would offer greater precision, it would significantly extend the trial duration — in our case, from 45 minutes to an estimated 6 hours and 45 minutes per participant — raising concerns about fatigue, dropout, and attention drift. Furthermore, even with this separation, it would still be an approximation, as existing research shows that attention and perception vary across all cardinal directions [17]. Moreover, these asymmetries vary with deviation from the vertical meridian [18], adding further complexity and extending testing time. To comprehensively investigate the interaction between these three key spatial characteristics without excessively prolonging the user study or imposing undue burden on participants, we choose to utilize a unified eccentricity term.

With eccentricity, depth, and visual angle constrained to two or three unique values, we treat them as categorical variables and conduct an exploratory three-way ANOVA to gain insight into the relationship between these spatial characteristics and the first attentional selection of the user. From these findings, we generate two models: a quasi-binomial generalized linear model (GLM) providing easy-to-understand coefficients and an ensemble machine learning (ML) model for capturing complex behaviors. To determine their viability, we evaluate and compare their performance on additional user data from realistic VR environments.

The results show that both models perform effectively at predicting attentional preference between two objects, even in complex scenes. To demonstrate their applicability, we develop a prototype visualization tool for VR designers to foresee people's attention allocation in the virtual environment given a specific object placement.

Our findings significantly benefit content creators; given a 3D scene, the models can predict likely perceptual patterns and user attention without actual user data. This information can benefit scene design, helping designers focus user attention on certain areas or objects. In addition, the prediction model can enhance foveated rendering techniques and complement existing attention analysis prediction algorithms [19].

Our contributions are summarized as the follows:

- We design a perceptual experiment to understand the choice process during perception, and examining the relationship between spatial attributes of 3D visual stimuli (eccentricity, depth, and visual angle size) and attention (Sec. II).
- We formulate and compare two quantitative models for attentional selection prediction: a GLM and an ML model (Sec. III).
- We design evaluation scenes to validate our models in detailed VR environments, finding strong performance for both, with the GLM matching the ML model, offering explainable coefficients, and running efficiently in resource-limited settings (Sec. IV).
- We design and present prototypes of attention prediction tools for content designers to estimate VR users' attention allocation during the design process (Sec. VI).
- We compile a dataset, and construct supplementary code which facilitates its visualization and analysis (Appendix D).

The dataset and supplementary code can be found at https://github.com/matcastellan/TVCG-2024-Attention-Prediction.

I. Related Work

The study of the human visual system is a broad research domain; our work focuses on selective visual attention models [11], [20]. We organize the relevant literature into three subsections, based on pertinent subcategories of attention analysis: foundational models from psychology and cognition, research on gaze data and its applications, and gaze-based attention models in practical scenarios.

A. Attention Models in Psychology and Cognition

Attention is defined as a procedure in which perceived information is selectively focused on and processed by the human nervous system [21]. Closely related to perception and cognition, it reflects engagement in an activity [22], [23]. Researchers have developed several models to capture aspects of this mechanism of attention. Anderson [24] has reviewed the evidence for the value-driven mechanism of attentional selection, concluding that reward learning underpins both goal-driven and salience-driven attentional selection. The drift-diffusion model proposes a conceptual framework regarding an arbitrary choice process in consumption tasks [12], where binary choices are determined based on the choice imbalance measured using the relative entropy of the probability of choosing one of the two targets. Baek et al. [25] have developed an observer model to study how spatial cues impact visual search, while Tsotsos et al. [26] have proposed a selective tuning model for visual attention.

B. Gaze and Attention

Advances in eye-tracking equipment and techniques have enhanced our understanding of the underlying mechanisms behind the human gaze, and in turn, their effect on attention. Two types of eye movement play a fundamental role in gaze behavior: saccades and fixations. A *saccade* is a type of rapid eyeball movement with a speed up to $500^\circ/s$ that typically lasts less than $100ms$; during this time, visual sensitivity to external visual stimuli decreases [5]. In contrast, a *fixation* involves more stable eye movements that maintain visual clarity [27], Fixational eye movement greatly influences visual perception; ocular drift increases spatial acuity, and microsaccades relocate the target in the fovea, causing a momentary adjustment in vision [28]. Due to their importance in understanding attention [29], numerous techniques have been developed to detect saccades and fixations [7], [30]–[32]. However, this information alone is not sufficient to determine a user's attention — gaze origin and direction are needed also to compute the gaze-object intersection. This data can be captured and computed using a variety of existing techniques and systems [33], [34].

To support the research community with gaze-based attention analysis research, we introduce a gaze visualization tool capable of fixation detection and tracking gaze-object interactions from VR and AR log data. This tool enables researchers to effectively analyze where users focus their attention on specific objects within the virtual space.

C. Gaze-Based Attention Models in Applications

Applications of gaze-based attention models have garnered significant interest in fields like graphics and computer vision. Some of this research focuses on lower-level characteristics and behaviors of the eye. For example, the influence of 2D image features (i.e., contrast, frequency, and eccentricity) on reaction time and saccade behavior has been studied and modeled in Duinkharjav et al. [35]. Similarly, Krajancich et al. [36] have measured participants' contrast sensitivity in their peripheral view and demonstrated that their contrast tolerance is higher when they concentrate on a task visible within their fovea. To improve saccade prediction across different environments, Arabadzhiyska et al. have investigated the effect variability of saccade orientation in 3D space and the smooth pursuit of eye-motion among users [37]. Our research focus shares some similarities with Tursun et al. [38], which presented new psychophysical experiments to measure sensitivity to spatial-temporal stimuli across a wide field of view. However, we concentrate on spatial attributes of visual stimuli — and in particular, on finding the attention pattern when multiple perceptually-comparable visual stimuli appear simultaneously.

Higher-level gaze analysis research develops gaze-based attention models for application-level use. As immersive displays are becoming more accessible and widespread, recent research has focused on attention-related applications for visual media such as 360° images [8], [39]–[41], 360° videos [9], [10], [42]–[45], and 3D dynamic scenes [19], [46]–[48]. Some have developed methods for real-time gaze prediction [19], [45], [47], [49], [50], many of which incorporate deep learning architectures; others have built up relevant datasets [48], [51], [52]. Our work can be used to supplement these approaches. Additionally, it can provide insight into predicted user attention when there is no real-time gaze available. Moreover, because of our use of a GLM, our solution is more straightforward and interpretable, and can easily be integrated into lower-level applications or applications with resource constraints.

II. Perceptual Experiment

Various attributes can affect visual attention, including color-based and structural features. As discussed in Sec. I, image features such as contrast, frequency, and eccentricity have been studied for their effects on visual attention. However, visual attention is also significantly influenced by the 3D characteristics of objects, such as an object spatial location and size. Given the lack of systematic research on 3D characteristics in this context, we focused on three fundamental attributes of a visual stimulus in 3D space: *eccentricity*, *depth*, and *size*, as previously defined.

Our goal was to answer the question: in a 3D environment, given two visual stimuli with the same mental workload, which draws the observer's attention first? To this end, we designed and conducted an experiment using VR HMDs, as their binocular views enable 3D visual perception and cognition. A high-level experiment description is provided in Fig. 3 for context. The core difference between our work and existing visual attention analysis studies is that we investigated

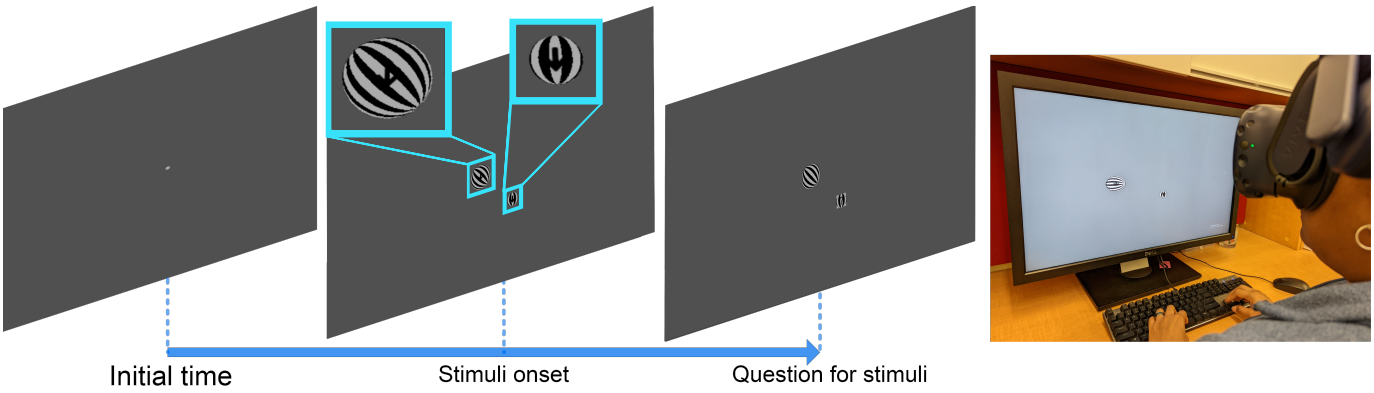


Fig. 3. From left to right: trial procedure. Prior to trial start (*Initial time*), the user focuses on a grey point in the middle of their field of view to set their initial gaze position. When the trial begins (*Stimuli onset*), the point is replaced by two spheres overlaid with arrow textures; each varies in its size and placement. After several seconds, the arrows disappear, and a question mark appears in each sphere (*Question for stimuli*), first in one, and then the other. User responses are recorded.

the spatial attributes of 3D stimuli, rather than 2D attributes (i.e., contrast, color distribution, and position on 2D views). The human binocular vision system naturally recognizes 3D content in both the real and simulated environments, allowing for distinction between objects with the same per-eye visual angle that are at different distances and have different sizes.

A. Experiment Design

To investigate how spatial attributes influence user selective attention, we designed an experiment that measures perception during a cognitive task (i.e., recognizing arrow directions on a sphere). Specifically, we wanted to identify the correlation between different spatial attributes in a 3D virtual environment.

During the study, subjects remained seated and viewed stimuli through their HMD. Before each experiment, we performed two eye-tracking calibration procedures: a hardware-level calibration included with the device, and an application-level one-point eye-tracking calibration to account for each subject's focused position variations, as discussed by Sipatchin et al. [53].

B. Visual Stimuli and Task

The experiments consisted of several series of trials where the subject must determine and remember the orientation (up, down, left, or right) of an arrow presented in each sphere (stimulus). For each trial, participants were asked to sit still and stare at a fixed point in the scene, and when the stimuli appeared, they were free to explore the scene to complete their task. The arrow was obscured by a spinning black strip pattern that hindered subjects from determining the orientation from their peripheral vision. This pattern utilized a variable number of strips depending on the sphere's size, guaranteeing a similar visual appearance between spheres in terms of stripe width and stripe frequency, which can impact performance on visual attention tasks [54]. Subjects were instructed to look at the stimuli naturally and had the liberty to move their heads to do so. After identifying the arrows, the subject input the observed arrow orientations using the arrow keys on a keyboard, and their answer was logged.

To guarantee maximum visibility, a neutral background was used. All sphere-related textures were monochrome, and the background was a solid dark grey, RGB(80,80,80), minimizing contrast to reduce user discomfort [55] while also meeting W3C contrast visibility guidelines for large-scale text [56].

The design of this arrow recognition task was motivated by two major objectives. Our first objective was to observe the natural selective attention of the subject based on the controlled variables, as indicated by their gaze behavior. To achieve this, we carefully controlled the sphere's eccentricity, depth, and visual angle size in the field of view, as well as a necessary rotation angle around the center gaze line (*theta*) for placement. When generating a trial, the spheres are first placed at the desired depth, eccentricity, and theta, and then scaled to the desired visual angle size to guarantee that the latter property will not be affected by the former. We also avoided introducing conscious gaze action, such as driving gaze to a specific object. Additionally, we avoided varying aspects of the sphere other than the controlled variables and theta; for example, regardless of sphere characteristics, the visual subtended angle of the arrow remained a constant 2°.

Our second objective was to ensure that the task could not be directly achieved using the subject's peripheral view alone. The task must have a balanced mental workload in which focused observation is needed, yet little effort should be required to recognize the stimulus within the subject's central vision. To achieve this, we added the spinning black-and-white strips overlaying the surface of each sphere.

C. Procedure

We had the participation of n = 16 subjects (13 men and 3 women, ages 18 - 60+) with normal or corrected-to-normal vision; six subjects wore glasses. All experiments were conducted under the approval of the institutional review board (IRB) of Stony Brook University's Office of Research Compliance. All subjects provided informed consent prior to participating in the experiments.

Trials were split into five groups. A short break was offered after each group, with subjects able to extend the break as

needed. Participants could end the trial at their discretion. The average duration was 45 minutes. All subjects completed the full experiment. No cybersickness was reported. To avoid the inter-trial effect, the order of trials was randomized among subjects. Additionally, eye-tracking calibration was re-conducted every time the headset was put on.

There were 171 trials, which captured every condition combination of the three attributes: eccentricity $\{5°, 15°, 25°\}$, depth $\{2.0m, 5.0m, 8.0m\}$, and size $\{4°, 8°\}$. Additionally, the angular rotation of each sphere position around the subject's center gaze (*theta*) was randomly generated ($0° - 359°$) to avoid a focus on specific orientation combinations. We avoided controlling for *theta* to minimize the experiment duration.

Each trial began with a small white sphere displayed in the center of the subject's field of view; this sphere was used to center the subject's gaze. The sphere was 1m in diameter, but was positioned 100m away from the user, thus appearing to the user as a small dot with a visual angle size of $0.573°$. Once their gaze was centered, the sphere disappeared, and two spheres with arrows were displayed. The arrows vanished after three seconds, and question marks appeared, one at a time, in a random order over the two spheres. When a question mark appeared over a sphere, participants were asked to press the arrow key on a keyboard corresponding to the arrow orientation previously displayed on that sphere. An overview of the procedure can be seen in Fig. 3.

III. Perceptual model

Using the gaze data from the perceptual experiment, our goal was to fit or learn a model that, given two foreground objects, would predict for each the probability that a user's visual attention would be drawn to it first. To achieve this goal, as an initial step, the first attention was extracted based on gaze motion in the raw data. Subsequently, the correlation between first attention and the spatial characteristics of the foreground objects was computed, after which two quantitative models were constructed for attention probability prediction.

A. Data Pre-Processing

User data was segmented by trial. It was comprised of raw gaze data, the global position of stimuli, the user's head position and orientation, and the calculated origin and direction of the user's gaze. Each trial was tagged with a user ID (a randomized series number) and a condition ID that marked the eccentricity, depth, and size values tested in the trial.

The classification of the first-looked-at sphere in each trial was completed using a custom program, the "Gaze Visualizer" (Fig. 4). The program imports and reconstructs trial data, recreating the subject's gaze as either a 2D dot or a 3D cone. Users can replay the data at various speeds, while moving the camera freely or locking it to the subject's head position and orientation. Keyboard shortcuts allow for quickly switching between subjects and trials. If calibration information has been included, gaze calibration can be toggled on and off.

The classification itself was accomplished using the Gaze Visualizer's "auto-classify" mode, where the trials are automatically replayed, and the first looked-at sphere is automatically recorded based on a pre-determined, customizable metric. For our experiments, we used the following metric: if the gaze vector was less than 3° from the vector between the gaze origin and the sphere center, and the fixation was measured during that time, the user was considered to have looked at that sphere first. Assuming a foveal vision of around 1-2°, this indicated that the user's gaze intersected the central 4° of the sphere. Fixations were measured using the Identification by Velocity Threshold (IVT) algorithm [32], where the velocity threshold is 40°/s. If no intersection occurred, the threshold angle was increased by 1° and the trial was re-run; this process repeated until an intersection was made. If two spheres were intersected simultaneously, the sphere that was uniquely intersected first was chosen.

Because of known issues with the eye tracking in the HTC Vive Pro Eye, especially when used on glasses-wearing users [57], the trials were manually reviewed for accuracy; when applicable, the answer determined by auto-classification was adjusted. The eye-tracking accuracy issue, as well as the manual classification process, is discussed in Sec. V-C.

Once classified, the data from the trials was filtered into a more compact dataset for training our quantitative models. The models predict a percentage indicating the likelihood that an object with specific eccentricity, depth, and size will attract first attention when compared to another object with its own eccentricity, depth, and size. We refer to the object being evaluated as the "potential object of interest" and the object it is compared to as the "competing object." To simplify notation, we label these objects as "A" (the object of interest) and "B" (the competing object).

Each data point includes the eccentricities, sizes, and depths for each sphere (A_{ecce}, A_{depth}, A_{size}, B_{ecce}, B_{depth}, B_{size}), as well as a value indicating the percentage of users who preferred "A" over "B" in the condition - $GroundTruthPercentPreferringA$. The full dataset consists of 324 data points. Since most conditions involve two spheres with unique combinations of eccentricity, depth, and size, either sphere can serve as the object of interest. As a result, most conditions produce two data points: one with the first sphere as "A" and the second sphere as "B," and one with their roles reversed.

As an example, consider a hypothetical condition where one sphere's characteristics are an eccentricity of 5°, a size of 4°, and a depth of 8m (abbreviated as "(5, 4, 8)"), and the other sphere has characteristics (15, 4, 2). In a test of 10 users, 7 may have looked at the first sphere first. This condition provides two data points:

- Out of 10 trials where "A" is a sphere with characteristics (5, 4, 8) and "B" is a sphere with characteristics of (15, 4, 2), "A" was preferred 7 times (70%).
- Out of 10 trials where "A" is a sphere with characteristics (15, 4, 2), and "B" is a sphere with characteristics of (5, 4, 8), "A" was preferred 3 times (30%).

This symmetric representation of the data is crucial for training our models, as either foreground object in a pair can

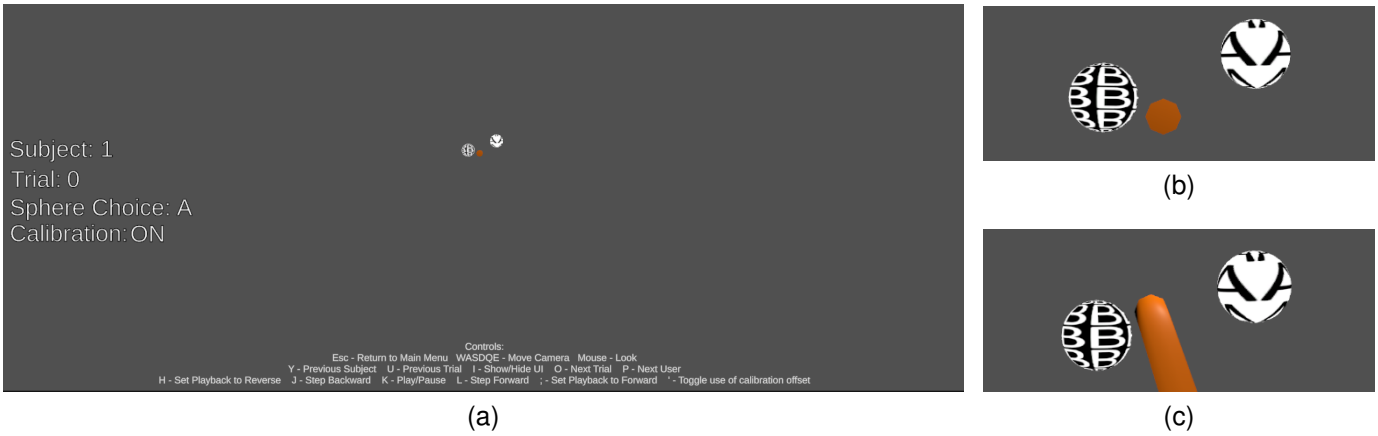


Fig. 4. Screenshots from the Gaze Visualizer. Users can control the camera, replay trials, and log which of the two spheres the subject looked at first. (a) The full screen view, displaying a reconstructed trial and the accompanying metadata. (b) A close-up of the same view, with the gaze represented as an orange dot. (c) An alternate visualization of the same timestamp of the same trial, with the gaze represented as a 3D cone; the camera has been moved to a different perspective than the user's original view.

serve as the potential object of interest or the competing object, and the model must be able to handle both cases.

In some conditions, the eccentricities, sizes, and depths of the two spheres were the same; these provided a single data point for training where the percentage is always 50%. Consider the hypothetical condition where both spheres' characteristics are (5, 4, 8), and out of 10 users, 7 looked at one sphere first, and 3 looked at the other first. Because the characteristics are equivalent, this provides the following data point:

- Out of 20 trials where "A" is a sphere with characteristics (5, 4, 8), and "B" is a sphere with characteristics of (5, 4, 8), "A" was preferred 10 times (50%).

These conditions were less useful for training our particular model, but benefit the dataset as a whole by providing additional data that may be beneficial for future work (e.g., determining the sequential effect of various conditions on attention).

B. Data Analysis

Although the variables of eccentricity, depth, and size are continuous, each trial used discrete values for each variable, enabling us to treat them as categorical variables and perform an exploratory data analysis using an analysis of variance (ANOVA). We performed a three-way ANOVA and constructed a interaction plot to determine the effect of these three independent variables on participants' gaze attention. The results (Appendix B) indicate that eccentricity and visual angle size have a statistically significant effect on user attention, while depth does not. Additionally, there appears to be no statistically significant relationship between the variables.

To further explore the results from these initial findings, we designed and analyzed a generalized linear model (GLM) [58] to gain further insight into the effects of eccentricity, depth, and size on a user's sphere preference. Given that our response variable represents a percentage that must be represented on the closed unit interval $[0, 1]$, a GLM is a suitable choice for a prediction model; we can employ the logit function as a link function that transforms probabilities into the range of all real numbers — ideal for linear combinations of predictors — and vice-versa. Using the generated coefficients, we can also identify and interpret the effects of predictor changes, which is not possible with many other machine learning techniques. As we are working with proportional data, we construct a GLM utilizing the quasibinomial family.

Our initial model was formulated as $\mathit{GroundTruthPercentPreferringA} \sim A_{ecce} + A_{depth} + A_{size} + B_{ecce} + B_{depth} + B_{size}$. The results from constructing the GLM with this model confirmed our findings from the ANOVA that depth was statistically insignificant, so the model was adjusted to $GroundTruthPercentPreferringA \sim A_{ecce} + A_{size} + B_{ecce} + B_{size}$. We experimented with variations on this looking for versions which generated optimal t-values and p-values. The final model predicts a logit of a probability p, utilizing Equation 1:

$$\begin{aligned} logit(p) = -0.1572\ A_{ecce} + 0.1313\ A_{size} \\ +\ 0.1572\ B_{ecce} - 0.1313\ B_{size} \end{aligned} \quad (1)$$

from which p can be recovered using

$$p = \frac{e^{logit(p)}}{1 + e^{logit(p)}} \quad (2)$$

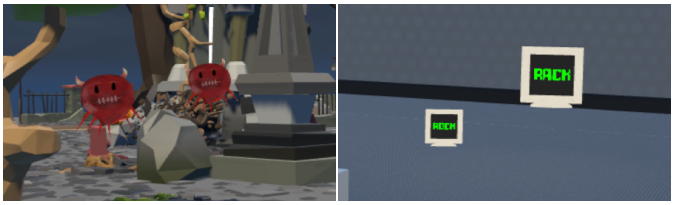
p can be considered as either the likelihood that A draws a user's first attention, or the estimated percentage of users that will have their first attention drawn to A.

Table I provides statistical details regarding the model's computed coefficients. The coefficient pairs of A_{ecce}/B_{ecce} and A_{size}/B_{size} have equal magnitude, as either object in a foreground object pair could be input to the model as "A" or "B", and the output should be logically consistent regardless. However, the sign differs as we wish to compute the log odds of first attention of whatever was input as "A"; thus, it is negative for A_{ecce} and B_{size}, since increasing these properties decreases the likelihood that "A" will be draw first attention.

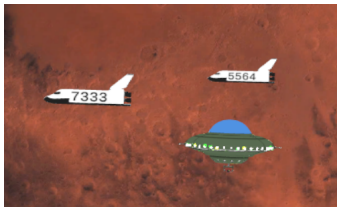
The dispersion parameter was 0.0591, indicating no significant under- or over-dispersion. Null deviance is 163.83 on

TABLE I
COEFFICIENTS OF THE QUASIBINOMIAL REGRESSION MODEL.

	Estimate	***SE***	***t***	***p***
Intercept	-1.01e-15	0.1680	0.00	1.000
A_{ecce}	-0.1572	5.02e-3	-31.32	< 0.001
A_{size}	0.1313	0.01736	7.561	< 0.001
B_{ecce}	0.1572	5.02e-3	31.32	< 0.001
B_{size}	-0.1313	0.01736	-7.561	< 0.001



(a) (b)



(c)

Fig. 5. The stimuli for our non-realistic (Figs. 5a, 5c) and realistic (Fig. 5b) testing scenes.

323 degrees of freedom, while residual deviance is 18.54 on 319 degrees of freedom, indicating that the model explains a significant portion of the variability in the response variable.

In addition to the GLM, we constructed a stacking-based ensemble model [59]. As the design of our experiment utilizes two to three discrete levels for each variable, it is difficult to infer trends from the raw data, including potential non-linear behavior that a GLM would be unable to capture. As with the GLM, the inputs were A_{ecce}, A_{size}, B_{ecce}, and B_{size}, and the output was $GroundTruthPercentPreferringA$.

IV. EVALUATION

We conducted an additional perceptual experiment to evaluate and compare the performance of our GLM and stacking ensemble models through their capability of predicting attention allocation in various practical VR environments.

A. Experiment Design

We used Unity to prepare three virtual environments for a second perceptual experiment. As perception is affected by familiarity with stimuli [60], we constructed two types of environments: novel, unfamiliar "non-realistic environments", and a more familiar "realistic environment" (Fig. 5). For the former, we prepared two scenes: a cartoonish Halloween amusement park, and outer space ; for the latter, we prepared an office environment.

Each scene has two semantically-appropriate objects for the subject to focus on and distinguish between for their task; the objects are identical, with the exception of a minimal task-related difference, to avoid influencing the user's choice of object . The Halloween amusement park scene has two ghosts with a variable number of horns that subjects must count to determine whether the total number is even or odd. The space scene has two space shuttles with a four digit number on the side; the subject must read each number digit-by-digit. The office scene has two computer monitors, each of which has a word onscreen; the words differ by one letter, and the subject must indicate the position of the letter that differs. Each of these objects of interest use modified materials and shaders to avoid being affected by lighting and shadows. With regards to the shape of the objects, the monitors have horizontal symmetry and near vertical symmetry; the ghosts have near vertical symmetry, and the shuttles have neither horizontal nor vertical symmetry, allowing us to test a variety of shape types. The shuttles also have elongated shapes, as opposed to the other two stimuli which are more spherical in nature.

Each scene presents the subject with three different tests; in each test, the tasks are the same but the level of detail differs. One contains no background objects, providing a simple environment to test the model with no confounding factors; another includes a full background, which provides static objects that could serve as potential distractions; a third includes a moving object, which could potentially draw their attention away. In total, across the three scenes, each subject performs nine tests, which are presented to them in a randomized order.

Each test has 18 conditions, in which eccentricity, depth, size, and theta values in the continuous real number range are randomly generated for each of the two stimuli. For the shuttles, which have an elongated shape, the visual angle is determined by the longest dimension of the shuttle. We constrain generated values to be between the minimum and maximum values listed in Sec. II-C. Each of the nine tests contains a unique set of conditions; the conditions are shared between subjects, but the order is randomized when presented. Trial data is logged, and the ground truth of which object the subjects preferred to look at first is determined using the Gaze Visualizer program (Fig. 4). To evaluate performance of our models, the test conditions are provided as input and a probability is generated, which is compared against the actual ground truth probability.

B. Procedure and Data Processing

We had the participation of a different set of n = 16 subjects (9 men and 7 women, ages 18-60+) with normal or corrected-to-normal vision; ten subjects wore glasses. Each trial began with subjects re-centering their gaze on and re-orienting their head toward a target object in the scene. We manually triggered the condition, displaying the two foreground objects, and then verbally asked a scene-related question related to their task; the subject responded verbally. During each trial, the object transforms, head position, head rotation, gaze origin, and gaze direction were logged at a rate of 120Hz.

The logged data was auto-classified using the Gaze Visualizer program, as described in Sec. III-A. Due to the

TABLE II
GLM MODEL PERFORMANCE BY CATEGORY

Metrics	Scene Type			Experiment Type		
	Park	Office	Space	Simple	Static	Motion
MSE	$4.20\%^2$	$4.67\%^2$	$5.94\%^2$	$5.54\%^2$	$4.99\%^2$	$4.30\%^2$
RMSE	20.50%	21.61%	24.38%	23.54%	22.33%	20.72%
Accuracy	87.04%	81.48%	81.48%	83.33%	87.04%	79.63%
Precision	87.12%	81.76%	81.48%	83.45%	87.11%	79.67%
Recall	87.03%	81.48%	81.48%	83.33%	87.04%	79.63%
F1	87.08%	81.62%	81.48%	83.39%	87.07%	79.65%
p-val	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001

TABLE III
ENSEMBLE MODEL PERFORMANCE BY CATEGORY

Metrics	Scene Type			Experiment Type		
	Park	Office	Space	Simple	Static	Motion
MSE	$4.46\%^2$	$4.92\%^2$	$6.14\%^2$	$5.79\%^2$	$5.30\%^2$	$4.36\%^2$
RMSE	21.13%	22.19%	24.78%	24.34%	23.01%	20.88%
Accuracy	83.33%	81.48%	81.48%	83.33%	83.33%	79.63%
Precision	83.88%	82.48%	81.48%	83.45%	83.40%	81.07%
Recall	83.33%	81.48%	81.48%	83.33%	83.33%	79.63%
F1	83.35%	79,69%	81.52%	83.36%	83.34%	79.66%
p-val	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001

task in the Halloween amusement park scene, where subjects counted spikes at the very bottom of the ghost, the auto-classification was modified to measure fixations anywhere on a sphere encompassing the stimulus. As users' gazes and head orientations could not be perfectly centered on the target object, the data used for training is taken from an average of the user data, excluding any outliers defined as trials with eccentricities beyond 1.5 times the interquartile range (IQR).

C. Model Performance

The overall performance of the two models can be seen in Tabs. II - IV. We measure the performance of our models using Mean Squared Error (MSE) and Root Mean Squared Error (RMSE), indicating how much the models' predictions deviated from the true percentages. For the remaining statistics, as a key task of the models is to indicate which object between the two will be preferred by subjects, we apply a binary classification approach, setting a threshold of 50% per condition. If the model correctly predicts which object the majority of the users prefer for a condition, it is classified as a success, and as a failure otherwise; the accuracy, precision, recall, F1-score, and p-value are computed using these results at the condition level — within scenes, within experiment types, and among the experiment as a whole.

Performance is comparable for both the GLM and the ensemble model, with the GLM predictions hewing closer to the actual percentages and the ensemble performing slightly better in terms of accuracy. This provides additional evidence that the formulation of the model in the GLM is appropriate.

Performance varies by scene, with the park and space scenes slightly outperforming the office, though the space scene has the highest RMSE. Considering performance by experiment type, RMSE falls slightly as scene complexity increase, with static background elements having little effect, but with motion elements affecting performance.

TABLE IV
OVERALL MODEL PERFORMANCE COMPARISON

Metrics	GLM	Ensemble
Mean Squared Error	$4.94\%^2$	$5.16\%^2$
Root Mean Squared Error	22.22%	22.72%
Accuracy	83.33%	81.48%
Precision	83.34%	81.77%
Recall	83.33%	81.48%
F1-Score	83.33%	81.52%
p-value	< 0.001	< 0.001

On average, percentage predictions by the models deviate from actual values by about 22-23%. However, with regards to accuracy, the models are able to predict which object is more likely to be looked at by the subject 81-83% of the time across all scenes, and 79-80% of the time in scenes with motion. Overall, the foundational models exhibit a relatively strong performance in its task of predicting visual attention.

V. DISCUSSION

We discuss the performance of the models (Sec. V-A), general observations from reviewing the dataset (Sec. V-B), issues encountered and the limitations of our work (Sec. V-C), and potential areas for future research (Sec. V-D).

A. Model Analysis

A review of the coefficients in Tab. I suggests that eccentricity has the most significant negative correlation with object preference - as the eccentricity of object "A" increases (or object "B" decreases), the likelihood that it first draws the user's gaze decreases significantly. A 1° move away from center view, with all other variables held constant, decreases the log odds that the user will prefer that sphere by -0.1572. Size also has a similarly strong positive correlation; a 1° increase in visual angle increases the log odds of the user looking at it first by 0.1313.

In contrast, depth appears to not be significant in commanding users' visual attention, at least within the tested range. We hypothesized that depth might play a role in stereoscopic views, as users might distinguish between two objects with the same visual angle but different points of convergence. However, the effect is negligible.

With regards to model performance: given that the two models were comparable, utilizing the GLM seems to be the preferable approach, given the explainability and relative ease of implementation. However, this presumes objects remain within the property ranges tested in the initial perceptual experiment. If the models are trained with data outside those ranges, the ensemble model may prevail if there are non-linear complexities to the relationship between gaze and attention.

Accuracy in the space scene, which utilized non-symmetrical elongated stimuli, matched or exceeded that of the other scenes, suggesting that the model can handle non-symmetrical shapes. Performance remained stable with the addition of a detailed static background, but accuracy decreased

with the addition of background motion. This suggests that while the model can handle realistic environments, certain elements that might draw user's gaze and attention could affect accuracy. However, the non-motion detailed settings are similar to many practical environments where foreground object attention is critical and distractions are minimized — e.g. AR/VR applications in work- or study-focused settings. In these contexts, our findings are directly applicable and provide actionable insights into how foreground objects are perceived and interacted with.

Regarding the precision of the percentage predictions in Tab. IV, the RMSE of 22-23% is relatively high, suggesting that the model could introduce errors when determining first attention if the true likelihood of an object attracting user attention is near 50%. This high RMSE also reduces the practical influence of size in the current model. According to the GLM findings, the impact of each spatial property is positively correlated with the magnitude of the discrepancy between that property for the two objects. The tested range of size (4-8°) is smaller than that of eccentricity (5-25°); combined with eccentricity's higher coefficient, the contribution of size to the final log odds is diminished. However, larger discrepancies in size outside of these ranges could cause this property to exert more influence - as could further improvements to the model described in Sec. V-D.

Despite this, the model was successful predicting 81-83% of conditions, suggesting that it captures fundamental patterns driving user preferences in these scenarios. Given that the model achieved this with only basic information about eccentricity and visual angle size, expanding this approach to incorporate additional information (as discussed in Section V-D) could improve the precision and accuracy of the predictions while maintaining a simple, clear, optimized implementation.

B. Dataset Patterns

During the review of subject trials, we identified some intriguing behavioral patterns that could warrant further exploration. One such pattern was that subjects' gazes would sometimes travel not to one sphere or the other, but to a point directly in between and fixate, before traveling to a sphere. The pattern appears more common when the spheres are equidistant from the subject's current gaze, suggesting indecision on which sphere to view. This behavior is consistent with findings by Hüttermann et al. [61], who determined that individuals may fixate between two targets that are equally demanding of their attention and process them peripherally. Interestingly, in our data, this can also occur when the spheres are not equidistant from the subject's gaze - the subject will actually travel over one sphere to fixate on the midpoint before selecting a target.

Another observed pattern emerged in the Halloween amusement park scene, where it is easy to see when subjects are actually performing the counting task — their gaze travels from spike to spike. In some cases, the object that attracts their attention first is not actually the object that they use to start working on the task; their gaze makes a brief visit to one ghost, before traveling to another to begin counting, and back to the first to count the remaining spikes.

C. Issues and Limitations

Our early testing of the HTC Vive Pro Eye found issues with gaze accuracy and noise during the processing of gaze-related data with the Gaze Visualizer, especially for glasses-wearing subjects, similar to those found by Schuetz and Fiehler [57]. We incorporated their solution of wiping down glasses prior to testing, and included a cleaning of the illuminators and cameras within the headset for all subjects; this improved the quality of the data. Still, we encountered some issues with noisy and imprecise gaze data in the final dataset affecting auto-classification results. Since our experiments required subjects to focus only on two objects within the scene, it was possible in most cases for a manual review to identify what was the most likely candidate for first fixation.

Manual overriding of the auto-classified result was performed in the following scenarios where classifications should have occurred, but did not:

- The subject's gaze exhibited clear travel towards and focus on or very near one sphere, in a situation where the focus could not be directed towards the other, but did not trigger an auto-classification due to excessive noise, fixation behavior outside the angle threshold, or a combination of these two reasons.
- The subject's gaze started on the sphere slightly outside the angle threshold, but exhibited extended fixation behavior.
- The gaze appears to be "offset" by a fixed amount, and removing this offset would change the result — e.g., the relative motion between two spheres would indicate that the user is looking between them, but the actual recorded gaze pattern is several degrees to the right.

Additionally, overriding occurs in the following scenarios where classifications should not have occurred, but did:

- Movement of a single frame which does not following the gaze trajectory — attributable to noise — triggers a classification erroneously.
- Excessive noise in gaze position during beginning of trial triggering a classification.

Most trials required no manual adjustment. In Section II, only 2.81% of trials were manually classified — 3.70% for glasses-wearers and 2.28% for non-wearers. In Section IV, 1.81% were manually classified, with 2.16% for glasses-wearers and 1.23% for non-wearers. Our dataset includes a list of which subjects wore glasses during their testing, allowing them to be filtered out or focused on if desired.

Another limitation encountered was dataset size. Due to our dependence on a metric based on user percentage per condition, the overall sample size used for training the model is relatively small, and training data is dependent upon the total number of conditions. The sample size could be improved with additional trials capturing additional conditions in the future.

Additionally, our model relies on certain assumptions. We consider sphere placements that are within a "reasonable" distance from the viewer (e.g., not too far/small to see, and not so close/big that they obscure the field of view). Predictions by our model on edge cases such as these may fail to accurately

predict sphere preference. Similarly, partial obscuring of one or both of the objects may affect predictions as well.

We also assume that human perceptual ability is maintained when perceiving visual stimuli at differing eccentricities and depths, and that the perceptual abilities and patterns among the participants in our experiments are following a statistically consistent route.

Finally, while our model takes into account spatial properties of objects, there are other object properties which can affect user attention, including but not limited to: sound, interactivity, opacity, contrast, frequency, and color. Such properties in real-world use cases could affect the predictive capabilities of our core model - e.g., a small, bright-red object at the periphery could draw attention away from a dull-grey object towards center view. Despite this, we believe our existing core model can already be applied to a variety of real-world scenarios — see Sec. VI.

D. Future Work

Our model distinguishes attentional preference between two objects; a useful area for further research would be to determine how the model might change when users are provided with three or more objects of interest.

The performance decrease with the addition of motion suggests that future model iterations should incorporate this information in some way — perhaps by treating the moving object as another object in the model, or by establishing some metric for background motion in the field of view. The effect of other potential distractions, such as changes in light, color, and sound, should also be studied.

As mentioned in Sec. V-C, edge case conditions (i.e., extremely close/far objects, objects in the center/far periphery, and extremely small/large objects) are not likely to perform as well under the current models. Additional research into these conditions could help develop a more comprehensive model of selective attention; for example, adding conditions where the object is extremely close to the user — where the effects of binocular vision may play a greater role — could reveal that depth affects visual attention within certain ranges .

Expanding the dataset with conditions that consider other object properties — for example, visual properties such as contrast, frequency, opacity, and color — could enable the generation of a more comprehensive model with a higher level of accuracy; so, too, could breaking up and refining existing parameters — for example, a more complex eccentricity parametrization that considers horizontal and vertical components separately. Visual angle size could potentially be measured using even more precise methods - e.g., measuring the percentage of pixels that the object takes up within the view. If the GLM continues to perform well with these extensions, the coefficients could provide insight into the collective relationship between these variables and visual attention.

We include subjects with and without glasses in our trial to capture a more diverse dataset. We hypothesize that, in the same way that the presence of glasses can affect eye tracking devices, it may also affect the attentional preferences of users, potentially by obscuring or distorting the field of view.

The spheres in our study are placed on-screen simultaneously; further research is warranted into how attentional preference might change given one object appearing after another, as the latter object may have already been observed and processed by the user.

We studied the relationship between 3D spatial attributes and selective attention at the beginning of the user's observation. However, it is also critical to identify how attention travels in 3D space over time, given multiple spatial layouts of foreground objects. Our gaze dataset contains not only data about first attention but also raw gaze data captured continuously over the entire session. Utilizing such data to study the prediction of attentional preference over time could prove fruitful as well.

VI. Application scenarios

The ability to predict which of two foreground objects draws first visual attention based on spatial attributes is valuable across a wide range of VR and AR applications. As our model is capable of predicting attention during the content design phase, it is significantly more useful than a model that can only operate at runtime. VR content designers can quickly and easily compare objects within the scene to understand how each one will draw a user's attention. Furthermore, by adjusting the position of an object's position, designers can observe not just how the adjustment affects the saliency of the object itself, but also that of a competing object.

This capability enables designers to effectively guide user attention, significantly enhancing their ability to achieve higher-level goals — for example, strategically directing users to key areas in guided experiences and ensuring engagement with important content at crucial moments. This is especially vital in immersive narratives, where highlighting key plot points or characters shapes the overall experience.

In navigation, this functionality can help users more easily identify critical waypoints, while in retail, it can optimize product placement to ensure key items capture attention effectively. Similarly, in education and training, directing focus to essential tools or information can improve learning outcomes.

All of this is accomplished without the reliance on actual users. User studies can be challenging to design and expensive to conduct, and a project may require multiple rounds of testing at various stages of development. While our existing model does not eliminate the need for user studies, it reduces it by facilitating an iterative testing approach.

In addition to predicting attention given a two-object layout in 3D space, our model can suggest optimal locations for placing objects to maximize or minimize the chance of attracting a user's attention, given the fixed placement of another object. We present two prototype examples of such a tool; one in Fig. 1, and another in Appendix A; improvements discussed in Sec. V-D could further improve its accuracy and precision by utilizing additional data.

Another potential application for our model is as a supplementary source of user attention estimates, which can augment existing algorithms and programs — including foveated rendering algorithms and existing gaze prediction techniques

based on deep learning. The GLM in particular can easily be incorporated into a wide range of applications, including low-level embedded software that may be ill-suited for more complex prediction models; it performs at a comparable level to the ensemble model with much less complexity and overhead.

VII. Conclusion

In this work, we designed and conducted perceptual experiments for investigating the relationship between 3D spatial attributes of visual targets and the gaze behavior and attentional patterns of human observers. We built up a dataset using the collected gaze data from the experiment, usable for future experiments involved in understanding the psychological, physiological, and cognitive processes of visual attention.

Based on the dataset, we evaluated the effect of three spatial attributes on attention, and determined that two (eccentricity and visual angle size) were significant within the ranges tested, and one (depth) was not. With this information, we successfully trained, tested and compared two prediction models, which demonstrated accuracy in predicting user attention between two objects. Finally, we explored potential applications of these models.

While the initial results are promising, there is room for improvement. Additional expansions to our dataset and model — capturing additional conditions, considering additional variables, and testing more complex scenarios — could generate a more robust and useful attention prediction model that is better able to capture the relationship between gaze and attention.

Acknowledgments

This project was supported in part by NSF grants OAC-1919752, ICER-1940302, and IIS-2107224.

References

[1] A. Patney, M. Salvi, J. Kim, A. Kaplanyan, C. Wyman, N. Benty, D. Luebke, and A. Lefohn, "Towards foveated rendering for gaze-tracked virtual reality," *ACM Transactions on Graphics*, vol. 35, no. 6, pp. 1–12, Dec. 2016.

[2] R. K. Mantiuk, G. Denes, A. Chapiro, A. Kaplanyan, G. Rufo, R. Bachy, T. Lian, and A. Patney, "Fovvideovdp: A visible difference predictor for wide field-of-view video," *ACM Transactions on Graphics*, vol. 40, no. 4, pp. 1–19, Jul. 2021.

[3] A. T. Duchowski, "Gaze-based interaction: A 30 year retrospective," *Computers & Graphics*, vol. 73, no. C, pp. 59–69, 2018.

[4] M. Versino, G. Castelnovo, R. Bergamaschi, A. Romani, G. Beltrami, D. Zambarbieri, and V. Cosi, "Quantitative evaluation of saccadic and smooth pursuit eye movements. is it reliable?" *Investigative Ophthalmology & Visual Science*, vol. 34, no. 5, pp. 1702–1709, 1993.

[5] E. Matin, "Saccadic suppression: a review and an analysis." *Psychological Bulletin*, vol. 81, no. 12, pp. 899–917, 1974.

[6] P. Termsarasab, T. Thammongkolchai, J. C. Rucker, and S. J. Frucht, "The diagnostic value of saccades in movement disorder patients: a practical guide and review," *Journal of Clinical Movement Disorders*, vol. 2, no. 1, pp. 1–10, 2015.

[7] V. Sitzmann, A. Serrano, A. Pavel, M. Agrawala, D. Gutierrez, B. Masia, and G. Wetzstein, "Saliency in VR: how do people explore virtual environments?" *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 4, p. 1633–1642, Apr. 2018.

[8] X. Sui, Y. Fang, H. Zhu, S. Wang, and Z. Wang, "Scandmm: A deep markov model of scanpath prediction for 360° images," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 6989–6999.

[9] E. Bernal-Berdun, D. Martin, D. Gutierrez, and B. Masia, "Sst-sal: A spherical spatio-temporal approach for saliency prediction in 360° videos," *Comput. Graph.*, vol. 106, no. C, p. 200–209, Aug. 2022.

[10] E. Bernal-Berdun, D. Martin, S. Malpica, P. J. Perez, D. Gutierrez, B. Masia, and A. Serrano, "D-sav360: A dataset of gaze scanpaths on 360° ambisonic videos," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 11, p. 4350–4360, Oct. 2023.

[11] S. Kastner and L. G. Ungerleider, "Mechanisms of visual attention in the human cortex," *Annual Review of Neuroscience*, vol. 23, no. 1, pp. 315–341, 2000.

[12] D. Fudenberg, W. Newey, P. Strack, and T. Strzalecki, "Testing the drift-diffusion model," *Proceedings of the National Academy of Sciences*, vol. 117, no. 52, pp. 33 141–33 148, 2020.

[13] D. McCready, "On size, distance, and visual angle perception," *Perception & Psychophysics*, vol. 37, no. 4, pp. 323–334, 1985.

[14] J. T. Holladay, "Proper method for calculating average visual acuity," *Journal of Refractive Surgery*, vol. 13, no. 4, pp. 388–391, 1997.

[15] M. Carrasco, C. P. Talgar, and E. L. Cameron, "Characterizing visual performance fields: Effects of transient covert attention, spatial frequency, eccentricity, task and set size," *Spatial vision*, vol. 15, no. 1, pp. 61–75, 2001.

[16] Y. Tsal, "Attending to horizontal, diagonal, and vertical positions in space," *Bulletin of the Psychonomic Society*, vol. 27, no. 2, pp. 133–134, 1989.

[17] S. Klatt, B. Noël, and R. Schrödter, "Attentional asymmetries in peripheral vision," *British Journal of Psychology*, vol. 115, no. 1, pp. 40–50, 2024.

[18] J. Abrams, A. Nizam, and M. Carrasco, "Isoeccentric locations are not equivalent: The extent of the vertical meridian asymmetry," *Vision research*, vol. 52, no. 1, pp. 70–78, 2012.

[19] Z. Hu, S. Li, C. Zhang, K. Yi, G. Wang, and D. Manocha, "Dgaze: CNN-based gaze prediction in dynamic scenes," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 5, pp. 1902–1911, 2020.

[20] G. Rizzolatti, L. Riggio, and B. M. Sheliga, "Space and selective attention," in *Attention and Performance XV: Conscious and Nonconscious Information Processing*, Aug. 1994, pp. 231–265.

[21] E. I. Knudsen, "Fundamental components of attention," *Annual Review of Neuroscience*, vol. 30, no. 1, pp. 57–78, 2007.

[22] J. Driver, "A selective review of selective attention research from the past century," *British Journal of Psychology*, vol. 92, no. 1, pp. 53–78, 2001.

[23] C. Peters, G. Castellano, and S. de Freitas, "An exploration of user engagement in hci," in *Proc. AFFINE*, 2009, pp. 1–3.

[24] B. A. Anderson, "A value-driven mechanism of attentional selection," *Journal of Vision*, vol. 13, no. 3, p. 7, Apr. 2013.

[25] J. Baek, B. A. Dosher, and Z.-L. Lu, "Visual attention in spatial cueing and visual search," *Journal of Vision*, vol. 21, no. 3, p. 162, Sep. 2021.

[26] J. K. Tsotsos, S. M. Culhane, W. Y. Kei Wai, Y. Lai, N. Davis, and F. Nuflo, "Modeling visual attention via selective tuning," *Artificial Intelligence*, vol. 78, no. 1, pp. 507–545, 1995.

[27] S. Martinez-Conde, S. L. Macknik, and D. H. Hubel, "The role of fixational eye movements in visual perception," *Nature Reviews Neuroscience*, vol. 5, no. 3, pp. 229–240, 2004.

[28] R. J. Krauzlis, L. Goffart, and Z. M. Hafed, "Neuronal control of fixation and fixational eye movements," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 372, no. 1718, pp. 1–12, 2017.

[29] J. E. Hoffman, "Visual attention and eye movements," *Attention*, pp. 119–153, 2016.

[30] F. Behrens, M. MacKeben, and W. Schröder-Preikschat, "An improved algorithm for automatic detection of saccades in eye movement data and for calculating saccade parameters," *Behavior Research Methods*, vol. 42, no. 3, pp. 701–708, 2010.

[31] S. Stuart, A. Hickey, R. Vitorio, K. Welman, S. Foo, D. Keen, and A. Godfrey, "Eye-tracker algorithms to detect saccades during static and dynamic tasks: a structured review," *Physiological Measurement*, vol. 40, no. 2, p. 02TR01, 2019.

[32] D. D. Salvucci and J. H. Goldberg, "Identifying fixations and saccades in eye-tracking protocols," in *Proc. ETRA*, 2000, pp. 71—-78.

[33] V. Sundstedt, M. Bernhard, E. Stavrakis, E. Reinhard, and M. Wimmer, "Visual attention and gaze behavior in games: An object-based approach," in *Game Analytics: Maximizing the Value of Player Data*, 2013, pp. 543–583.

[34] A. Kar and P. Corcoran, "A review and analysis of eye-gaze estimation systems, algorithms and performance evaluation methods in consumer platforms," *IEEE Access*, vol. 5, pp. 16 495–16 519, 2017.

[35] B. Duinkharjav, P. Chakravarthula, R. Brown, A. Patney, and Q. Sun, "Image features influence reaction time: a learned probabilistic perceptual model for saccade latency," *ACM Transactions on Graphics*, vol. 41, no. 4, pp. 1–15, 2022.

[36] B. Krajancich, P. Kellnhofer, and G. Wetzstein, "Towards attention–aware foveated rendering," *ACM Transactions on Graphics*, vol. 42, no. 4, pp. 1–10, Jul. 2023.
[37] E. Arabadzhiyska, C. Tursun, H.-P. Seidel, and P. Didyk, "Practical saccade prediction for head-mounted displays: Towards a comprehensive model," *ACM Transactions on Applied Perception*, vol. 20, no. 1, pp. 1–23, Jan. 2023.
[38] C. Tursun and P. Didyk, "Perceptual visibility model for temporal contrast changes in periphery," *ACM Transactions on Graphics*, vol. 42, no. 2, pp. 1–16, Nov. 2022.
[39] Y. Zhu, G. Zhai, and X. Min, "The prediction of head and eye movement for 360 degree images," *Signal Processing: Image Communication*, vol. 69, pp. 15–25, 2018.
[40] M. Assens, X. Giro-i Nieto, K. McGuinness, and N. E. O'Connor, "SaltiNet: scan-Path prediction on 360 degree images using saliency volumes," in *Proc. ICCVW*, 2017, pp. 2331–2338.
[41] D. Martin, A. Serrano, A. W. Bergman, G. Wetzstein, and B. Masia, "Scangan360: A generative model of realistic scanpaths for 360° images," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 5, pp. 2003–2013, May 2022.
[42] H.-N. Hu, Y.-C. Lin, M.-Y. Liu, H.-T. Cheng, Y.-J. Chang, and M. Sun, "Deep 360 pilot: Learning a deep agent for piloting through 360° sports videos," in *Proc. CVPR*, 2017, pp. 3451–3460.
[43] C.-L. Fan, J. Lee, W.-C. Lo, C.-Y. Huang, K.-T. Chen, and C.-H. Hsu, "Fixation prediction for 360° video streaming in head-mounted virtual reality," in *Proc. NOSSDAV*, 2017, pp. 67—72.
[44] Y. Xu, Y. Dong, J. Wu, Z. Sun, Z. Shi, J. Yu, and S. Gao, "Gaze prediction in dynamic 360° immersive videos," in *Proc. CVPR*, 2018, pp. 5333–5342.
[45] Q. Hu, J. Zhou, X. Zhang, Z. Shi, and Z. Gao, "Viewport-adaptive 360-degree video coding," *Multimedia Tools and Applications*, vol. 79, no. 17, pp. 12 205–12 226, 2020.
[46] S. Rothe, D. Buschek, and H. Hußmann, "Guidance in cinematic virtual reality-taxonomy, research status and challenges," *Multimodal Technologies and Interaction*, vol. 3, no. 1, p. 19, 2019.
[47] Z. Hu, C. Zhang, S. Li, G. Wang, and D. Manocha, "Sgaze: A data-driven eye-head coordination model for realtime gaze prediction," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 5, pp. 2002–2010, 2019.
[48] K. J. Emery, M. Zannoli, J. Warren, L. Xiao, and S. S. Talathi, "OpenNEEDS: a dataset of gaze, head, hand, and scene signals during exploration in open-ended VR environments," in *Proc. ETRA*, 2021, pp. 1–7.
[49] G. K. Illahi, M. Siekkinen, T. Kämäräinen, and A. Ylä-Jääski, "Real-time gaze prediction in virtual reality," in *Proc. MMVE*, 2022, pp. 12—18.
[50] G. A. Koulieris, G. Drettakis, D. Cunningham, and K. Mania, "Gaze prediction using machine learning for dynamic stereo manipulation in games," in *Proc. VR*, 2016, pp. 113–120.
[51] Z. Hu, A. Bulling, S. Li, and G. Wang, "Fixationnet: Forecasting eye fixations in task-oriented virtual environments," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 5, pp. 2681–2690, Mar. 2021.
[52] D. Lohr, S. Aziz, L. Friedman, and O. V. Komogortsev, "GazeBaseVR, a large-scale, longitudinal, binocular eye-tracking dataset collected in virtual reality," *Scientific Data*, vol. 10, no. 1, pp. 1–10, Mar. 2023.
[53] A. Sipatchin, S. Wahl, and K. Rifai, "Accuracy and precision of the htc vive pro eye tracking in head-restrained and head-free conditions," *Investigative Ophthalmology & Visual Science*, vol. 61, no. 7, p. 5071, 2020.
[54] J. L. Gobell, C.-h. Tseng, and G. Sperling, "The spatial distribution of visual attention," *Vision Research*, vol. 44, no. 12, pp. 1273–1296, 2004.
[55] S. M. Haigh, L. Barningham, M. Berntsen, L. V. Coutts, E. S. Hobbs, J. Irabor, E. M. Lever, P. Tang, and A. J. Wilkins, "Discomfort and the cortical haemodynamic response to coloured gratings," *Vision Research*, vol. 89, pp. 47–53, 2013.
[56] W3C, "Web content accessibility guidelines (WCAG) 2.0," web, World Wide Web Consortium (W3C), Recommendation, Dec. 2008. [Online]. Available: http://www.w3.org/TR/WCAG20/
[57] I. Schuetz and K. Fiehler, "Eye tracking in virtual reality: Vive pro eye spatial accuracy, precision, and calibration reliability," *Journal of Eye Movement Research*, vol. 15, no. 3, pp. 1–18, Sep. 2022.
[58] J. A. Nelder and R. W. Wedderburn, "Generalized linear models," *Journal of the Royal Statistical Society Series A: Statistics in Society*, vol. 135, no. 3, pp. 370–384, 1972.
[59] S. Džeroski and B. Ženko, "Is combining classifiers with stacking better than selecting the best one?" *Machine Learning*, vol. 54, pp. 255–273, Mar. 2004.
[60] P.-L. Yang and D. M. Beck, "Familiarity influences visual detection in a task that does not require explicit recognition," *Attention, Perception, & Psychophysics*, vol. 85, no. 4, pp. 1127–1149, 2023.
[61] S. Hüttermann, D. Memmert, D. J. Simons, and O. Bock, "Fixation strategy influences the ability to focus attention on two spatially separate objects," *PLoS One*, vol. 8, no. 6, pp. 1–8, 2013.

Matthew Castellana is currently pursuing a Ph.D. degree in Computer Science at Stony Brook University. He received his B.S. in Computer Science and B.E. in Computer Systems and Engineering from Rensselaer Polytechnic Institute. His research interests include virtual and augmented reality, novel VR/AR peripherals and interfaces, remote collaboration, graphics, and related areas.

Ping Hu received her Ph.D. degree in Computer Science from Stony Brook University (2022). Her research focuses include visual computing in VR/AR, generative graphics, and scientific visualization.

Doris Gutierrez is currently pursuing a Ph.D. degree in Computer Science at Stony Brook University. She received her B.S. in Computer Science and M.Sc. Information and Communications Technologies from Technological University of Panama. Her research interests include virtual and augmented reality, immersive facilities, graphics, and related areas.

Arie E. Kaufman is a Distinguished Professor of Computer Science, Director of Center of Visual Computing, and Chief Scientist of Center of Excellence in Wireless and Information Technology at Stony Brook University. He served as Chair of Computer Science Department, 1999-2017. He has conducted research for >40 years in visualization, VR and graphics and their applications, and published >350 refereed papers. He was the founding Editor-in-Chief of IEEE TVCG, 1995-98. He is an IEEE Fellow, ACM Fellow, National Academy of Inventors Fellow, recipient of IEEE Visualization Career Award (2005), and inducted into Long Island Technology Hall of Fame (2013) and IEEE Visualization Academy (2019). He received his Ph.D.in Computer Science from Ben-Gurion University, Israel (1977).