Predicting Gaze-based Target Selection in Augmented Reality Headsets based on Eye and Head Endpoint Distributions

Yushi Wei
Xi'an Jiaotong-Liverpool University
Suzhou, Jiangsu, China
yushi.wei21@student.xjtlu.edu.cn

Rongkai Shi
Xi'an Jiaotong-Liverpool University
Suzhou, Jiangsu, China
rongkai.shi19@student.xjtlu.edu.cn

Difeng Yu
University of Melbourne
Melbourne, Victoria, Australia
difengy@student.unimelb.edu.au

Yihong Wang
Xi'an Jiaotong-Liverpool University
Suzhou, Jiangsu, China
yihong.wang@xjtlu.edu.cn

Yue Li
Xi'an Jiaotong-Liverpool University
Suzhou, Jiangsu, China
yue.li@xjtlu.edu.cn

Lingyun Yu
Xi'an Jiaotong-Liverpool University
Suzhou, Jiangsu, China
lingyun.yu@xjtlu.edu.cn

Hai-Ning Liang*
Xi'an Jiaotong-Liverpool University
Suzhou, Jiangsu, China
haining.liang@xjtlu.edu.cn

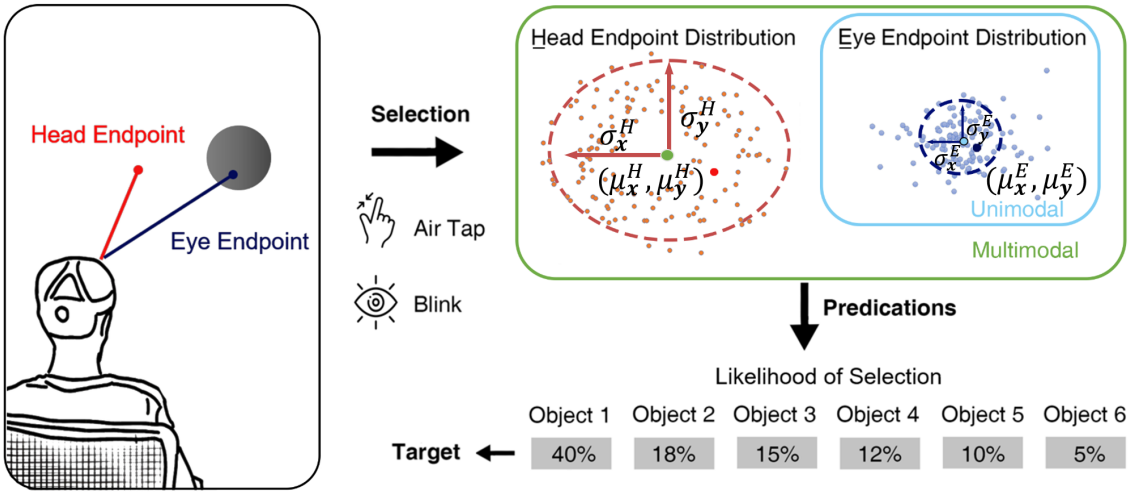

Figure 1: LEFT: a user is selecting a target with eyes in an AR headset. A head endpoint and an eye endpoint are recorded after each selection. RIGHT: Multiple repetitive selections form a head endpoint distribution and an eye endpoint distribution. The distributions are used by our models to predict the likelihood of selecting each object. The object with the highest selection probability will be chosen as the predicted target.

ABSTRACT

Target selection is a fundamental task in interactive Augmented Reality (AR) systems. Predicting the intended target of selection in such systems can provide users with a smooth, low-friction interaction experience. Our work aims to predict gaze-based target selection in AR headsets with eye and head endpoint distributions, which describe the probability distribution of eye and head 3D orientation when a user triggers a selection input. We first conducted a user study to collect users' eye and head behavior in a gaze-based pointing selection task with two confirmation mechanisms (air tap and blinking). Based on the study results, we then built two models: a unimodal model using only eye endpoints and a multimodal model using both eye and head endpoints. Results from a second user study showed that the pointing accuracy is improved by approximately 32% after integrating our models into gaze-based selection techniques.

*Corresponding author.

CHI '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9421-5/23/04…$15.00
https://doi.org/10.1145/3544548.3581042

CCS CONCEPTS

• **Human-centered computing → HCI theory, concepts and models**; **Mixed / augmented reality**; **User studies**.

KEYWORDS

Augmented reality, target selection, selection modeling, eye input, error prediction

ACM Reference Format:
Yushi Wei, Rongkai Shi, Difeng Yu, Yihong Wang, Yue Li, Lingyun Yu, and Hai-Ning Liang. 2023. Predicting Gaze-based Target Selection in Augmented Reality Headsets based on Eye and Head Endpoint Distributions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3544548.3581042

1 INTRODUCTION

Target selection is one of the most common and fundamental tasks in Augmented Reality (AR) systems. Modern AR Head-Mounted Displays (HMDs), such as Magic Leap and HoloLens 2, allow users to select virtual objects via handheld controllers or mid-air gestures. However, there are situations where users cannot interact with AR HMDs using controllers or gestures because one or both of their hands are preoccupied with other tasks, such as carrying tools or other items. Such cases would become common as AR HMDs gradually become more incorporated into people's daily lives in the same way as personal computers or mobile phones in the foreseeable future. With eye-tracking capabilities increasingly integrated into AR HMDs, gaze-based input can help solve this dilemma as a complement to the hand-based selection approaches. Eye gaze is generally fast as it could scan a wide range of areas with little physical effort [9, 64]. It is also lightweight and often indicates selection intention [16]. However, gaze-based selection can often be inaccurate due to the noisy patterns of gaze points and limitations in today's tracking devices.

There have been extensive discussions about strategies to improve the accuracy of eye-based selection, such as expanding the target size [48], adding supportive visual cues [40], and integrating multi-stage refinements [35]. However, due to frequent visual changes, target expansion or visual cue strategies inevitably sacrifice the consistency of a user interface and lower users' immersion [68]. On the other hand, a multi-stage technique may introduce extra complexity to the selection procedure. We deem that a user behavior modeling approach can potentially avoid these shortcomings. It helps understand a user's target selection process and their behavioral pattern, and ultimately, supports target selection processes with probabilistic-based model predictions [6, 25, 38, 63, 67]. This could implicitly improve the target selection accuracy that does not involve visual changes or additional interaction steps, which eventually saves users' efforts in a target selection task and make the process more usable and efficient.

Several models based on movement dynamics and selection endpoints have been explored for predicting the intended target of selection [6, 24, 63]. More relevant to our work are endpoint-based models that describe the probability distribution of a selection pointer when a user triggers a selection input (i.e., the likelihood of the user selecting a target with specific cursor positions/orientations). However, prior work has mainly focused on endpoints-based models for one-dimensional (1D) or two-dimensional (2D) target selection tasks (e.g., [5, 26, 33, 39]). While in immersive three-dimensional (3D) virtual environments, the endpoints-based models are mainly about hand-/head-based selections [15, 63]. To the best of our knowledge, limited research has explored an endpoints-based model for gaze-based target selection tasks in AR. Due to the promises and uniqueness of gaze as an input modality, it is essential to understand how interface features (e.g., object size and distance) and input mechanisms (e.g., eye blinking and air tap) can influence users' gaze selection endpoints and to further provide prediction assistance in AR systems.

To fill this gap, we present two models for gaze-based target selection. One considers only the eye endpoint when a user triggers a selection, i.e., the endpoint distribution is with eyes-only (see Figure 1, the blue scattered points). The other model adds the head endpoint to provide additional information in addition to the eye-only model, which further explains how a user's head and eyes collaborate during the selection process (see Figure 1, the multimodal uses both eye and head endpoints). To do this, we first conducted a user study to investigate how factors including target width, distance, and selection confirmation mechanisms could influence endpoint distributions and collect data for our models. We then conducted another user study to compare our target prediction models with a baseline. We show that the pointing accuracy can be improved approximately by 32% after integrating our models into gaze-based selection techniques. We also find that our unimodal model offers slightly better performance than the multimodal model.

The primary contributions of this paper include:

- Two endpoint distribution models for predicting gaze-based target selection: a unimodal model with eye endpoints only and a multimodal model with both eye and head endpoints.
- Results and interpretations of the behavioral characteristics of the eye and head during gaze-based target selection. We recorded participants' behavioral data in our studies (such as selection time, trajectories, head direction, and endpoints coordinates). We have also evaluated the influence of various factors on the outcomes (such as target width, movement amplitude, and confirmation mechanism).
- Open-sourced gaze-based selection dataset collected from our two user studies which contain 20160 data trials.

2 RELATED WORK

2.1 Gaze Behavior and Eye-Head Coordination

Gaze behavior is a well-studied and trending research domain. Prior work has leveraged eye-gaze as a single input modality in conventional interactive systems, such as desktops, phones, and tablets [2, 30, 39]. In these studies, researchers have investigated how users' eyes could be used to select targets on 2D screens with limited sizes at certain distances. More recently, researchers have considered the role of the head during gaze-based interaction, especially in HMDs, which are attached close to users' eyes. Different from conventional 2D interfaces, when interacting with HMDs, users "carry" the displays when searching and selecting the target.

This search and selection process often involves coordination between the eyes and head [51, 52]. For example, previous research has utilized head movement to support gaze-based selection [52] or to predict probable gaze positions [11]. Therefore, the role of the head becomes vital to understand users' gaze behavior while using HMDs.

The existing literature provides some understanding of how users' eyes and head coordinate with each other to achieve a comfortable viewing experience. The visual field of the human eyes is usually around 210° horizontally and 120° vertically [51]. When an eye movement is greater than 30° from the center of the sight, people tend to move their head together with eye movements [36]. For short-distance eye movements that are smaller than 30°, the head typically contributes one-third of the whole amplitude, and the percentage of head movements in the whole amplitude increases as the distance gets longer [50]. When people intend to look at a target, they typically move their eyes towards the target position, followed by head movements in about the same direction [7]. When the remaining amplitude is comfortable enough for eye movements, the head movement may stop its assistance [21]. Such 'assistance' behavior varies from person to person, where some people tend to move their heads to keep their eyes in a narrower and more comfortable range while some others do not [51]. This difference in eye-head coordination can have a significant effect on the use of HMDs, which typically requires head-supported gaze movements to search and select the target, especially for those targets located outside a user's field-of-view (FoV).

Prior work has identified that when the head is not disturbed by extra factors, the amplitude of saccadic eye movements (i.e., rapids movement of the eye between fixation points) can be expressed as a function of the head's speed [23]. However, it is unknown whether the HMD's weight would affect the head movement speed, and thus affect eyes and/or head movements. While previous work has discussed head-eye collaboration in immersive VR HMDs [51, 52], it is still unclear whether the position and depth of the holographic projections generated in realistic scenes in AR devices will have an impact on users' eye and head movements.

2.2 Gaze-Based Selection

With the upgrade of eye-tracking technology and the increasing demand for using implicit eye data as an additional modality, a growing number of studies have focused on gaze-based selection. Gaze-based selection has been applied to various scenarios, including but not limited to gaming [18, 53], object manipulation [55, 64], text input [20, 43, 44], and text selection [41, 47]. In gaze-based selection, users are able to locate the target using their eyes in a fast and natural way but may suffer from low accuracy. Considerable attention has been devoted to improving gaze-based selection accuracy. One aspect is to design and enhance the confirmation mechanism to make it stable and efficient [19, 32, 49]. Similar to other pointing-based selection, gaze-based selection is normally a two-step procedure. In the first step, a user's eyes work as an indicator to search, navigate to, and locate the target. While in the second confirmation step, the user needs to trigger the selection, either using the eyes or an additional modality [51, 52].

In addition to the eyes, gaze-based selection is often combined with other parts of people's bodies, such as the head [60], hand gestures [17, 62], handheld devices [35], or body movements [32]. For example, Klamka et al. [32] incorporated gaze with foot strokes so that when the user needs to zoom in on a 2D picture or map, instead of dragging the mouse to select the target, the area where the eye is currently looking at is the zoom center and the foot strokes are the triggers. In a multi-modal method, the eye is often used for controlling the cursor to locate the target, while a second modality is adopted for confirmation. Given the difficulty of controlling eye movements precisely, many prior studies have used eyes as a coarse localizer and performed fine-tuning of the selection via other interaction methods to improve the selection accuracy [12, 35, 60]. This approach is effective in improving selection accuracy, but it also lowers the naturalness and intuitiveness of the selection procedure with higher learning efforts and interrupted experience.

When considering using only the eyes to perform a gaze-based selection, dwell is a common and usable confirmation mechanism. After aiming at the target, a user can stay or hold the gaze at the target for a certain amount of time to confirm the selection (i.e., a dwell time). The efficiency of using dwell to select targets mainly depends on the dwell criterion [29, 39]. It is usually restricted to specific tasks, such as text selection or selecting static targets in sparse environments. First, due to their nature, people's eyes constantly have motions when gazing at a target, which results in tremors, drifts, and microsaccades [18]. This means the eyes cannot maintain a steady gaze for a long period [18]. In complex scenarios when objects are small, moving, or occluded by each other, selecting the target becomes more complicated and possibly causes serious eye fatigue [4, 59]. Blinking is another typical confirmation mechanism in gaze-based selection [43]. It can be classified as spontaneous (short) and voluntary (long) blinking, where a spontaneous blinking typically lasts around 100 milliseconds, but a voluntary blinking may last over 200 milliseconds [34]. A camera and an algorithm are needed to determine an intended blink for selecting a target [13]. When AR HMDs do not have built-in eye-tracking camera(s), an external USB eye-tracking camera is normally required to utilize blink behavior for interaction (e.g., the setup of HoloLens headset with Pupil Lab's eye tracker [35]). It should be noted that the selection using dwell or multi-modal methods is usually continuous, whereas blinking creates an interruption to the process. For a blinking action, a user would close the eyes and this ceases the indication of the gaze point in the selection process. It is unknown whether these two types of confirmation mechanisms have a direct effect on gaze-based pointing selection.

2.3 Modeling Target Selection

One of the most common ways to determine if a target has been selected is the Visual Boundary Criterion (VBC), where a target is considered to be selected only if the cursor (controlled by hand, stylus, etc.) falls within the target's boundary. An ongoing research topic is designing new target selection methods to achieve selection efficiency and accuracy using VBC. Several approaches have been proposed, such as modifying the size of targets [3, 8, 46, 48, 65], or using visual cues to indicate where the target is [40]. These approaches inevitably require visual changes in the immersive

environment and break the consistency of the interaction, which ultimately lowers the usability and interactivity [56, 68].

One open challenge is how to improve the selection accuracy without changing or distorting the interface and achieve a better understanding of human behavior in the selection process. Fitts' law is one of the well-known models for understanding human behavior when selecting a target. Its extensive generalization allows it to calculate the movement time (MT) in different scenarios [5, 57, 58]. Fitts' law has also been applied to eye-based selection. For example, Zhang [66] proposed the IDeye model, which accurately predicts the eye's MT and pointing time (EMT) in 2D screens. Isomoto et al. [29] proposed a gaze model based on Fitts' law that effectively reduces the dwell time of acquiring a target.

Fitts' law helps to understand human behavior in target selection; however, improving the efficiency and accuracy of target selection cannot be achieved only with Fitts' law, especially in a dense environment, or the targets are moving or small. Understanding the selection endpoints can be one potential solution to fill this gap. Bi et al. [6] proposed an approach applying Bayesian theory for touch screens, which treats the finger touch input as an uncertain process and reduces the selection errors compared to a VBC approach by calculating the Bayesian Touch Distance (BTD). Bi et al.'s later work has shown that the endpoint distribution with finger pointing follows a bivariate Gaussian distribution and it is possible to predict selection behavior, such as selection accuracy, using this approach [5]. These models have been generalized in several application scenarios in 2D interfaces, including pointing-based selection for 1D moving targets [27], 2D moving targets [28], 2D moving targets with arbitrary shapes [67], and crossing-based selection for moving targets [26]. In addition, they have been adapted to pointing selection for static targets in 3D VR environments with head- and hand-based approaches [63]. Zhu et al. [68] used endpoint distribution as a likelihood model in their Bayesian model. Li et al. [38] also built the Bayesian-based pointers to select targets with 2D movements.

Our work provides some unique contributions as compared to prior work: (1) While prior endpoint distribution-based models have discussed head- and hand-based interaction in 3D VEs [63], our work extends this to gaze-based selection in AR HMD scenarios. (2) Previous eye-based models considered the eye as a single input modality [39]. However, we consider both the eye and head and combine them to build a multi-modal model. (3) Prior eye-based models utilized users' eyes' trajectories but not the endpoint distributions. In contrast, our work focuses on the endpoint distribution of gaze-based selection, which represents the probability distribution of pointing direction when a user triggers a selection input.

3 RESEARCH OVERVIEW

3.1 Models

In this research, we aim to build two endpoint distribution models (a *unimodal model* and a *multimodal model*) for gaze-based target selection in AR headsets. For the unimodal model, we consider endpoint distribution as the probability distribution of gaze-pointing direction in 3D space when a user triggers a selection input. In this case, the cursor orientation (i.e., the Raycasting direction) is the same as where the eyes are looking at. For the multimodal model, we incorporate the probability distribution of the head-pointing direction into the existing eye-based model to provide additional information for target prediction. We hypothesize that the endpoint distributions of both modalities (i.e., eyes and head) to be bivariate Gaussian distribution, and different task factors including target size and target distance could influence the parameters of the distribution.

3.1.1 Unimodal Model. The basic idea of our unimodal model is to automatically select the target with the highest probability based on their gaze-pointing direction once a user triggers a selection input. The concept is similar to previous works that leverage Bayes' theorem for target prediction [6, 27, 63]. Supposing a list of potential targets in the environment $T_1, T_2, ..., T_n$, the probability of selecting a particular target T with a gaze-pointing direction G is $P(T \mid G)$. According to Bayes' theorem (Equation 1), $P(T \mid G)$ is based on **(1)** $P(T)$, the probability of selecting a target which is usually set to $1/n$; **(2)** $P(G)$, the probability of gaze pointing at the particular direction G, which is assumed to be a constant value; and **(3)** $P(G \mid T)$, the likelihood of selecting a particular target T with gaze-pointing direction G, which can be calculated based on the eyes endpoint distribution probability density function.

$$P(T \mid G) = \frac{P(G \mid T)P(T)}{P(G)} \tag{1}$$

Therefore, our model needs to learn the endpoint distribution of users' eyes when they select potential targets with different widths and distances to derive the likelihood of selecting a particular target given a gaze-pointing direction. The T with the highest probability will be recognized as the intended target T^* (Equation 2). The likelihood function should be derived from an empirical user study. To this end, our first study served as a data source to help us quantify the distribution of endpoints. The derived parameters of the likelihood function are described in Section 5.1.

$$T^* = \arg\max_T P(T \mid G) = \arg\max_T P(G \mid T) \tag{2}$$

3.1.2 Multimodal Model. Our multimodal model further extends the unimodal model by incorporating head-pointing direction to determine the intended target of selection. In this case, the probability of selecting a target T with a gaze pointing direction G and head pointing direction H can be expressed as $P(T \mid G, H)$. Similarly, according to Bayes' theorem, $P(T \mid G, H)$ can be expanded based on Equation 3.

$$P(T \mid G, H) = \frac{P(G, H \mid T) \cdot P(T)}{P(G, H)} \tag{3}$$

We can use the conditional probability formula on $P(G, H)$ to further expand the equation. Therefore, we derive Equation 4 as follows.

$$P(T \mid G, H) = \frac{P(G, H \mid T) \cdot P(T)}{P(H \mid G) \cdot P(G)} \tag{4}$$

To determine $P(T \mid G, H)$, we need **(1)** $P(T)$, the probability of each object being selected (set to 1/n); **(2)** $P(G)$, the probability of gaze pointing at the direction G (set to a constant value); **(3)** $P(H \mid G)$, the probability of head pointing direction H, given that

the gaze pointing direction G. We treat this value as a constant within a particular 3D space. **(4)** $P(G, H \mid T)$, the probability of gaze pointing direction G and head pointing direction H, for a given target. The likelihood of gaze and head pointing direction $P(G, H)$ should be different for each potential target T.

Therefore, to determine $P(T \mid G, H)$, our primary concern is to compute $P(G, H \mid T)$, meaning the likelihood of gaze pointing at G and head pointing at H for each potential target (i.e., $P(G, H)^T$ for each T). We further expand the term through the condition probability formula (Equation 5).

$$P(G, H)^T = P(H \mid G)^T \cdot P(G)^T \tag{5}$$

Here, we thus want to determine **(1)** $P(H \mid G)^T$, the probability of head pointing direction H, give a gaze pointing direction G for a specific target. This can be calculated based on the relative probability distribution of the head by treating the gaze pointing direction as the origin for selecting a target. **(2)** $P(G)^T$, the probability of gaze pointing direction G when selecting a target. Both terms can be determined from an empirical user study where we require users to perform gaze-based target selection on objects of different sizes and distances. After computing these values from the probability density functions, our multimodal prediction model can then pick the object with the highest $P(G, H)^T$ as the intended target T^* (Equation 6). The derived parameters of the likelihood functions are described in Section 5.2.

$$\begin{aligned} T^* &= \arg\max_T P(T \mid G, H) \\ &= \arg\max_T P(G, H \mid T) \\ &= \arg\max_T P(H \mid G)^T \cdot P(G)^T \end{aligned} \tag{6}$$

3.2 Study Outline

To build the models, we first conducted a data collection study to explore how factors including target size and target distance can influence endpoint distributions of users' eyes and head. Additionally, we investigate two confirmation mechanisms (blinking and air tap) as they could result in different endpoint distributions. The data from the first study would allow us to fit both models for target prediction. In a second study, we compared three gaze-based selection techniques: target selection with visual boundary criterion (baseline), target prediction with our unimodal model, and target prediction with our multimodal model. This would allow us to quantify the performance improvement brought by the prediction models.

4 USER STUDY 1: DATA COLLECTION

The goal of this study was to collect data on eyes and head endpoints when performing gaze-based target selection in AR HMDs. With the collected data, we were able to better understand how interface features (object size and distance) and input mechanisms (air tap and blinking) may influence endpoint distributions. We could also learn about how users' eyes and head collaborate in a selection. Moreover, we derived two models based on the eye and head endpoints.

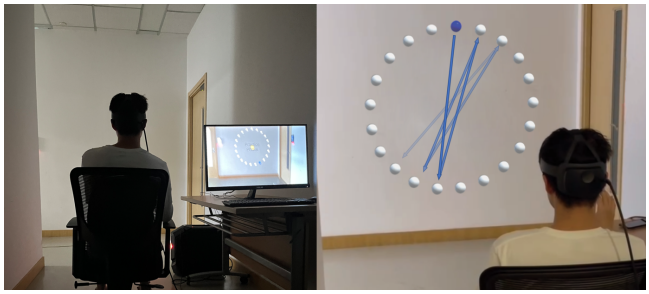

Figure 2: LEFT: Experimental physical environment setup. RIGHT: Holographic projection in the scene. Participants selected the targets (in blue) based on the order indicated by the arrows. The arrows are only for illustration and were not shown to participants.

4.1 Participants and Apparatus

Twenty participants (4 females and 16 males) were recruited from a local university. They were aged between 19 and 25 years ($mean$ = 23.10, $s.d.$ = 1.29). Eight participants had prior experiences with AR HMDs. All of them reported they were able to clearly see all the virtual objects during the experiment.

We used a Microsoft HoloLens 2 AR HMD for this user study. It has a horizontal field-of-view (FoV) of 43°, a vertical FoV of 29°, and a 2K display resolution. While the program was deployed and run on the HoloLens 2, we connected the headset to a desktop PC to monitor participants' behavior in real time. The program was developed using C# in Unity3D with MRTK (Microsoft Mixed Reality Toolkit).

Prior work has suggested that luminance levels of the environment would affect human eyes' pupil variation, and as such affect target selection accuracy [14]. We thus conducted this user study in an empty room without natural light, where participants would face a white wall approximately 2.5 meters away from them, and the holographic projection would be projected two meters in front of the participants. Participants sat in a chair to complete the experiment (see Figure 2.LEFT).

4.2 Task

We adopted a similar Fitts' ring design (ISO9241-9) as in previous works [54, 63]. There were 21 virtual grey spheres in front of a participant. In each trial, one sphere would turn blue, representing the target for that trial, and the participant needed to move the gaze-based cursor to select it. If a target was selected, the next target would appear opposite to it, which yielded a consistent movement amplitude for all selections. Overall, the alternating sequence proceeded in a clockwise manner (see Figure 2. RIGHT). To simulate users' selection behavior in a real-world scenario, we asked them to select their target comfortably and naturally.

4.3 Input Mechanisms

We included two input confirmation mechanisms: air tap and blinking—both of them were common in gaze-based selection but could potentially lead to different eye and head endpoints. Air tap requires participants to perform a pinch action with their dominant hand.

Blinking requires participants to close and open their eyes to confirm the selection.

We followed a previous approach [43, 44] to detect eye blinks: the system checked if the HoloLens 2 headset lost eye tracking data (which happened when eyes were closed) for a certain period of time. We run a pilot study with 5 participants to determine the time duration threshold for blinking with 0.1s, 0.15s, and 0.2s, and asked participants to complete an additional condition using air tap for comparison. We collected 900 trials for each condition. Our results showed that using a 0.1s time threshold would lead to a high error rate (9.7%), most of which were false positives due to users' spontaneous blinking [34]. A threshold of 0.15s received a 6.2% error rate, which was comparable to air tap (6.4%). A 0.2s threshold received the lowest error rate (4.6%). However, participants reported that using a 0.2s threshold could cause extra eye fatigue compared to a 0.15s threshold. Thus, we chose a threshold of 0.15s for the formal experiments. This threshold also aligns with a previous work [43].

Head and eye endpoints were recorded if a selection input was triggered. For air tap, the endpoints were logged at the moment of the registration of a pinch action. For blinking, the endpoints were determined based on the timestamp that the HoloLens 2 headset received the last valid gaze-pointing data before the blinking action. When a user closes her eyes, the HoloLens 2 loses the eye tracking data. During this period, the eye cursor would stay stationary at the position right before the eye-close event. We carefully pre-processed the collected data to ensure that irregular data generated from the two input mechanisms were not used for further analyses (discuss more in Section 4.5).

4.4 Design and Procedure

Our study used a $6 \times 3 \times 2$ within-subjects design with three factors: movement amplitude A (14°, 17°, 20°, 23°, 26°, and 29°), target width W (1.5°, 1.75°, and 2.0°)[1], and confirmation mechanisms CM (air tap and blinking). As mentioned, due to the limited FoV of an AR HMD, it is common that a virtual object appears outside of users' FoV. We conducted preliminary tests to determine the values of A and W so that three different cases were tested. (1) When A was 14° or 17°, users could observe and select the objects without moving their head, i.e., the entire Fitts' ring was inside the FoV. (2) When A was 20° or 23°, it was possible for users to select the target without head movement but mild eye fatigue/discomfit may occur. The head movements were dependent on individual preference [51]. (3) When A was 26° or 29°, users could not see the target from the initial position, so they had to move their head to support the target selection. We adjusted the values of W accordingly to ensure that no object occlusions would happen in any condition.

The experiment was divided into two sessions by CM, the first session used air tap, and the second used blinking. This was to prevent eye fatigue generated in the blinking condition to affect the selection performance in the air tap condition in reverse order. We also ensured that participants had enough rest between the two sessions to prevent potential order effects. Each $A \times W$ combination contained 21 targets, but the data of the first target was removed because of the short amplitude (the distance between the start point and the first target was 0.5^*A). In each CM condition, we applied a Latin-Square design to W and A accordingly. Eventually, we collected 6 $A \times 3$ $W \times 2$ $CM \times 20$ participants $\times$ 20 repetitions = 14,400 trials of data, including eye trajectories, head trajectories, and endpoints. The trajectories and endpoints were in angular representations using a spherical coordinate system, which consisted of an x-axis representing the direction of movement and a y-axis perpendicular to it. The origin was placed at the center of the target [63].

Before the experiment, participants were first asked to complete a questionnaire to collect their demographic information. Then, they were briefed about the research goal, task, and controls, followed by an introduction to the AR HMD. After that, the participants wore the HMD and run an eye calibration procedure. A training session without a time limit was given before the formal trials. As a result of the inaccuracy of the calibration system, we took the following steps to carefully calibrate the eye tracking system: (1) maintaining the same physical environment for all participants to minimize the influence of external factors, such as lighting and background distractions; (2) requiring participants to follow the calibration procedure strictly; (3) during the training phase, we proactively asked the participants if they felt the cursor did not follow their eye movements or was inaccurate when looking at a certain position, and if so, they were asked to re-calibrate their eyes. Participants could practice as much as they wanted to ensure that they got familiar with the device and the controls.

We required participants to orally say 'start experiment' before aligning their head and starting the formal trials. The head positions were calibrated at the beginning of the training session and formal trials. The whole experiment lasted approximately 50 minutes per participant.

4.5 Outliers Removal

We removed the following three types of outliers: (1) Data trials that exceeded three times the standard deviations from the averaged results in the x-axis, y-axis, or movement time following previous works [25, 63]. Through this process, 155 (or 2.15% of the total number of trials) and 144 trials (1.58%) were removed from the air tap and blinking condition, respectively. (2) After implementing step 1, we could still find data trials that contained abnormally short trajectories, which could be because of participants mistakenly confirming the selection twice (i.e., double-confirmation). Therefore, We further removed 286 (3.79%) in the air tap condition and 436 (6.05%) trials in the blinking condition if the selection time was smaller than 0.4s or the distance between the eye endpoint and the previous target center was less than 4°. (3) We also removed the sub-sequential data trial of the double-confirmation trial if the participant did not select the target correctly. We wanted to minimize the impact of double-confirmation altering the movement trajectory of the participants once they detected the mis-triggering of the selection. We removed 269 (3.73%) and 349 (4.84%) trials from the air tap and the blinking condition, respectively. In total, 710 trials of data were removed in the air tapping condition, and 899 trials of data were removed in the blinking condition, representing 9.8% and 12.4% of total trials in each condition.

[1]The absolute distance were 488.60mm, 539.30mm, 698.00mm, 802.70mm, 907.40mm, and 1012.70mm for A, 52.34mm, 61.08mm, and 69.80mm for W, and 2m for *depth*.

In the following section, we present empirical results based on eye endpoints (Section 4.6), relative head endpoints (Section 4.7), and two confirmation mechanisms (Section 4.8). We then provide our interpretation/discussion of the results (Section 4.9).

4.6 Results on Eye Endpoints

4.6.1 Normality tests and data processing. We first run Kolmogorov-Smirnov (KS) tests to examine the normality of the distribution of endpoints in the x- and y-axis. The KS tests were not able to reject the hypothesis that the data followed a Gaussian distribution. Therefore, for simplicity, we treated all sets of endpoints to be normally distributed. We then used maximum likelihood estimation (MLEs, `mle()` in MATLAB) to obtain the mean μ and standard deviation σ of the Gaussian distribution. The correlation ρ between the endpoints in two axes was calculated using the function of `corrcoef()` in MATLAB. We also calculated the average movement time of target selection ($Time$) for each condition ($A \times W \times CM$) as an additional measure. We derive the values of the following six dependent variables based on our data: μ_x, σ_x, μ_y, σ_y, ρ, and $Time$. Figure 3 shows example distributions of eye endpoints when W is 1.5°.

4.6.2 Significance Tests. We conducted repeated-measures ANOVA (RM-ANOVA) regarding the six dependent variables. Table 1 summarizes the factors that had a statistically significant main effect on the dependent variables (please refer to our supplementary materials for the full results). The degree of freedom was adjusted by Greenhouse-Geisser if the sphericity was not met. Overall, the results indicated that W had a significant main effect on μ_x and σ_x; A had a significant main effect on μ_x and σ_y; while CM had a significant main effect on σ_y. Additionally, we found a significant interaction effect of $W \times CM$ on ρ.

4.6.3 Linear Regressions. We found strong linear relationships between the factors and the dependent variables through linear regressions (see Figure 4). According to the linear regression results, we identified that: $\mu_x = -0.1541W + 0.1399$ ($R^2 = 0.9275$), $\mu_x = 0.0084A - 0.3101$ ($R^2 = 0.8824$), $\sigma_x = 0.1243W + 0.4115$ ($R^2 = 0.9829$), and $\sigma_y = 0.0071A + 0.4307$ ($R^2 = 0.9444$).

Table 1: Significant results of RM-ANOVAs for the eye-based endpoints in Study 1 ($\alpha = .05$). DV represents dependent variable.

Factor	*DV*	df_{effect}	df_{error}	F	p	η_p^2
W	μ_x	1.722	32.716	3.549	.047	.157
A	μ_x	3.938	74.822	6.071	.000	.242
W	σ_x	2	35.728	4.304	.023	.185
A	σ_y	3.305	63.642	5.318	.002	.219
CM	σ_y	1	19	9.28	.007	.328
$W \times CM$	ρ	1.985	37.712	3.432	.043	.153

4.7 Results on Relative Head Endpoints

The head often moves in conjunction with the eyes when selecting a target using the eyes. The technique of using the eyes and head simultaneously for target selection has previously been explored [52].

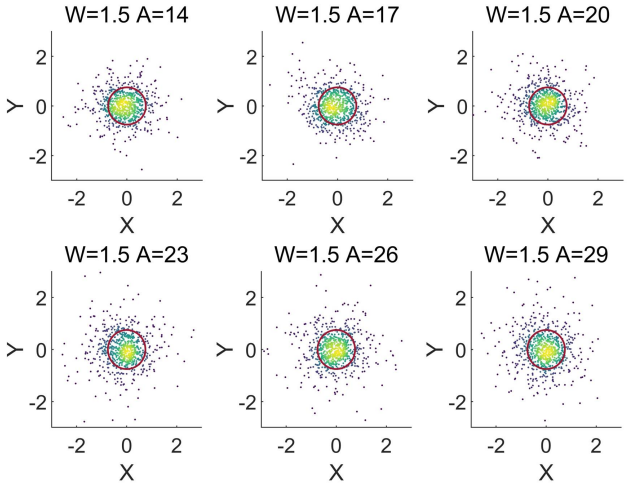

Figure 3: Distribution density map of eye endpoints when W is 1.5°. The red circles represent the boundaries of the targets.

To explore their potential collaborative behavior, we define a *relative head endpoint* as the relative position (in angular representation) of the head endpoint with regard to the corresponding eye endpoint in a selection trial. That is, $(x^{RHE}, y^{RHE}) = (x^{HE}, y^{HE}) - (x^{EE}, y^{EE})$ where RHE stands for Relative Head Endpoint, HE stands for Head Endpoint, and EE stands for Eye Endpoint. This allowed us to quantify the movement behavior of the head in relation to the eyes. Figure 5 shows examples of distributions of relative head endpoints when W is 1.5°. We treated the direction of gaze movement towards a target as the positive direction of the x-axis.

4.7.1 Normality tests and data processing. Similar to eye endpoints data, we run KS tests to examine the normality of distribution of relative head endpoints in the x- and y-axis. The KS tests were not able to reject the normality hypothesis, and we treated all sets of endpoints to be normally distributed. We also calculated μ_x^{head}, σ_x^{head}, μ_y^{head}, σ_y^{head}, and ρ^{head} values as in the previous section. Note that, we use a superscript of *head* in the math expressions hereafter to represent relative head endpoints.

4.7.2 Significance tests. Results of RM-ANOVAs showed that W, A, and CM all had significant main effects on μ_x^{head}. Meanwhile, W and A had an interaction effect on μ_x^{head}. Additionally, A had a significant main effect on σ_x^{head}. Factors that led to statistically significant main effects were summarized in Table 2. Please refer to the supplementary materials for the full results.

4.7.3 Linear Regressions. Due to the interaction effect of $W \times A$, we did not perform linear regressions of A and W on μ_x^{head} separately but combined both of them into a single regression as shown in Figure 6. The expression for the air tap condition is $\mu_x^{head} = -1.8068 - 1.0664W - 0.1960A$ ($R^2 = 0.8248$), and for the blinking condition is $\mu_x^{head} = -2.4852 - 0.9959W - 0.2069A$ ($R^2 = 0.8620$). We also performed linear regression of the main statistical effect between A and σ_x^{head}, with the final expression as $\sigma_x^{head} = 5.4419 - 0.0142A$ ($R^2 = 0.7307$).

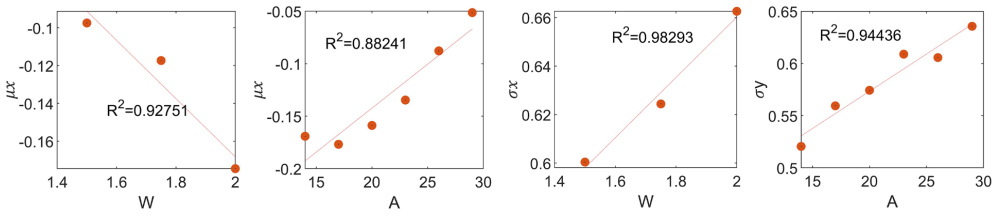

Figure 4: LEFT: Linear regression of the mean of μ_x on *W*. MIDDLE-LEFT: Linear regression of the mean of μ_x on *A*. MIDDLE-RIGHT: Linear regression of the mean of σ_x on *W*. RIGHT: Linear regression of the mean of σ_y on *A*.

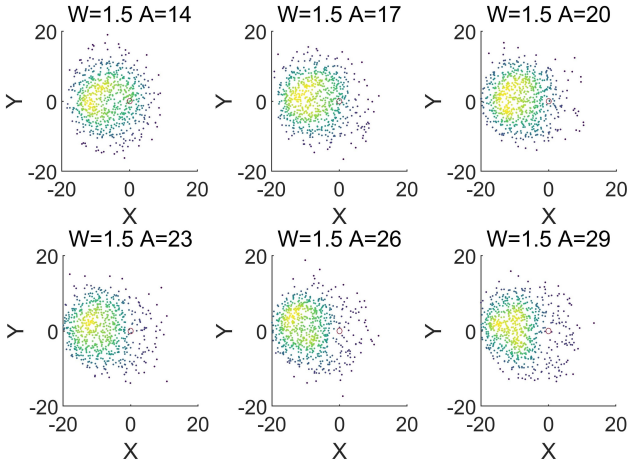

Figure 5: Distribution density map of relative head endpoints.

Table 2: Significant results of RM-ANOVAs for the relative head-pointing endpoint in Study 1 ($\alpha = .05$).

Factor	*DV*	df_{effect}	df_{error}	*F*	*p*	η_p^2
W	μ_x^{head}	1.095	20.797	18.728	.000	.496
A	μ_x^{head}	2.763	52.488	5.262	.004	.217
CM	μ_x^{head}	1	19	5.280	.033	.217
$W \times A$	μ_x^{head}	1.999	37.879	3.371	.009	.151
A	σ_x^{head}	3.029	57.556	2.752	.050	.127

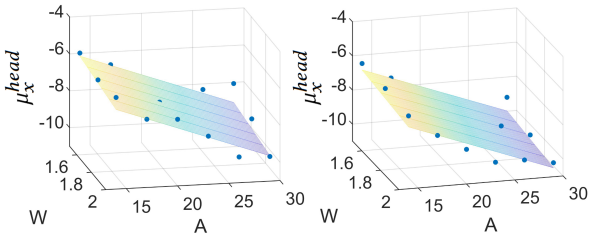

Figure 6: Regressions for relative head orientation in Study 1. μ_x^{head} was fitted to different $W \times A$ combinations using planes. LEFT: Fitting results for the air tap condition. RIGHT: Fitting results for the blinking condition.

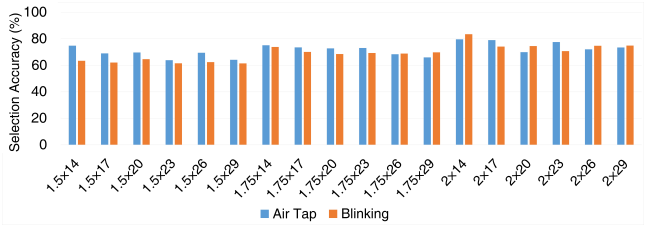

Figure 7: Comparison of selection accuracy for both air tap and blinking conditions in each $W \times A$ combination.

4.8 Results on the Two Confirmation Mechanisms

4.8.1 Eye and Relative Head Endpoints. From Table 1, *CM* has a significant main effect on σ_y in eye-based endpoint distribution. Results showed that gaze-based selection using air tap as the confirmation mechanism ($mean = 0.54, s.d. = 0.19$) led to a smaller σ_y than using blinking ($mean = 0.63, s.d. = 0.25$). In addition, *CM* also has a significant main effect on μ_x^{head} (Table 2), with air tap ($mean = -7.8, s.d. = 3.15$) has a larger μ_x^{head} than blinking ($mean = -8.68, s.d. = 2.52$).

4.8.2 Movement Time (MT). We verified Fitts' law by performing a linear regression between the index of difficulty (ID) and movement time (MT). The fitting results of air tap $R^2 = 0.8281$ and blinking $R^2 = 0.8917$ were high, indicating that gaze-based selection with both confirmation mechanisms followed Fitts' law. Results from a paired sample t-test showed that there was a significant effect of *CM* on MT ($t_{17} = -5.153, p < .001$). Air tap ($mean = 1.31s, s.d. = .16$) led to a shorter time than blinking ($mean = 1.38s, s.d. = .14$).

4.8.3 Selection Accuracy. We compared the two confirmation mechanisms (*CM*) regarding their accuracy (i.e., whether an endpoint falls within the target boundaries). The selection accuracy in each $W \times A$ combination regarding the two *CM* conditions is summarized in Figure 7. The average selection accuracy was 71.8% for air tap and 69.3% for blinking. Results from a paired sample t-test revealed a significant main effect of *CM* on selection accuracy ($t_{17} = -2.278, p = .036$).

4.9 Discussion

4.9.1 Gaze-based selection endpoint distributions. An eye selection endpoint was produced once a user confirmed the selection in a trial. After the repetitive selection process, the eye endpoints formed

a distribution, which was shown to follow a bivariate Gaussian (normal) distribution. Our significant tests and linear regression analyses suggested that the distribution mean μ and standard deviation σ varied as we changed W and A.

Our results suggest that both W and A could influence μ_x, potentially in a linear relationship. Specifically, increasing W decreased μ_x (negative linear relationship $R^2 = 0.93$), and increasing A increased μ_x (positive linear relationship $R^2 = 0.88$). The influence of W on μ_x could be potentially attributed to the "lazy effect", as identified by a previous work [63]. That is, participants were more inclined to shorten their movement distance by performing the selection earlier. This phenomenon got amplified as the target width increased. However, the relationship between A and μ_x showed that the larger the A, the closer the eye-based endpoint was to the center of the target along the x-axis. This was unexpected and in contrast to the relationship of A and μ_x for head-based selection mentioned in [63].

The positive correlation between A and μ_x may not be solely due to the eye behavior—head and eye-head coordination may be another factor. As A increases, μ_x^{head} is further away from its origin while μ_x gets closer to the eye endpoint (see Figures 4.MIDDLE-LEFT and 6). This means participants moved their head more when A increased, because the limited FoV of the AR headset forced participants to move their head when A was large and targets were outside of their vision. With the support of head movements, participants can select the target with their eye gaze more comfortably; thus, this effect could lead to a positive correlation between A and μ_x, with μ_x getting closer to the target center when the distance was larger. On the other hand, when A was small and the targets were within the FoV, participants may not move their head, and the μ_x did not show such a trend (see A=14 and 17 in Figure 4.MIDDLE-LEFT).

Our results also showed that W has a significant main effect on σ_x (see Table 1). Figure 4.MIDDLE-RIGHT revealed a positive correlation between W and σ_x. These results are in line with the findings in finger-based selection in touch screens [6] and head-based selection in VR HMD [63]. In addition, A has a significant main effect on σ_y: the greater the A, the higher the σ_y. That is, as A increases, the eye endpoint in the perpendicular direction of movement becomes more spread. This finding is also in line with prior work [22].

4.9.2 Relative head endpoints distribution. A relative head endpoint represents the relative position of the head endpoint with regard to the corresponding eye endpoint. We found significant main effects of W, A, and $W \times A$ on μ_x^{head}. From the results, one observation was that the values of μ_x^{head} were always smaller than zero. This resonated with the findings by Sidenmark et al. [51, 52] who showed that while head movements are often coupled with eye movements, users seldom move their head fully toward a target to which they have shifted their gaze. Since we treated the eye movement direction as the positive axis, the relative head endpoints always resulted in negative values.

From Figure 6, we also observed that for both *CM* conditions, relative head endpoints became closer to the eye endpoints as A increased. Previous research suggested that users could move their heads earlier than their eyes to help locate the target if the target was outside of the FoV [51]. In our case, a larger A makes the target to be further away from the center of the view, even outside of the FoV. Therefore, we hypothesized that with a large A, the head may provide more assistance in locating the target, resulting in a close distance between the head and eye endpoint. However, the influence of W and $W \times A$ on μ_x^{head} was hard to parse from our results. It was maybe because the difference between different levels of W (1.5°, 1.75°, and 2.0°) was too small to produce meaningful patterns.

Furthermore, we observed a negative linear correlation between A and σ_x^{head}. This suggested that when the participants utilized their heads to assist with target selection, the variances in their head endpoints became smaller when the target was located further away from the center of the view.

4.9.3 The impact of different confirmation mechanisms. *CM* has a significant main effect on σ_y in the eye-based endpoint distribution and μ_x^{head}. Gaze-based selection using air tap led to fewer variances of endpoint in the y-axis (i.e., the perpendicular direction of movement) and made the relative head endpoints closer to the gaze endpoints than using blinking. Additionally, gaze-based selection using air tap was slightly faster and more accurate than using blinking.

All these findings imply that participants were able to perform a more stable, efficient, and accurate gaze-based selection using the air tap as the input confirmation mechanism. This is in line with our expectation—air tap uses hand as an additional confirmation modality which does not interrupt the eye-based target pointing procedure. On the contrary, blinking uses the eyes for both pointing and confirmation so it became difficult for the eyes to fine-tune their direction for accurate selection. Furthermore, since blinking is an eye-based approach, it is also affected by the eye-tracking modules. Thus, users may have a more diverse behavior pattern and worse performance when using blinking as the input mechanism in gaze-based selection. Given the significant main effect of *CM* on σ_y and μ_x^{head}, we constructed two dedicated models for the two confirmation mechanisms (see next section).

The average selection accuracy was 71.8% for air tap and 69.3% for blinking in our study, which was relatively low compared to previous works [35]. One possible reason is that our targets were quite small (from 1.5° to 2.0°) which represented a more challenging scenario for gaze-based selection with our current equipment [63].

4.9.4 Summary - Study 1. In this study, we collected eye and head endpoint data in a gaze-based selection task in AR HMD and explored how the factors including W, A, and *CM* can affect the endpoint distributions. We found target widths W, movement amplitudes A, and input confirmation mechanisms *CM* all had significant effects on either center or variance of the eyes or head endpoint distributions. According to these results, our next step is to fit the unimodal model and multimodal model.

5 MODEL FITTING

In Section 3.1, we described that to calculate the likelihood of a user selecting each target (i.e., making a target prediction), we needed to collect the following information from the data: (1) the endpoint distribution of eyes when users select potential targets with different widths and distances for both unimodal and multimodal models; and (2) the endpoint distribution of the head by treating the gaze

pointing direction as the origin (i.e., the relative head endpoint distribution) when users select potential targets for the multimodal model. These two endpoint distributions were modeled as bivariate Gaussian distributions as in previous works [6, 63], so we needed to obtain $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ for each of them, where $\boldsymbol{\mu}$ represents the center of the distribution (a 2D vector, in both x- and y-axis) and Σ is the covariance matrix, containing the variances of the distribution and the correlations (a 2×2 matrix).

For both models, we first identified the factors that had significant effects on the response variables (e.g., from the first study, we found W and A had significant main effects on μ_x). As in previous research [63], we then applied a linear regression to quantify the relationship between the factors and the response variable. For example, we built a linear regression model $\mu_x = eW + fA + g$ for eye endpoint distribution and determined variable e, f, g based on the empirical data from Study 1. We assumed the correlations between the two axes to be 0 for simplicity.

5.1 Modeling the likelihood function of endpoint distribution for eyes-only

Based on the significance testing results from the first study, we constructed a bivariate-Gaussian distribution model for gaze-based pointing selection in AR HMD. This endpoint distribution model of eyes can be used as the likelihood model in Bayesian theory, thus forming our final unimodal model. The model parameters are shown below.

$$\boldsymbol{\mu} = \begin{bmatrix} eW + fA + g \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} (a\,\mathrm{W} + b)^2 & 0 \\ 0 & (c\,\mathrm{A} + d)^2 \end{bmatrix} \quad (7)$$

W and A are the target width and the movement amplitude, while a, b, c, d, e, f, g are constants. Using the data collected in the first study, as mentioned in Section 4.6.3, we have: $a = 0.1243$, $b = 0.4115$, $e = -0.1541$, $f = 0.0084$ and $g = -0.1702$. Since the input confirmation mechanisms CM has a significant main effect on σ_y, we have $c = 0.0078$, $d = 0.371$ when CM is air tap, and $c = 0.0064$, $d = 0.4904$ when CM is blinking.

5.2 Modeling the likelihood function of relative endpoint distribution of head

To construct our multimodal model, we need a relative head endpoint distribution model in addition to the eye endpoint distribution model. Therefore, we quantified how factors may influence the relative head endpoint distribution through another bivariate Gaussian distribution by treating the eye position as an origin. The model parameters are presented below.

$$\boldsymbol{\mu} = \begin{bmatrix} dW + eA + f \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} (a\,\mathrm{A} + b)^2 & 0 \\ 0 & c \end{bmatrix} \quad (8)$$

a, b, c, d, e, f are constant values. Given the statistical main effect between the CM for μ_x, we have two sets of d, e, f for the two CM. Our data suggest, as mentioned in Section 4.7.3, $a = -0.0142$, $b = 5.4419$. $c = 5.2460$ (the mean value of σ_y), $d = -1.0664$, $e = -0.1960$, $f = -1.8968$ for air tap, and $d = -0.9959$, $e = -0.2069$, $f = -2.4852$ for blinking.

6 USER STUDY 2: SELECTION PERFORMANCE EVALUATION

The purpose of this study is threefold. (1) To assess whether our two models have enhanced selection accuracy compared to the approach based on Visual Boundary Criterion (VBC). (2) To verify the reliability and robustness of our models in complex environments. (3) To check the predictive capability of our models in different FoVs, especially in the out-of-view cases.

6.1 Participants and Apparatus

Sixteen participants were recruited (3 females and 13 males; aged between 22 and 26, $mean = 23.50, s.d. = 1.366$). Thirteen participants had prior experience with AR HMDs. All of them reported they could see the virtual environment clearly during the experiment. The experiments were conducted using the same apparatus and in the room with the same setup as in the first study.

6.2 Task, Design, and Procedure

We compared three techniques in this experiment: VBC-based selection (baseline scenario, the gaze pointer can successfully select an object only if it is within the object), target prediction based on the unimodal model, and target prediction based on the multimodal model. VBC has been used as a common interaction mechanism in virtual environments, such as text entry [44], UI interaction [10] and game context [37]. Notably, VBC is a single-step approach that only requires users to perform one selection input. There are other eye-based selection techniques that could be more accurate [31, 35]. However, many of them used a multi-step approach, where users determine the selection region first and then fine-tune their selection in a second stage. In this work, we chose VBC-based selection as our baseline for comparison because it is simple and prevalent in off-the-shelf applications.

We modified the Fitts' ring with four distractors surrounding the intended target according to a previous work [63]. The intended target was in blue, while the distractors were in grey and displayed at the target's top, down, left, and right sides (see Figure 8). The widths of the distractors were set to 2° in all trials. The distances between the distractors and the target were set the same as the width of the target (W) in each trial. In other words, if the W decreases, the distance between the distractor and the target also decreases, rendering a more challenging selection task. Distractors may alter a user's selection behavior to avoid potential errors and we wanted to verify if our models could still be useful in this case. Like in the first study, the participants were asked to select the target comfortably and naturally.

A $3 \times 3 \times 2$ within-subjects design was used in this study with three factors: movement amplitude A (20°, 30°, and 40°), target width W (0.5°, 1°, and 2°)[2], and CM (air tap and blinking). We have included an extreme condition of $A = 40°$ in this study as we wanted to test the generalizability of our model to conditions with longer head movement distances. In terms of target width, we also included another extreme case of $W = 0.5°$, where the target was very small and the distractors were very close to the target. These conditions represent complex environments that users may face

[2]The absolute distance were 698mm, 1047mm and 1396mm for *A*, 8.75mm, 17.45mm, and 34.90mm for *W*, and 2m for *depth*.

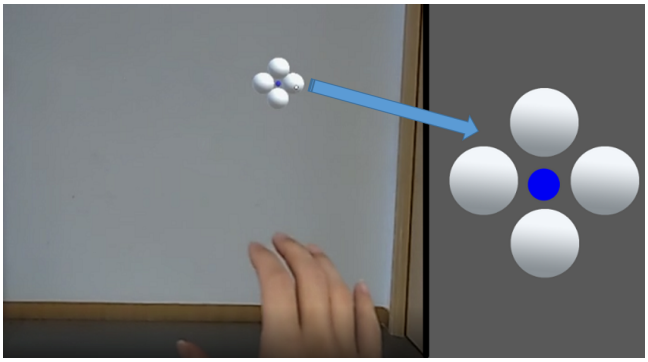

Figure 8: An example of an experimental task in Study 2. The target is surrounded by four distractors.

in a real application (e.g., immersive analytics [45]. Each $A \times W$ combination still contained 21 targets, while only the intended target and surrounded distractors would appear in each trial to avoid the occlusions. The order of the conditions was counterbalanced as in the first study. In addition, we followed the same experimental procedure. Participants took approximately 15 minutes to complete the whole experiment. In total, we collected 3 $A\times$ 3 $W\times$ 2 $CM\times$ 16 participants $\times$ 20 repetitions = 5760 trials of data in this study.

Once the participants confirmed a selection, the program automatically calculated its correctness based on the endpoint(s) according to three selection methods (VBC, unimodal model, and multimodal model; the variable referred to as *technique*). We decided not to provide visual feedback of a selection (e.g., highlighting an object if the cursor is "on" the target) because (1) rapid, noisy eye movements may cause the feedback to bounce between objects when selecting a target in a dense environment, which could be confusing and mentally fatiguing [42, 61] and (2) we wanted to test our proposed techniques without altering the visual interfaces that provide implicit assistance.

6.3 Results

We pre-processed the data as in the first study. In total, we removed 65 trials (2.2%) in the air tap condition and 139 trials (4.8%) in the blinking condition. Figure 9 shows the average selection accuracy of VBC-based selection, unimodal model-supported selection, and multimodal model-supported selection in each $W \times A$ combination.

Results of RM-ANOVAs showed that A, W, and CM all had a significant main effect on selection accuracy ($F_{2,30} = 224.171$, $p < .001$; $F_{2,30} = 64.771$, $p < .001$; and $F_{1,15} = 10.945$, $p = .005$, respectively). In addition, *technique* also had a significant main effect on selection accuracy ($F_{2,30} = 13.854$, $p < .001$). The average accuracy of VBC = 62.43%, unimodal model = 95.02%, and multimodal model = 94.28%.

Furthermore, we performed RM-ANOVAs to investigate whether there are significant differences between the three techniques in terms of selection accuracy in each $W \times A \times CM$ condition. Results of post-hoc pairwise comparisons adjusted by Bonferroni correction are summarized in Figure 9. Overall, we found significant performance differences between VBC and the two prediction-based approaches (unimodal and multimodal models) in almost all conditions for both confirmation mechanisms.

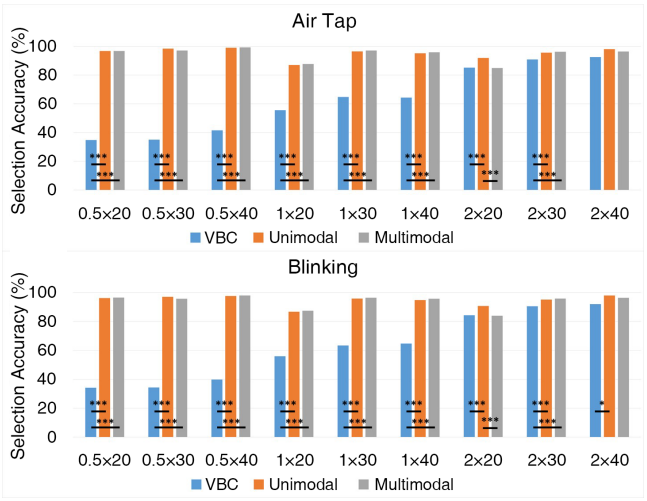

Figure 9: Selection accuracy of the three techniques in air tap (TOP) and blinking (BOTTOM) conditions for each $W \times A$ combination.

For small targets, i.e., with a width of 0.5°, the accuracy of target selection using VBC is particularly low (below 41.50%). Compared to VBC, the two prediction-based approaches (unimodal model and multimodal model) increased the selection accuracy considerably. Notably, for the condition W=0.5° and A=20°, the accuracy improvement was the most significant—the accuracy gains of our unimodal model reached 61.98% and 61.93% for air tap and blinking, respectively. The multimodal model produced similar results to the unimodal model. For the condition of W=0.5° and A=40°, the accuracy of both our models in air tap is greater than 99%. Both examples illustrate that our models could work particularly well for small targets. For larger targets, with a maximum target width of 2° in our setting, VBC can achieve an accuracy of around 89.2%. Our model further improved the accuracy to around 93.3%.

6.4 Discussion

6.4.1 VBC vs. target prediction models. We found that while decreasing W greatly affected the accuracy for VBC-based selection, the target prediction accuracy based on our models remained stable. Our models significantly improved target selection accuracy in conditions where the target size was less than or equal to 1°. Notably, for the smallest target (W = 0.5°), the models achieved 99% accuracy with air tap and over 95% with blinking when both A = 30° and A = 40°. These results demonstrated that our models can achieve highly accurate target prediction without altering the visual appearance of a user interface and can significantly boost the accuracy for small targets compared to the VBC-based approach.

6.4.2 Unimodal model vs. multimodal model. We designed the multimodal model to further incorporate the head information to help eliminate the noise of eyes or their tracking devices. Our results showed that the unimodal model and the multimodal model achieved comparable performance across different experimental conditions. We found that the multimodal model could be more useful for scenarios that require longer head movement distances.

When targets were placed far away (e.g., A was larger than 20°), the multimodal model demonstrated excellent performance regardless of confirmation mechanisms, especially for small targets. However, when the target was closer to a user's front-facing direction, e.g., A was less than 20°, the head movement did not contribute much to the gaze-pointing movement in such cases and created more uncertainty to the prediction. In other words, the uncertainty of head pointing direction was high when A is small, thus reducing the prediction accuracy of the multimodal model. For example, the selection accuracy for the multimodal model in the condition $W = 2°$ and $A = 20°$ was relatively low as compared to the unimodal model (Figure 9). Therefore, we recommend using the unimodal model for simplicity. The multimodal model may achieve superior performance when eyes are actively looking for targets, so the head-pointing information may provide more confirmation on the "region" the target might locate. However, future work is required to verify this assumption.

In summary, both target prediction models are more robust and stable than the baseline. They can be applied to assist gaze-based target selection in scenarios with a variety of targets and different confirmation mechanisms.

6.4.3 Generalizability and Reproducibility. Our models leveraged linear regressions to approximate how interface features like object size and distance may influence eye or head endpoint distributions. Linear regression is a relatively more robust approach compared to nonlinear-based models which could cause overfitting. Therefore, we are confident that our model can be generalized to other similar conditions. Notably, we considered extreme cases (e.g., $W = 0.5°$ and $A = 40°$) and distractors in the second study, which were not counted when we fit the model with the data from the first study. Our models were still able to produce highly accurate predictions in these unseen conditions.

Future work may verify our models with new eye-tracking devices and interface layouts. We hypothesized that while the fitting parameters may be slightly different, the general gaze-based selection behavior may not change significantly. But this assumption needs to be evaluated in future research. We have opensourced our dataset collected from the two studies that are available at https://github.com/Yushi-Wei/Modeling-Endpoint-Distributions. Therefore, future research can reproduce our models and compare their results with ours.

7 LIMITATIONS AND FUTURE WORK

We identified some limitations and possible avenues for future research. First, while our model has been verified to produce promising results regarding static targets with a circular shape in a controlled environment with simplified background and lighting, we plan to extend our findings to targets with arbitrary shapes and complex backgrounds and lighting environments in the future. We also want to further investigate how visual feedback (e.g., highlighting an object when the cursor is "on" the object) may influence endpoint distributions.

Second, we used a controlled experimental scheme (the Fitts' ring), where participants could easily predict the next target that they needed to select. This was reasonable to collect data on pointing behavior as searching was minimized. However, in many AR target selection scenarios, searching might be necessary and a user's eye and head behavior might be different. In some other cases, targets might not even be located in front of a user; instead, they may appear behind them. In these cases, the user must rotate their body to select it. Previous work has demonstrated that there is a collaboration between the limbs and the head and eyes during target selection [51]. It will be interesting to see how the behavior of the searching and the movements of limbs can be incorporated into the models.

Third, we used the most commonly accepted VBC as a benchmark when evaluating our model for better understanding, and in the future, we prefer to use more diverse ways (e.g., selecting the target closest to the cursor) to compare with the model to achieve a more comprehensive understanding for different application scenarios.

Fourth, previous research indicated that the default eye calibration system of the HoloLens 2 could lead to a wide range of horizontal and vertical inaccuracies [1]. While we followed a strict process of eye-tracking calibration, we could not fully eliminate all errors. Therefore, our models might have incorporated those inaccuracies when predicting the targets. Future work should replicate the study when using a different eye tracker (e.g., HTC Vive Pro Eye).

8 CONCLUSION

In this paper, we presented two novel models applicable to AR HMD devices for predicting gaze-based selection. Both models (unimodal and multimodal) are based on endpoint distributions and the Bayesian theory. In contrast to previous work, we propose the concept of multi-modality in one of the models, considering the collaborative role of the eyes and the head. We built our models through a data collection study and tested their effectiveness in another follow-up study. Our models significantly improved the accuracy of object selection for small targets, achieving nearly 61% improvement for targets with a visual width of 0.5° and approximately 32% improvement for targets with a visual width of 1°. The study results also indicated that the multimodal model based on eye and head endpoint distribution achieved similar performance as the unimodal model with only the eye endpoint distribution.

ACKNOWLEDGEMENT

We thank the volunteers who participated in our studies. We also thank the reviewers for their valuable time and insightful comments that helped improve this paper. This research was partly funded by Xi'an Jiaotong-Liverpool University Special Key Fund (#KSF-A-03), the National Science Foundation of China (#62272396; #62207022), and the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (#22KJB520038).

REFERENCES

[1] Samantha Aziz and Oleg Komogortsev. 2022. An Assessment of the Eye Tracking Signal Quality Captured in the HoloLens 2. In *2022 Symposium on Eye Tracking Research and Applications* (Seattle, WA, USA) *(ETRA '22)*. Association for Computing Machinery, New York, NY, USA, Article 5, 6 pages. https://doi.org/10.1145/3517031.3529626

[2] R. Bates and H. O. Istance. 2002. Why Are Eye Mice Unpopular? — A Detailed Comparison of Head and Eye Controlled Assistive Technology Pointing Devices.

In *Universal Access and Assistive Technology*, Simeon Keates, Patrick Langdon, P. John Clarkson, and Peter Robinson (Eds.). Springer London, London, 75–86.

[3] Benjamin B. Bederson. 2000. Fisheye Menus. In *Proceedings of the 13th Annual ACM Symposium on User Interface Software and Technology* (San Diego, California, USA) *(UIST '00)*. Association for Computing Machinery, New York, NY, USA, 217–225. https://doi.org/10.1145/354401.354782

[4] Matthias Bernhard, Efstathios Stavrakis, Michael Hecher, and Michael Wimmer. 2014. Gaze-to-Object Mapping during Visual Search in 3D Virtual Environments. *ACM Trans. Appl. Percept.* 11, 3, Article 14 (aug 2014), 17 pages. https://doi.org/10.1145/2644812

[5] Xiaojun Bi, Yang Li, and Shumin Zhai. 2013. FFitts Law: Modeling Finger Touch with Fitts' Law. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) *(CHI '13)*. Association for Computing Machinery, New York, NY, USA, 1363–1372. https://doi.org/10.1145/2470654.2466180

[6] Xiaojun Bi and Shumin Zhai. 2013. Bayesian Touch: A Statistical Criterion of Target Selection with Finger Touch. In *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology* (St. Andrews, Scotland, United Kingdom) *(UIST '13)*. Association for Computing Machinery, New York, NY, USA, 51–60. https://doi.org/10.1145/2501988.2502058

[7] Emilio Bizzi. 2011. Eye-Head Coordination. *Comprehensive Physiology* (2011), 23–24.

[8] Renaud Blanch and Michael Ortega. 2011. Benchmarking Pointing Techniques with Distractors: Adding a Density Factor to Fitts' Pointing Paradigm. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 1629–1638. https://doi.org/10.1145/1978942.1979180

[9] Jonas Blattgerste, Patrick Renner, and Thies Pfeiffer. 2018. Advantages of Eye-Gaze over Head-Gaze-Based Selection in Virtual and Augmented Reality under Varying Field of Views. In *Proceedings of the Workshop on Communication by Gaze Interaction* (Warsaw, Poland) *(COGAIN '18)*. Association for Computing Machinery, New York, NY, USA, Article 1, 9 pages. https://doi.org/10.1145/3206343.3206349

[10] Jonas Blattgerste, Patrick Renner, and Thies Pfeiffer. 2018. Advantages of Eye-Gaze over Head-Gaze-Based Selection in Virtual and Augmented Reality under Varying Field of Views. In *Proceedings of the Workshop on Communication by Gaze Interaction* (Warsaw, Poland) *(COGAIN '18)*. Association for Computing Machinery, New York, NY, USA, Article 1, 9 pages. https://doi.org/10.1145/3206343.3206349

[11] Riccardo Bovo, Daniele Giunchi, Ludwig Sidenmark, Hans Gellersen, Enrico Costanza, and Thomas Heinis. 2022. Real-Time Head-Based Deep-Learning Model for Gaze Probability Regions in Collaborative VR. In *2022 Symposium on Eye Tracking Research and Applications* (Seattle, WA, USA) *(ETRA '22)*. Association for Computing Machinery, New York, NY, USA, Article 6, 8 pages. https://doi.org/10.1145/3517031.3529642

[12] Ishan Chatterjee, Robert Xiao, and Chris Harrison. 2015. Gaze+Gesture: Expressive, Precise and Targeted Free-Space Interactions. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction* (Seattle, Washington, USA) *(ICMI '15)*. Association for Computing Machinery, New York, NY, USA, 131–138. https://doi.org/10.1145/2818346.2820752

[13] Michael Chau and Margrit Betke. 2005. *Real time eye tracking and blink detection with usb cameras.* Technical Report. Boston University Computer Science Department.

[14] Iakov Chernyak, Grigory Chernyak, Jeffrey K. S. Bland, and Pierre D. P. Rahier. 2021. Important Considerations of Data Collection and Curation for Reliable Benchmarking of End-User Eye-Tracking Systems. In *ACM Symposium on Eye Tracking Research and Applications* (Virtual Event, Germany) *(ETRA '21 Full Papers)*. Association for Computing Machinery, New York, NY, USA, Article 1, 9 pages. https://doi.org/10.1145/3448017.3457383

[15] Tor-Salve Dalsgaard, Jarrod Knibbe, and Joanna Bergström. 2021. Modeling Pointing for 3D Target Selection in VR. In *Proceedings of the 27th ACM Symposium on Virtual Reality Software and Technology* (Osaka, Japan) *(VRST '21)*. Association for Computing Machinery, New York, NY, USA, Article 42, 10 pages. https://doi.org/10.1145/3489849.3489853

[16] Brendan David-John, Candace Peacock, Ting Zhang, T. Scott Murdison, Hrvoje Benko, and Tanya R. Jonker. 2021. Towards Gaze-Based Prediction of the Intent to Interact in Virtual Reality. In *ACM Symposium on Eye Tracking Research and Applications* (Virtual Event, Germany) *(ETRA '21 Short Papers)*. Association for Computing Machinery, New York, NY, USA, Article 2, 7 pages. https://doi.org/10.1145/3448018.3458008

[17] Shujie Deng, Nan Jiang, Jian Chang, Shihui Guo, and Jian J. Zhang. 2017. Understanding the impact of multimodal interaction using gaze informed mid-air gesture control in 3D virtual objects manipulation. *International Journal of Human-Computer Studies* 105 (2017), 68–80. https://doi.org/10.1016/j.ijhcs.2017.04.002

[18] Andrew T. Duchowski. 2018. Gaze-based interaction: A 30 year retrospective. *Computers & Graphics* 73 (2018), 59–69. https://doi.org/10.1016/j.cag.2018.04.002

[19] Anna Maria Feit, Shane Williams, Arturo Toledo, Ann Paradiso, Harish Kulkarni, Shaun Kane, and Meredith Ringel Morris. 2017. Toward Everyday Gaze Input: Accuracy and Precision of Eye Tracking and Implications for Design. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 1118–1130. https://doi.org/10.1145/3025453.3025599

[20] Wenxin Feng, Jiangnan Zou, Andrew Kurauchi, Carlos H Morimoto, and Margrit Betke. 2021. HGaze Typing: Head-Gesture Assisted Gaze Typing. In *ACM Symposium on Eye Tracking Research and Applications* (Virtual Event, Germany) *(ETRA '21 Full Papers)*. Association for Computing Machinery, New York, NY, USA, Article 11, 11 pages. https://doi.org/10.1145/3448017.3457379

[21] Edward G Freedman. 2008. Coordination of the eyes and head during visual orienting. *Experimental brain research* 190, 4 (2008), 369–387.

[22] Tovi Grossman and Ravin Balakrishnan. 2005. A Probabilistic Approach to Modeling Two-Dimensional Pointing. *ACM Trans. Comput.-Hum. Interact.* 12, 3 (sep 2005), 435–459. https://doi.org/10.1145/1096737.1096741

[23] D Guitton and M Volle. 1987. Gaze control in humans: eye-head coordination during orienting movements to targets within and beyond the oculomotor range. *Journal of neurophysiology* 58, 3 (September 1987), 427—459. https://doi.org/10.1152/jn.1987.58.3.427

[24] Rorik Henrikson, Tovi Grossman, Sean Trowbridge, Daniel Wigdor, and Hrvoje Benko. 2020. Head-Coupled Kinematic Template Matching: A Prediction Model for Ray Pointing in VR. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376489

[25] Jin Huang, Xiaolan Peng, Feng Tian, Hongan Wang, and Guozhong Dai. 2018. Modeling a Target-Selection Motion by Leveraging an Optimal Feedback Control Mechanism. *Science China Information Sciences* 61, 4 (2018). https://doi.org/10.1007/s11432-017-9326-8

[26] Jin Huang, Feng Tian, Xiangmin Fan, Huawei Tu, Hao Zhang, Xiaolan Peng, and Hongan Wang. 2020. Modeling the Endpoint Uncertainty in Crossing-Based Moving Target Selection. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376336

[27] Jin Huang, Feng Tian, Xiangmin Fan, Xiaolong (Luke) Zhang, and Shumin Zhai. 2018. Understanding the Uncertainty in 1D Unidirectional Moving Target Selection. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3173811

[28] Jin Huang, Feng Tian, Nianlong Li, and Xiangmin Fan. 2019. Modeling the Uncertainty in 2D Moving Target Selection. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology* (New Orleans, LA, USA) *(UIST '19)*. Association for Computing Machinery, New York, NY, USA, 1031–1043. https://doi.org/10.1145/3332165.3347880

[29] Toshiya Isomoto, Toshiyuki Ando, Buntarou Shizuki, and Shin Takahashi. 2018. Dwell Time Reduction Technique Using Fitts' Law for Gaze-Based Target Acquisition. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications* (Warsaw, Poland) *(ETRA '18)*. Association for Computing Machinery, New York, NY, USA, Article 26, 7 pages. https://doi.org/10.1145/3204493.3204532

[30] Robert J. K. Jacob. 1991. The Use of Eye Movements in Human-Computer Interaction Techniques: What You Look at is What You Get. *ACM Trans. Inf. Syst.* 9, 2 (apr 1991), 152–169. https://doi.org/10.1145/123078.128728

[31] Anjali K Jogeshwar, Gabriel J Diaz, Susan P Farnand, and Jeff B Pelz. 2020. The Cone Model: Recognizing gaze uncertainty in virtual environments. *Electronic Imaging* 2020, 9 (2020), 288–1.

[32] Konstantin Klamka, Andreas Siegel, Stefan Vogt, Fabian Göbel, Sophie Stellmach, and Raimund Dachselt. 2015. Look & Pedal: Hands-Free Navigation in Zoomable Information Spaces through Gaze-Supported Foot Input. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction* (Seattle, Washington, USA) *(ICMI '15)*. Association for Computing Machinery, New York, NY, USA, 123–130. https://doi.org/10.1145/2818346.2820751

[33] Yu-Jung Ko, Hang Zhao, Yoonsang Kim, IV Ramakrishnan, Shumin Zhai, and Xiaojun Bi. 2020. Modeling Two Dimensional Touch Pointing. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '20)*. Association for Computing Machinery, New York, NY, USA, 858–868. https://doi.org/10.1145/3379337.3415871

[34] Aleksandra Królak and Paweł Strumiłło. 2012. Eye-blink detection system for human–computer interaction. *Universal Access in the Information Society* 11, 4 (2012), 409–419. https://doi.org/10.1007/s10209-011-0256-6

[35] Mikko Kytö, Barrett Ens, Thammathip Piumsomboon, Gun A. Lee, and Mark Billinghurst. 2018. Pinpointing: Precise Head- and Eye-Based Target Selection for Augmented Reality. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3173655

[36] Michael Land and Benjamin Tatler. 2009. *Looking and Acting: Vision and eye movements in natural behaviour.* Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198570943.001.0001

[37] Michael Lankes and Barbara Stiglbauer. 2016. GazeAR: Mobile gaze-based interaction in the context of augmented reality games. In *International conference on augmented reality, virtual reality and computer graphics*. Springer, 397–406.

[38] Nianlong Li, Feng Tian, Jin Huang, Xiangmin Fan, and Hongan Wang. 2018. 2D-BayesPointer: An Implicit Moving Target Selection Technique Enabled by Human Performance Modeling. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI EA '18)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3170427.3188520

[39] Zhi Li, Maozheng Zhao, Yifan Wang, Sina Rashidian, Furqan Baig, Rui Liu, Wanyu Liu, Michel Beaudouin-Lafon, Brooke Ellison, Fusheng Wang, IV Ramakrishnan, and Xiaojun Bi. 2021. BayesGaze: A Bayesian Approach to Eye-Gaze Based Target Selection. In *Proceedings of Graphics Interface 2021* (Virtual Event) *(GI 2021)*. Canadian Information Processing Society, 231 – 240. https://doi.org/10.20380/GI2021.35

[40] Jen-Shuo Liu, Carmine Elvezio, Barbara Tversky, and Steven Feiner. 2021. Using Multi-Level Precueing to Improve Performance in Path-Following Tasks in Virtual Reality. *IEEE Transactions on Visualization and Computer Graphics* 27, 11 (2021), 4311–4320. https://doi.org/10.1109/TVCG.2021.3106476

[41] Xinyi Liu, Xuanru Meng, Becky Spittle, Wenge Xu, BoYu Gao, and Hai-Ning Liang. 2023. Exploring Text Selection in Augmented Reality Systems. In *Proceedings of the 18th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and Its Applications in Industry* (Guangzhou, China) *(VRCAI '22)*. Association for Computing Machinery, New York, NY, USA, Article 35, 8 pages. https://doi.org/10.1145/3574131.3574459

[42] Feiyu Lu, Difeng Yu, Hai-Ning Liang, Wenjun Chen, Konstantinos Papangelis, and Nazlena Mohamad Ali. 2018. Evaluating Engagement Level and Analytical Support of Interactive Visualizations in Virtual Reality Environments. In *IEEE International Symposium on Mixed and Augmented Reality, ISMAR 2018, Munich, Germany, October 16-20, 2018*, David Chu, Joseph L. Gabbard, Jens Grubert, and Holger Regenbrecht (Eds.). IEEE, 143–152. https://doi.org/10.1109/ISMAR.2018.00050

[43] Xueshi Lu, Difeng Yu, Hai-Ning Liang, and Jorge Goncalves. 2021. IText: Hands-Free Text Entry on an Imaginary Keyboard for Augmented Reality Systems. In *The 34th Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '21)*. Association for Computing Machinery, New York, NY, USA, 815–825. https://doi.org/10.1145/3472749.3474788

[44] Xueshi Lu, Difeng Yu, Hai-Ning Liang, Wenge Xu, Yuzheng Chen, Xiang Li, and Khalad Hasan. 2020. Exploration of Hands-free Text Entry Techniques For Virtual Reality. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 344–349. https://doi.org/10.1109/ISMAR50242.2020.00061

[45] Kim Marriott, Falk Schreiber, Tim Dwyer, Karsten Klein, Nathalie Henry Riche, Takayuki Itoh, Wolfgang Stuerzlinger, and Bruce H Thomas. 2018. *Immersive analytics*. Vol. 11190. Springer.

[46] Michael J. McGuffin and Ravin Balakrishnan. 2005. Fitts' Law and Expanding Targets: Experimental Studies and Designs for User Interfaces. *ACM Trans. Comput.-Hum. Interact.* 12, 4 (dec 2005), 388–422. https://doi.org/10.1145/1121112.1121115

[47] Xuanru Meng, Wenge Xu, and Hai-Ning Liang. 2022. An Exploration of Hands-free Text Selection for Virtual Reality Head-Mounted Displays. In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 74–81. https://doi.org/10.1109/ISMAR55827.2022.00021

[48] Darius Miniotas, Oleg Špakov, and I. Scott MacKenzie. 2004. Eye Gaze Interaction with Expanding Targets. In *CHI '04 Extended Abstracts on Human Factors in Computing Systems* (Vienna, Austria) *(CHI EA '04)*. Association for Computing Machinery, New York, NY, USA, 1255–1258. https://doi.org/10.1145/985921.986037

[49] Ken Pfeuffer, Benedikt Mayer, Diako Mardanbegi, and Hans Gellersen. 2017. Gaze + Pinch Interaction in Virtual Reality. In *Proceedings of the 5th Symposium on Spatial User Interaction* (Brighton, United Kingdom) *(SUI '17)*. Association for Computing Machinery, New York, NY, USA, 99–108. https://doi.org/10.1145/3131277.3132180

[50] Sohrab Saeb, Cornelius Weber, and Jochen Triesch. 2011. Learning the Optimal Control of Coordinated Eye and Head Movements. *PLOS Computational Biology* 7, 11 (2011). https://doi.org/10.1371/journal.pcbi.1002253

[51] Ludwig Sidenmark and Hans Gellersen. 2019. Eye, Head and Torso Coordination During Gaze Shifts in Virtual Reality. *ACM Trans. Comput.-Hum. Interact.* 27, 1, Article 4 (dec 2019), 40 pages. https://doi.org/10.1145/3361218

[52] Ludwig Sidenmark and Hans Gellersen. 2019. Eye&head: Synergetic eye and head movement for gaze pointing and selection. In *Proceedings of the 32nd annual ACM symposium on user interface software and technology*. 1161–1174.

[53] J. David Smith and T. C. Nicholas Graham. 2006. Use of Eye Movements for Video Game Control. In *Proceedings of the 2006 ACM SIGCHI International Conference on Advances in Computer Entertainment Technology* (Hollywood, California, USA) *(ACE '06)*. Association for Computing Machinery, New York, NY, USA, 20–es. https://doi.org/10.1145/1178823.1178847

[54] R William Soukoreff and I Scott MacKenzie. 2004. Towards a standard for pointing device evaluation, perspectives on 27 years of Fitts' law research in HCI. *International journal of human-computer studies* 61, 6 (2004), 751–789.

[55] India Starker and Richard A. Bolt. 1990. A Gaze-Responsive Self-Disclosing Display. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Seattle, Washington, USA) *(CHI '90)*. Association for Computing Machinery, New York, NY, USA, 3–10. https://doi.org/10.1145/97243.97245

[56] Ayoung Suh and Jane Prophet. 2018. The State of Immersive Technology Research: A Literature Analysis. *Computers in Human Behavior* 86 (2018), 77–90. https://doi.org/10.1016/j.chb.2018.04.019

[57] Eleftherios Triantafyllidis and Zhibin Li. 2021. The Challenges in Modeling Human Performance in 3D Space with Fitts' Law *(CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, Article 56, 9 pages. https://doi.org/10.1145/3411763.3443442

[58] Huawei Tu, Jin Huang, Hai-Ning Liang, Richard Skarbez, Feng Tian, and Henry Been-Lirn Duh. 2021. Distractor Effects on Crossing-Based Interaction. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 192, 13 pages. https://doi.org/10.1145/3411764.3445340

[59] Stephen Uzor and Per Ola Kristensson. 2021. An Exploration of Freehand Crossing Selection in Head-Mounted Augmented Reality. *ACM Trans. Comput.-Hum. Interact.* 28, 5, Article 33 (aug 2021), 27 pages. https://doi.org/10.1145/3462546

[60] Oleg Špakov, Poika Isokoski, and Päivi Majaranta. 2014. Look and Lean: Accurate Head-Assisted Eye Pointing. In *Proceedings of the Symposium on Eye Tracking Research and Applications* (Safety Harbor, Florida) *(ETRA '14)*. Association for Computing Machinery, New York, NY, USA, 35–42. https://doi.org/10.1145/2578153.2578157

[61] C Wingrave and D Bowman. 2005. Baseline factors for raycasting selection. In *Proceedings of HCI International*. 61–68.

[62] Wenge Xu, Xuanru Meng, Kangyou Yu, Sayan Sarcar, and Hai-Ning Liang. 2022. Evaluation of Text Selection Techniques in Virtual Reality Head-Mounted Displays. In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 131–140. https://doi.org/10.1109/ISMAR55827.2022.00027

[63] Difeng Yu, Hai-Ning Liang, Xueshi Lu, Kaixuan Fan, and Barrett Ens. 2019. Modeling Endpoint Distribution of Pointing Selection Tasks in Virtual Reality Environments. *ACM Trans. Graph.* 38, 6, Article 218 (nov 2019), 13 pages. https://doi.org/10.1145/3355089.3356544

[64] Difeng Yu, Xueshi Lu, Rongkai Shi, Hai-Ning Liang, Tilman Dingler, Eduardo Velloso, and Jorge Goncalves. 2021. Gaze-Supported 3D Object Manipulation in Virtual Reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 734, 13 pages. https://doi.org/10.1145/3411764.3445343

[65] Shumin Zhai, Stéphane Conversy, Michel Beaudouin-Lafon, and Yves Guiard. 2003. Human On-Line Response to Target Expansion. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Ft. Lauderdale, Florida, USA) *(CHI '03)*. Association for Computing Machinery, New York, NY, USA, 177–184. https://doi.org/10.1145/642611.642644

[66] Xinyong Zhang. 2021. Evaluating the Effects of Saccade Types and Directions on Eye Pointing Tasks. In *The 34th Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '21)*. Association for Computing Machinery, New York, NY, USA, 1221–1234. https://doi.org/10.1145/3472749.3474818

[67] Ziyue Zhang, Jin Huang, and Feng Tian. 2020. Modeling the Uncertainty in Pointing of Moving Targets with Arbitrary Shapes. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3334480.3382875

[68] Suwen Zhu, Yoonsang Kim, Jingjie Zheng, Jennifer Yi Luo, Ryan Qin, Liuping Wang, Xiangmin Fan, Feng Tian, and Xiaojun Bi. 2020. Using Bayes' Theorem for Command Input: Principle, Models, and Applications. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3313831.3376771