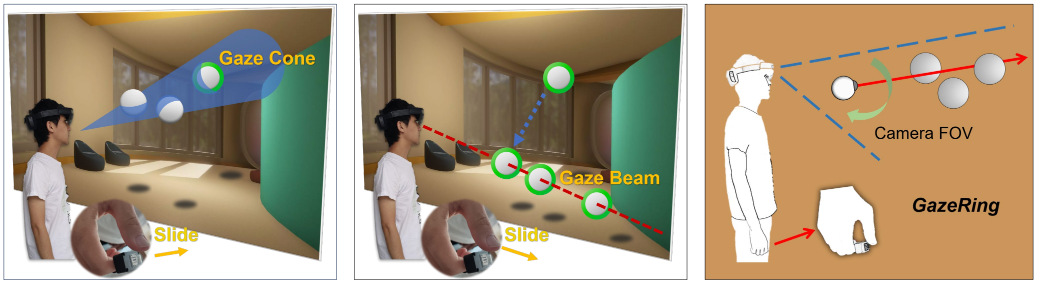
GazeRing: Enhancing Hand-Eye Coordination with Pressure Ring in Augmented Reality

Zhimin Wang[1] Jingyi Sun[1] Mingwei Hu[2] Maohang Rao[1] Weitao Song[2] Feng Lu[1]*

[1] State Key Laboratory of VR Technology and Systems, School of CSE, Beihang University, Beijing, China
[2] School of Optics and Photonics, Beijing Institute of Technology, Beijing, China



(a) Select within gaze cone (b) Translate along gaze beam (c) Allow hands outside camera view

Figure 1: We propose GazeRing, a novel multimodal interaction technique that combines eye-gaze tracking and ring-based touch. a) The user first selects occluded targets within the gaze cone by sliding on the ring. b) The user then translates the target by attaching it to and moving it along the gaze beam. c) Our GazeRing technique is subtle and private, allowing the user's hands complete freedom of movement.

Abstract

Hand-eye coordination techniques find widespread utility in augmented reality and virtual reality headsets, as they retain the speed and intuitiveness of eye gaze while leveraging the precision of hand gestures. However, in contrast to obvious interactive gestures, users prefer less noticeable interactions in public settings due to concerns about social acceptance. To address this, we propose ***GazeRing***, a multimodal interaction technique that combines eye gaze with a smart ring, enabling private and subtle hand-eye coordination while allowing users' hands complete freedom of movement. Specifically, we design a pressure-sensitive ring that supports sliding interactions in eight directions to facilitate efficient 3D object manipulation. Additionally, we introduce two control modes for the ring: finger-tap and finger-slide, to accommodate diverse usage scenarios. Through user studies involving object selection and translation tasks under two eye-tracking accuracy conditions, with two degrees of occlusion, GazeRing demonstrates significant advantages over existing techniques that do not require obvious hand gestures (*e.g.*, gaze-only and gaze-speech interactions). Our GazeRing technique achieves private and subtle interactions, potentially improving the user experience in public settings. A demo video can be found at zhimin-wang.github.io/GazeRing.html.

Index Terms: Augmented reality, object manipulation, hand-eye coordination, pressure ring

1 Introduction

Virtual Reality (VR), Augmented Reality (AR), and Mixed Reality (MR), collectively known as Extended Reality (XR) technologies, seamlessly bridge the virtual and physical realms, offering users an immersive experience. The increasing adoption of XR headsets, such as the Apple Vision Pro and Microsoft HoloLens, has spurred extensive exploration of their applications across various domains, including online education and remote healthcare [31, 6, 16].

Developing natural and effective interaction methods for XR headsets has been a critical research focus. Commercial VR devices, such as the HTC Vive and Meta Quest, primarily utilize hand-held controllers for interaction, tracked by capturing the array of infrared LEDs on them (*e.g.*, 24 infrared LEDs). However, this interaction poses challenges in outdoor environments due to interference from ambient infrared light [3, 12]. Consequently, more research has explored leveraging direct human inputs as interaction channels, such as hand gestures and eye gaze. For AR devices like the HoloLens, hand gesture interaction is favored for its precision (*e.g.*, an average positional error of 5.4 mm for finger-tracking [57]). Nevertheless, it suffers from visual field occlusion caused by hands [14] and arm fatigue [8, 20]. In contrast, eye gaze-based interaction offers significant advantages in terms of speed, with eyeball rotations reaching up to $700^\circ/\mathrm{s}$ and accelerations of $24,000^\circ/\mathrm{s}^2$ [60], and intuitiveness, as gaze naturally points to the desired target. However, this interaction faces the Midas Touch problem [33, 48] and insufficient accuracy [4]. Therefore, it underscores the need for enhancement of single-modal interactions.

Combining different modalities can leverage their complementary strengths. One intuitive idea is to combine hand and eye inputs for hand-eye coordination. Since eye gaze, which is fast and intuitive, complements the precision of hand gestures, many studies follow the principle of "gaze selects, hand manipulates" [24, 59, 10]. Hand-eye coordination is also crucial for the Apple Vision Pro, improving its spatial computing capabilities and accessibility [1]. This dual-modal interaction has been proven effective in numerous scenarios, *e.g.*, menu selection [36, 30], text input [19, 29, 61], and page browsing [1, 35], significantly reducing the learning curve and aligning the user experience more closely with natural human behaviors. Additionally, Bao *et al.* have explored how hand-eye coordination can aid in selecting and translating objects within environments with heavy 3D occlusions, further illustrating the practical applications and advantages of this interaction technique [4].

However, the aforementioned hand-eye coordination may raise concerns regarding public social acceptance, as it is only effective

*Corresponding Author: Feng Lu (lufeng@buaa.edu.cn) when the hands are within the view of XR headset's cameras (*e.g.*, a limited FoV for the Articulated Hand Tracking (AHAT) mode of the HoloLens 2's depth camera [46]). This means the XR headset may lose track of the hands if they move outside the camera's view, *e.g.*, being too low or behind the user's back. Tung *et al.* found that users preferred subtle interactions over obvious hand gestures in public [45]. Consequently, enhancing the privacy and convenience of hand-eye coordination techniques could significantly advance their practicality for public use.

To address these issues, we explore the use of smart rings. Interaction between finger and smart ring is highly favored for its private and subtle nature [45]. Recently, ring-based interaction has seen extensive application in XR, such as fingertip micro-gestures recognized by cameras or electric field sensing [26, 11], touch contact sensing on physical surfaces with Inertial Measurement Units (IMUs) [17, 27, 41], and page scrolling enabled by flexible pressure sensors (FPS) [62, 28, 5, 47]. Rings utilizing FPS allow for sliding interactions in either two directions (left and right) or four cardinal directions (up, down, left, and right). These ring-based interactions potentially offer a compelling alternative to the traditional "hand" component of hand-eye coordination.

In this paper, we propose ***GazeRing***, a multimodal interaction technique combining eye gaze with a smart ring to enhance hand-eye coordination with private and subtle actions while allowing users' hands freedom of movement. We design a pressure-sensitive ring with a FPS, utilizing a thumb-to-index-finger trigger mechanism (Fig. 1) that supports sliding interactions in eight directions for rapid 3D object manipulation. We introduce two control modes: fingertip-tap and fingertip-slide, to support different usage scenarios. In a user study, we evaluate GazeRing under two tasks related to object selection and translation, with varying occlusion and eye-tracking accuracy conditions (1.5° and 4° error). We compare the GazeRing method with techniques that also do not require obvious hand gestures (*e.g.*, gaze-only interaction and gaze-speech interaction). The GazeRing demonstrates significant advantages in terms of speed and usability in scenarios involving occlusion or inaccurate eye tracking, and is preferred by participants. Our work enhances the privacy and convenience of hand-eye coordination interactions, potentially improving user experience in public.

In summary, the contributions of this paper are three-fold:

1. We propose GazeRing, a novel multimodal interaction technique that combines eye gaze with a smart ring, enabling private and subtle hand-eye coordination while allowing users' hands complete freedom of movement.
2. We design a pressure-sensitive ring, supporting sliding interactions in eight directions for rapid 3D object manipulation. Additionally, we also introduce two control modes (fingertip-tap and fingertip-slide) for the ring to accommodate diverse usage scenarios.
3. We evaluate the performance of GazeRing in object selection and translation tasks under two eye-tracking conditions, with two degrees of occlusion, demonstrating its advantages in scenarios involving occlusion or inaccurate eye tracking.

2 Related work

In this section, we review single-modal interaction in AR, multimodal interaction in AR, as well as a discussion of ring-based interaction.

2.1 Single-Modal Interaction in AR

Single-modal interaction has been widely explored, and we primarily discuss interactions closely related to this paper, including eye gaze, freehand gesture, and handheld controller-based interaction.

Eye gaze. This interaction requires less physical effort and provides a more natural experience [51, 54, 36, 49, 55]. Kytö *et al.* used gaze-only interaction to select targets in AR [24]. However, this interaction faces mainly three issues: 1) The Midas touch problem, which makes it difficult to distinguish between gaze selection and viewing [33]. 2) The inability to move objects in depth [59]. 3) The impact of insufficient gaze accuracy on interaction [4].

Freehand gesture. Gesture interaction is widely used in current AR systems [40, 50, 37], due to its precision and resemblance to daily behaviour. Users can interact through various gestures, such as pinch, tap, drag, and dual-hand interaction [8, 38]. However, gestures must be captured within the AR device's camera view, which can lead to muscle fatigue during prolonged use [8, 20].

Handheld controller. Although many VR devices use handheld controllers, AR devices also have this option [56, 15]. However, this interaction poses challenges outdoors due to ambient infrared light interference [3, 12]. Moreover, handheld controllers are generally large and inconvenient to carry in public.

In summary, single-modal interactions have their own advantages and limitations. As a result, it is vital to discover an optimized method that merges different interaction modalities to leverage their complementary properties.

2.2 Multimodal Interaction in AR

We discuss two multimodal interactions relevant to our work: hand-eye coordination and gaze-speech interaction.

Hand-eye coordination. In recent years, numerous studies have focused on hand-eye coordination [4, 53, 59, 24], as it leverages the speed and intuitiveness of eye gaze for target selection and the precision of hand gestures for manipulation. Apple's Vision Pro has adopted hand-eye coordination as its fundamental interaction method [1]. However, hand-eye coordination requires hands to be visible to the headset's cameras, making gestures obvious and noticeable, which raises concerns about public social acceptance [45].

Gaze-speech interaction. We consider this interaction to be private and subtle, as speech can be made silent and achieved through lip movements [7]. Gaze-speech interaction has been extensively explored in performing basic operations on computer interfaces. Elepfandt *et al.* combined gaze tracking with voice commands to manipulate objects in pictures displayed on a back projection canvas [13]. Similarly, Kaur *et al.* investigated the fusion of gaze tracking and speech for moving objects on a computer screen [21]. As speech interaction involves numerous voice commands, users may find it challenging to remember all of them.

In this paper, we design a multimodal interaction combining eye-gaze and ring-based touch, enhancing hand-eye coordination with subtle actions and allowing hands complete freedom of movement.

2.3 Ring-based Interaction

Recently, smart rings have become popular due to their subtle nature. Ring-based interactions can offer an alternative to the traditional "hand" component of hand-eye coordination. This paper primarily discuss three types of ring-based interactions as follows.

Fingertip micro-gestures recognized by cameras. Recent advancements in computer vision have enabled the recognition of fingertip micro-gestures using cameras [44, 23]. NailRing utilizes a camera mounted on a ring to capture and recognize subtle fingertip gestures [26], while EFRing employs electric field sensing to detect and classify micro-gestures performed on the fingertip [11].

IMUs-based Ring. IMUs have been integrated into ring-based devices to enable touch contact sensing on surfaces. Gu *et al.* utilized a ring-mounted IMU to detect tapping and sliding gestures on various surfaces [17]. DualRing incorporates two IMU-equipped rings, expanding the interaction modalities [27]. Shi *et al.* employed an IMU-based ring to provide a set of input gestures [41].

FPS-based Ring. FPS have been employed in ring-based devices to enable sliding interactions. Octa-Ring extends this concept by incorporating different levels of pressure around the ring, enabling five touch configurations [28]. ARO [5] and GestuRING [47] further demonstrate the potential of FPS-based rings for navigation tasks and web-based tool.

We present a pressure-sensitive ring equipped with a FPS that employs a thumb-to-index-finger trigger mechanism, enabling sliding interactions in eight directions for 3D object manipulation.

3 System Design

We propose GazeRing, a novel multimodal interaction technique that combines eye-gaze tracking and ring-based touch. This private and subtle interaction mechanism addresses concerns regarding public social acceptance. The remainder of this section is structured as follows. First, we introduce the hardware implementation of the pressure-sensitive ring and the system's hardware diagram in Section 3.1. Next, we describe two control modes for the ring, namely fingertip-slide and fingertip-tap, in Section 3.2. Finally, in Section 3.3, we present GazeRing, a set of strategies for the combined operation of eye-gaze tracking and ring-based touch.

3.1 Hardware Implementation: Pressure-sensitive Ring

As the design of thumb-to-index-finger interaction heavily relies on the pressure ring, it is essential to first introduce the hardware implementation of the ring and pressure sensing.

Design of Pressure Ring. Ideally, for optimal thumb-to-index-finger interaction, the sensor should be directly attached to the finger surface. This approach ensures a more private interaction, as no obvious hardware is visible. However, although previous research has explored skin-inspired flexible sensors for interactions [58, 22], these hardware solutions are not yet widely adopted. Currently available commercial FPS utilizes pressure-sensitive resistors. However, when these sensors are attached along the finger surface, the curved plane of the finger can cause sensor deformation, significantly interfering with pressure detection. To address this issue, we designed a finger-mounted ring for the FPS, as shown in Fig. 2 (c)(d). The ring consists of a rigid wearable band and a hollow box. The band is worn to the index finger, while the box, measuring $31 \times 19 \times 13$ mm^3, houses the Printed Circuit Board (PCB), battery and flexible printed circuit (FPC). The flat surface of the box allows the FPS to be attached smoothly without causing sensor deformation. Users interact by placing their thumb on the sensor attached to the box's surface.

Selection of FPS. We compare FPS of various sizes and selected one similar in size to a one-penny coin, as shown in Fig. 2 (a). The chosen FPS measures $14 \times 14 \times 0.1$ mm^3 and is based on distributed sensing technology. It features 16 distributed sensor units on its surface, with a spacing of 0.8 mm between each unit. Each sensor unit is a resistive pressure sensor that exhibits a decrease in resistance as the applied pressure increases. The sensor units convert the change in resistance into a voltage signal, allowing the determination of the pressure applied above each unit. By combining data from all 16 sensor units, the pressed area can be identified.

Design of PCB. The PCB provided by the FPS manufacturer measures $40 \times 26 \times 10$ mm^3, which is too bulky for our small-ring requirements. Therefore, we designed a more compact PCB with dimensions of $19 \times 14 \times 5$ mm^3, as shown in Fig. 2 (b). This custom-designed board is responsible for receiving the raw voltage signals, calculating the pressed amplitude for each sensor unit, and transmitting the processed data to a computer or HoloLens for subsequent GazeRing interaction. The board incorporates an ESP32 chip as the main control chip and is powered by a battery with an output of 3.7 V × 0.1 A. The board communicates with the PC via Bluetooth. The chip, Bluetooth module, battery, and FPS are all common products purchased online. The ESP32 chip is the Tensilica dual-mode dual-core processor, the Bluetooth module is the NRF52810, the battery is a 3.7 V soft-pack lithium battery, and the FPS is the RX-M0404S.

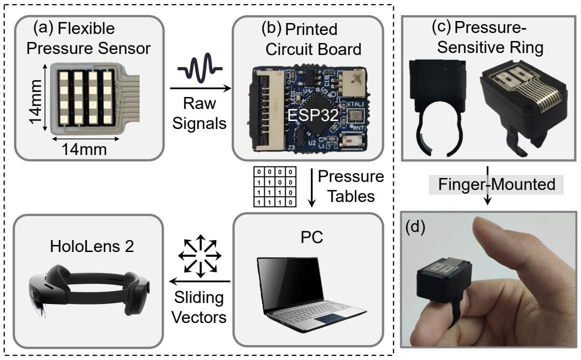


Figure 2: Hardware diagram (left): User presses FPS, generating raw voltage signal sent to PCB. PCB calculates pressed amplitude for each sensor unit and transmits data to PC. The PC calculates two control modes and sends sliding vectors to headset application. Custom Pressure Ring (right): FPS attached to ring's surface, enabling thumb-to-index-finger interaction.

Pressed amplitude calculation in PCB. The PCB needs to process the received raw voltage signal before sending it to the PC. The processing mainly includes three steps. 1) Baseline calibration. As shown in Fig. 2 (c), the bending of the FPC still interferes with the force applied to the FPS to a certain extent. We use the pressure values of each sensor unit when the user is wearing the ring without pressing as the baseline for subsequent voltage calculations. 2) Average processing. The PCB communicates with the PC every 50 ms, while the data accumulated by the sensor units are continuous in time. Let t denote the current time. Therefore, each sensor unit i $(i = 1, \cdots, 16)$ calculates the average of all the voltage values collected in the time interval (t - 50 ms, t), which is denoted as $V_i^{avg}(t)$. 3) Voltage normalization. Each sensor unit i records its maximum voltage value V_i^{max} that appears during the entire usage process. The voltage $V_i(t)$ sent to the PC at time t is actually a percentage of the maximum voltage value, *i.e.*, $V_i(t) = V_i^{avg}(t)/V_i^{max}$. Finally, we obtain a pressure table containing 16 normalized voltage values.

3.2 Two Control Modes of Pressure Ring

Based on the carefully designed pressure-sensitive ring, we can successfully obtain the pressure table reflecting the pressure amplitude of each sensor unit. However, another equally important problem is how to interact with virtual objects using these pressure tables from the ring. In this section, we design two control modes for the ring: the fingertip-slide mode and the fingertip-tap mode. The fingertip-slide mode better aligns with the characteristics of flexible sensing, perceiving the sliding of the fingertip, while the fingertip-tap mode more closely resembles the control method of traditional controllers, *i.e.*, clicking. Therefore, we can select a more appropriate mode depending on the specific AR application scenario.

3.2.1 Fingertip-Slide Mode

Existing commercial rings support sliding operations, such as scrolling [5, 47], in either two directions (left and right) or four cardinal directions (up, down, left, and right). However, their performance is less satisfactory in object selection and manipulation in 3D space, as objects may need to be moved in various directions, including diagonally. To address this limitation, we propose the fingertip-slide mode for rapid 3D object manipulation, which supports sliding operations in eight directions by adding four diagonal directions. Fingertip-slide mode enables users to interact primarily by sliding their fingertips across different areas of the flexible pressure sensor. Additionally, users can also perform a long-press on the sensor to confirm and select. This mode includes two steps, as shown in Fig. 3. We will provide a detailed introduction to the implementation of this mode in following sections.

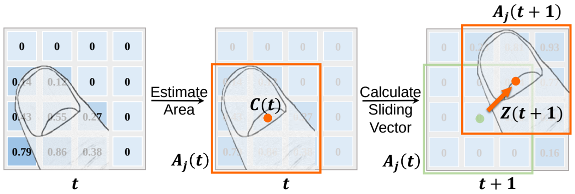


Figure 3: The fingertip-slide mode need to calculate the direction of sliding pointing from the previous pressing position to the current pressing position.

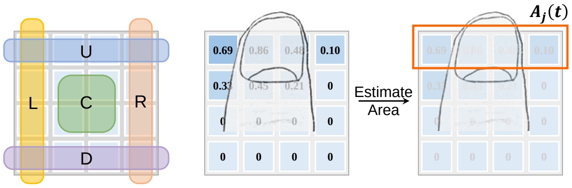


Figure 4: For fingertip-tap mode, the flexible pressure sensor are partitioned into five areas: up (U), down (D), left (L), right (R), and center (C), with each area corresponding to a specific operation.

$$\begin{aligned} \text{Idx} &= \{lu, ru, ld, rd\} \\ \text{ValidAreas} &= \{A_i | \sum_{k \in A_i} \mathbb{I}(V_k(t) > 0) > 3, i \in \text{Idx}\} \\ \text{VoltageSum} &= \{S_i(t) | S_i(t) = \sum_{k \in A_i} V_k(t), A_i \in \text{ValidAreas}\} \end{aligned} \tag{1}$$

Estimation of pressing areas and pressing centers. The sixteen sensor units on flexible pressure sensor are partitioned based on their location into four 3×3 grid areas, left-up (A_{lu}), right-up (A_{ru}), left-down (A_{ld}), and right-down (A_{rd}). The sensor unit located at the center of each area is designated as the center point of the corresponding area. With the pressure table containing voltages $V(t)$, the validity of each area A_i is determined and its voltage sum $S_i(t)$ is calculated, as shown in Eq. (1). Here, $\mathbb{I}(\cdot)$ represents the indicator function. Then, we identify the valid area $A_j(t)$ with the maximum voltage sum and recognize its center $C(t)$ as the estimated center of pressing area, which can be written as Eq. (2).

$$C(t) = \text{Center}(A_j(t)), \text{ s.t. } j = \arg\max_{i \in \text{Idx}} S_i(t) \tag{2}$$

Calculation of sliding vectors. The temporal change of pressing center indicates the sliding direction of the fingertip. Hence, we define the sliding vector $Z(t+1)$ as the vector originating from previous pressing center $C(t)$ to the current pressing center $C(t+1)$, as described in Eq. (3). If $Z(t+1)$ remains $\vec{0}$ for 1.5 seconds, the user is considered long-pressing to confirm and select. Otherwise, if the sliding vector remains constant over a period, the object's sliding speed will increase, thus accelerating the operation. By calculating the pressing center and sliding vector in this manner, we have implemented a kind of sliding operation interaction with a resolution of 2×2, allowing for movement in eight distinct directions.

$$Z(t+1) = \overrightarrow{C(t)C(t+1)} \tag{3}$$

It is worth mentioning that we had attempted to implement an algorithm to provide sliding vectors with higher resolution, with the method of searching extreme points and gradient descent. However, experiments have revealed that sliding vectors with higher resolution would result in more mis-operations, possibly due to the tiny size of flexible pressure sensor. Therefore, for the sake of stability and accuracy of the interaction, we ultimately designed a more straightforward sliding operation mode, as discussed above.

3.2.2 Fingertip-Tap Mode

We designed the fingertip-tap mode for two reasons. First, in certain AR application scenarios, such as menu selection, a button-like interaction method may be more convenient. Moreover, an interaction mode similar to traditional controllers may be more popular in some users who have difficulty in getting accustomed to fingertip-slide mode. Fingertip-tap mode enables users to execute corresponding operations by tapping on different areas of the FPS. Fig, 4 shows the detailed implementation of this mode.

Partition of tapping areas. Resembling traditional controllers, we designed five distinct operations for the fingertip-tap mode, including a confirmation operation and four directional operations. Therefore, the sixteen sensor units on the flexible pressure sensor are divided into five areas: left (A_{left}), right (A_{right}), up (A_{up}), down (A_{down}), and center (A_{center}), with each area corresponding to a specific operation. Tapping on the center area for a period of 1.5 seconds represents the confirmation operation.

Estimation of tapping areas. As the partitioned area decreases in size, areas with more than two valid sensor units are considered valid. Then, similar to the fingertip-slide mode, the area with the maximum voltage sum $S_j(t)$ among the valid areas is selected as the tapped area $A_j(t)$. Eq. (4) below describes the calculation process in detail.

$$\begin{aligned} \text{Idx} &= \{left, right, up, down, center\} \\ \text{ValidAreas} &= \{A_i | \sum_{k \in A_i} \mathbb{I}(V_k(t) > 0) > 2, i \in \text{Idx}\} \\ \text{VoltageSum} &= \{S_i(t) | S_i(t) = \sum_{k \in A_i} V_k(t), A_i \in \text{ValidAreas}\} \\ \text{TapArea} &= A_j(t), \text{ s.t. } j = \arg\max_{i \in \text{Idx}} S_i(t) \end{aligned} \tag{4}$$

3.3 Interaction Design for GazeRing

As selection and translation are the most prevalent operations with 3D objects [4], our interaction design primarily focuses on the two operations: first selecting a 3D object and then translating it to a specific location. By integrating gaze and the pressure ring, we develop a set of strategies for object selection and translation that combine operations of eye-gaze tracking and ring-based touch, *i.e.*, GazeRing interaction.

GazeRing interaction comprises GazeRing-Slide interaction (*GR-S*) and GazeRing-Tap interaction (*GR-T*), depending on the specific control mode of the pressure ring. We implement these two interaction techniques on Microsoft HoloLens2, which provides the necessary gaze data through its built-in gaze estimation module. It is important to note that these techniques can be applied to any Head-Mounted Display (HMD) device equipped with gaze estimation capabilities.

GazeRing-Slide interaction (*GR-S*). Our design of eye-gaze and ring-touch combined interaction mechanism of *GR-S* consists of two phases, each with four steps, as shown in Fig. 5. **Selection Phase:** (1) The user gazes at the target object and long-presses to establish a gaze cone with a 6° angle surrounding the gaze beam, as depicted in Fig. 5(a). This gaze cone, representing the viewing area, is designed to accommodate potential insufficiencies in eye tracking. (2) If the gaze cone is still positioned improperly, it can be refined by sliding on the pressure ring in eight different directions, as depicted in Fig. 5(b). (3) Once the gaze cone is refined, the objects within the cone become selectable options. The user can slide up and down to traverse occluding objects, which become transparent when not selected, until successfully picking the target object, as depicted in Fig. 5(c). (4) Subsequently, the user long-presses to confirm the selection of the target object. **Translation Phase:** (1) The user gazes at the destination and long-presses to attach the selected object onto the gaze beam, as presented in Fig. 5(d), providing rapid object translation. (2) Similar to the selection

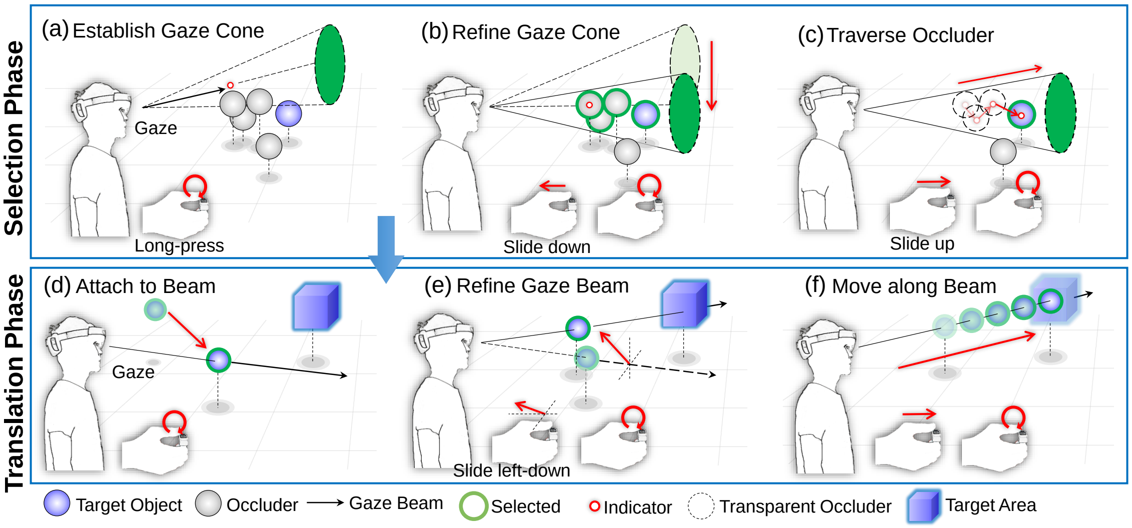


Figure 5: This paper presents GazeRing-Slide interaction, which consists of two phases. In the selection phase, users trigger a gaze cone and slide on the pressure ring to discretely select objects in the depth direction. In the translation phase, users refine the gaze beam direction using the pressure ring and controls objects by sliding. The GazeRing-Tap interaction follows a similar process.

phase, the gaze beam can be refined in the same manner as the gaze cone, as presented in Fig. 5(e). (3) The user then controls the object to move along the gaze beam by sliding up and down, allowing for depth adjustment to reach the destination, as presented in Fig. 5(f). (4) Eventually, when the user long-presses to release the target, the interaction process is accomplished.

GazeRing-Tap interaction (*GR-T*). *GR-T* shares the same interaction process as *GR-S*, differing only in the operation of the pressure ring. **In Selection Phase**, the user refines the gaze cone and traverse the occluders by tapping on the four directional areas. Similarly, **in Translation Phase**, the user taps on the center area to attach the object to the gaze beam and taps on the other areas to refine and move the object.

4 Comparisons of Interaction Modalities For Object Manipulation

The primary goal of this section is to evaluate whether GazeRing techniques can efficiently complete common AR tasks, such as object selection and translation, without requiring obvious hand gestures. The impact of different levels of occlusion on interaction efficiency is assessed as well. Furthermore, considering that eye-tracking accuracy often decreases during use [39], we also evaluate the influence of two eye-tracking accuracy conditions on GazeRing techniques. We compare GazeRing with interactions with a similarly private and subtle nature, *i.e.*, gaze-only interaction (*Gaze*) and gaze-speech interaction (*GS*). We propose two hypotheses:

H_1: *These GazeRing techniques have higher efficiency and usability than using Gaze and GS in object manipulation.*

H_2: *These GazeRing techniques are more attractive than using Gaze and GS in object manipulation.*

4.1 Participants and Apparatus

We recruited 16 participants (14 males, 2 females) from the laboratory and university campus, aged 21 to 27 years (M = 22.3, SD = 1.6). According to the pre-study questionnaire with a 5-point Likert scale, the participants have low prior familiarity with the pressure-sensitive ring (Mean = 1.9) and eye-tracker (Mean = 2.6), medium familiarity with AR (Mean = 3.1) and speech-based inputs (Mean = 3.3). All participants have normal or correct-to-normal vision and can read and speak English fluently. We conducted our experiments using Microsoft HoloLens 2, which is equipped with eye tracking and speech keyword spotting capabilities.

4.2 Task Design

Object manipulation, *e.g.*, selection and translation, is one of the most common tasks in AR [59, 52]. Previous studies have explored hand-eye coordination in object selection with heavy occlusions [4]. To evaluate the efficiency and usability of GazeRing, we design two tasks: No Occlusion (NO) Task and Heavy Occlusion (HO) Task, similar to Bao *et al.*'s work [4]. In both tasks, the goal is to find spheres of different colors in the scene and place them into the corresponding colored target areas, as shown in Fig. 7. Each sphere has a radius of 0.2 m, and each target area is a cube-shaped space with an edge length of 0.5 m. The task parameters are set with reference to the work of Bao *et al.* [4]. When a sphere is placed into the correct target area, the target area highlights to notify the participant. The participant then releases the sphere, which disappears to indicate successful placement.

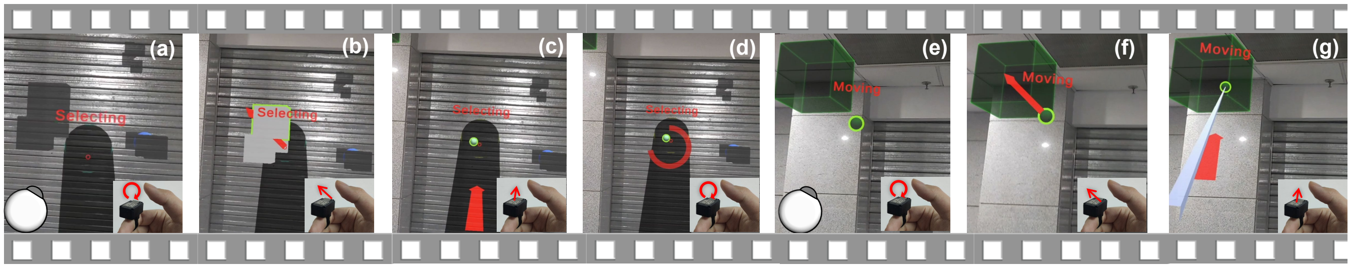
No Occlusion Task. In this task, there are four spheres in the scene, each with a different color. They are evenly distributed along a horizontal line 2.7m away from the participant, with a length of 1.2 m. The scene also contains four target areas with different colors, placed at various depths of 2.5 m, 4 m, 5.5 m, and 7 m away from the participant, at a height of 1.5 m.

Heavy Occlusion Task. In this task, four spheres are distributed within a space of 1.5m × 1.2 m, at distances ranging from 2 m to 5 m from the user. Each sphere is completely occluded by 3 to 4 interfering cubes. The four target areas are placed in the corners of a 3 m × 3 m rectangular area, 3 m away from the participant. Participants need to overcome the influence of the interfering objects using different interaction methods to select the task spheres and place them into the corresponding target areas.

4.3 Design of Two Eye-tracking Conditions

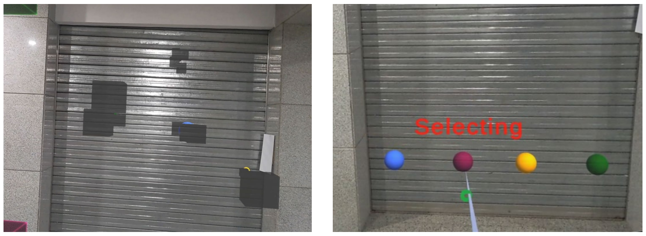
In AR, the accuracy of eye tracking is crucial for effective gaze-based interaction. Due to device slippage, achieving precise gaze-based interaction requires frequent recalibration by users, which can be time-consuming and energy-intensive [39]. Since all the interaction techniques used in the experiment are gaze-related, eye-tracking accuracy plays a vital role. Therefore, we design two eye-tracking conditions to investigate the influence of eye-tracking accuracy on different interactions.

Accurate Eye Tracking (*Acc-Eye*). The eye tracking data used in our study is directly obtained from the HoloLens. After users correctly perform the gaze calibration procedure, the eye tracking



Establish Gaze Cone Refine Gaze Cone Traverse Occluder Confirm Target Attach to Gaze Beam Refine Gaze Beam Move Along Beam

Figure 6: The strategies for the combined operation of eye-gaze tracking and ring-based touch are divided into two phases. The selection phase includes steps (a) through (d), while the translation phase encompasses steps (e) through (g).



(a) Heavy Occlusion Task (b) No Occlusion Task

Figure 7: Demonstration of two task scenarios. In the Heavy Occlusion (HO) Task, all target objects (4 spheres) are completely occluded.

system provides relatively accurate gaze data, with an eye tracking error of approximately 1.5° [32].

Insufficient Eye Tracking (*Ins-Eye*). We considered the impact of slight device slippage, such as that caused by eyebrow movement, on eye tracking accuracy. In this case, eye tracking accuracy slightly decreases, potentially reaching an error of 4° [2]. To simulate this, we added a random deviation of 2 to 3° to the gaze vector output by the HoloLens, with the direction being either 45° upward or downward to the right. During the participant's task execution, when the participant's gaze point undergoes a significant change (*i.e.*, a gaze jump), the original deviation is replaced by another random deviation; otherwise, this random deviation remains fixed.

4.4 Interaction Modalities

We compared four private and subtle interaction techniques: two GazeRing techniques proposed in this paper (*GR-S* and *GR-T*), which are introduced in Section 3.3, as well as two techniques introduced below, gaze-only interaction and gaze-speech interaction.

Gaze-only Interaction (*Gaze*). Gaze-based selection of occluded objects has been extensively explored [25, 42]. However, these methods are suitable only for partial occlusion [43] or require high accuracy in gaze depth estimation [42]. For the nearly complete occlusion and dense scenes presented, an effective solution has not yet been found. We adopt the gaze-only interaction based on dwell time. **In Selection Phase**, the user gazes at the target object and fixates on it for 1.5 seconds to select it. The dwell time was established based on two key factors: the one-second duration suggested by Wang *et al.* [52] and the accuracy reduction resulting from the *Ins-Eye* condition. If the gaze intersects with multiple objects, the nearest object to the user will be selected. **In Translation Phase**, the user gazes at the target area, and the object is quickly translated laterally. A menu bar appears around the object, containing four buttons representing “forward, backward, pause, and release”. Gazing at the forward/backward buttons makes the object move along the corresponding depth direction until the user gazes at the pause button. Once the object reaches its destination, the user can gaze at the release button to complete the translation.

Gaze-Speech Interaction (*GS*). Gaze-speech interaction has been explored in object selection and translation scenarios [52, 34], including scenarios with occluding object [9]. We define gaze-speech interaction as private, which may arouse different opinions because users need to speak in public. However, recent research has shown that users can achieve speech keyword detection with only lip movement without vocalization [7], making this method simple to implement. Thus, we consider gaze-speech interaction as a type of private interaction. This paper adopts the vocalization scheme, which can be replaced by a silent scheme. **In Selection Phase**, the user gazes at the target object and invokes the gaze cone through the ”confirm” voice command. The user can say ”up, down, left, right” to refine the gaze cone and ”forward/backward” to select objects in depth. **In Translation Phase**, the user gazes at the target area, says ”confirm”, and controls the object's movement using six directional commands, finishing by confirmation or release command.

4.5 Experiment Procedure

Participants complete a pre-study questionnaire assessing their familiarity with AR systems and interaction methods. They watch an introductory video explaining the interactions and experiment. Participants then wear the HoloLens 2, complete gaze calibration, and undergo training to become acquainted with all techniques. Users begin the experiments for the four interactions in sequence, with the order of techniques counterbalanced using Latin Square design. Each interaction involves 4 sessions (= 2 tasks × 2 conditions), each limited to 3 minutes with 30-second breaks. After each interaction, users complete a questionnaire evaluating that interaction and rest for 5 minutes before the next. Following all 4 interaction experiments, participants rank the interaction techniques in a post-study questionnaire. Each interaction takes about 30 minutes, with the entire experiment lasting around 130 minutes. Each subject performs 16 (= 4 techniques × 4 sessions) experimental sessions in total.

4.6 Evaluation Metrics

Objective Measures. We define five objective metrics to evaluate the performance of participants across different interaction techniques in each session.

- **Average Finish Time:** the total finish time divided by the number of completed targets. This time includes both the selection and translation phases.
- **Finish Rate:** the proportion of completed targets to the total target count. This represents the overall efficiency of different interaction techniques.
- **Invalid Selection Count:** the total count of inference cube selections. A higher number of selections conducted on inference cubes indicates more redundant operations in 3D space with occlusion.
- **Average Selection Time:** the total selection time divided by the number of completed targets. This is the time participants spend observing and selecting within each target.

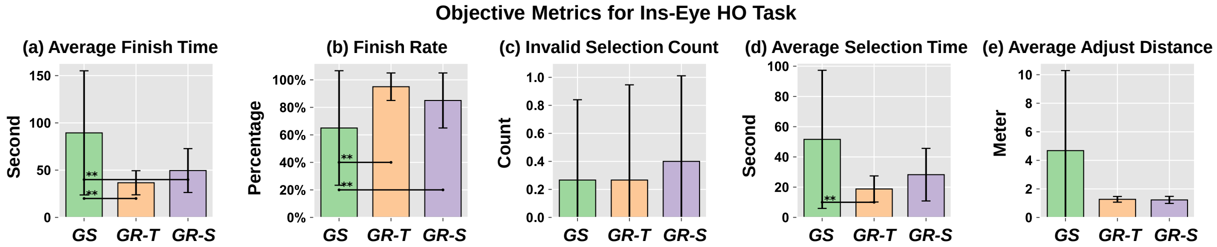


Figure 8: Bar charts of performance of multimodal interactions under different objective metrics for the Ins-Eye HO Task. Error bars indicate the standard error. The statistical significances are labeled with ** ($p < 0.05$).

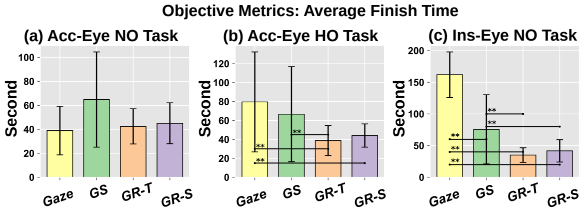


Figure 9: Bar charts of the interaction techniques' performance under Average Finish Time. Error bars indicate the standard error. The statistical significances are labeled with **($p < 0.05$)

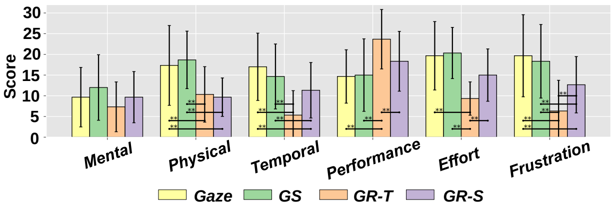


Figure 10: Bar charts of scores on the NASA-TLX questionnaire for the four interaction techniques, with error bars indicating the standard error. Statistical significances are denoted by ** ($p < 0.05$).

- **Average Adjust Distance:** the total adjustment distance in translation phase divided by the number of completed targets.

Subjective Measures. We also evaluate the techniques by some subjective measures of workload, frustration, occlusion impacts, and user preferences. Subjective measures are collected after participants accomplish all sessions with each interaction technique.

- **NASA-TLX [18]:** A 7-point Likert scale used to measure participants' mental demand, physical demand, temporal demand, effort, performance, and frustration level.
- **Occlusion Impact:** A 10-point Likert scale that measures the extent to which occlusion impacts interaction techniques. Higher scores indicate that the technique is more significantly affected by object occlusion.
- **Ring Observing Frequency:** A 10-point Likert scale that measures how frequently participants observe the ring. This scale is specifically designed for *GR-S* and *GR-T*. A lower score suggests that participants interact more intuitively.
- **User Preference:** A measure of participants' preferences across all interaction techniques. It is collected after completing all experiments. The ranks, from first to fourth place, are assigned scores of 10, 8, 6, and 4 points, respectively.
- **Open Question:** Open-ended questions that gather participants' general evaluations, perceptions of intuitiveness, suggestions for improvement, and assessments of interaction under different eye tracking conditions.

4.7 Results

Results of Objective Measures. We conducted repeated-measures ANOVAs ($\alpha = 0.05$) and post hoc pairwise t-tests to judge whether a certain metric is significantly different across interaction techniques. The tests for ANOVA assumptions are presented in the Supplementary Material.

Evaluation on Gaze-only Interaction. We analyzed the Average Finish Time of Gaze-Only interaction compared to other interactions across three sessions, as shown in Fig. 9. Repeated-measures ANOVAs showed significant differences in Average Finish Time among the four techniques for Acc-Eye HO and Ins-Eye NO Tasks ($p < 0.05$), but not for Acc-Eye NO Task ($p = 0.217$). **In Acc-Eye HO Task**, the performance of *Gaze* was significantly worse than that of *GR-S* and *GR-T* ($p = 0.017$, 0.003). Furthermore, **in Ins-Eye NO Task**, *Gaze* required the longest time to complete ($p < 0.001$ for all comparisons between *Gaze* and other techniques). Our findings indicate that even without occlusion, insufficient eye tracking significantly reduced the efficiency of *Gaze*, with its finish time being nearly three times that of *GR-S* and *GR-T*. These results suggest that *Gaze* lacks robustness and is significantly less efficient than *GR-S* and *GR-T*. Hence, we will not further analyze *Gaze* in the Ins-Eye HO Task.

Comparison among *GR-S*, *GR-T* and *GS*. We evaluated these three techniques across four sessions. **In Acc-Eye NO Task**, there were no significant differences among these techniques in terms of average finish time. **In Acc-Eye HO Task**, *GR-T* was significantly faster than *GS* ($p = 0.039$) and there was no significant differences between others. **In Ins-Eye NO Task**, both GazeRing interaction methods (*GR-T* and *GR-S*) were significantly faster than *GS* ($p = 0.016$, 0.027). Additional results are presented in the Supplementary Material. Next, we mainly analyzed these three interaction techniques in the most complex task, *i.e.*, **the Ins-Eye HO Task**, as shown in Fig. 8. Among the five objective metrics, the Invalid Selection Count and Average Adjust Distance did not exhibit significant differences among the techniques ($p = 0.808, 0.09$). Notably, *GR-S* and *GR-T* interactions demonstrated significantly shorter Average Finish Times than *GS* ($p = 0.017$, 0.008), with no significant difference between *GR-S* and *GR-T*. Consequently, this resulted in higher Finish Rate for both *GR-S* and *GR-T*. Furthermore, only *GR-T* exhibited a significantly faster selection speed than *GS* in terms of Average Selection Time ($p = 0.013$), indicating that *GR-T* can still select targets quickly under heavy occlusion. Overall, *GR-S* and *GR-T* interactions demonstrated good efficiency and accuracy even in the presence of occlusion or inaccurate eye tracking, with no significant difference between the two techniques.

Results of Subjective Measures. Repeated-measures ANOVAs on the NASA-TLX questionnaire revealed no significant differences among the four interaction techniques in terms of *mental demand*. However, significant differences were observed in the other five task loads, as presented in Fig. 10. Except for *mental demand*,

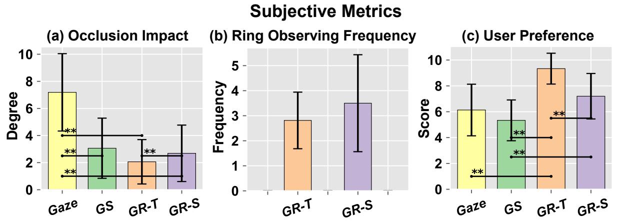


Figure 11: Bar charts of the techniques' performance under different subjective metrics. Error bars indicate the standard error. The statistical significances are labeled with ** ($p < 0.05$). The Ring Observing Frequency is only analyzed for *GR-S* and *GR-T*.

GR-T outperformed *Gaze* and *GS* in each task load with significant differences, indicating a reduction in the loads on participants. In *physical/temporal demand* and *frustration*, *GR-S* was significantly better than *Gaze* and *GS*. These findings suggest that interaction techniques with FPS are generally more user-friendly.

Repeated-measures ANOVAs on other subjective metrics demonstrated significant differences among the four interaction techniques in Occlusion Impact and User Preference. These results are shown in Fig. 11. Participants believed that occlusion had a greater impact on gaze-only interaction than the other three interactions, which is consistent with the results of the objective metric analysis. *GR-T* was significantly preferred by participants compared to the other three interactions, followed by *GR-S*, indicating the eye-gaze and ring-touch combined interaction mechanism is more preferred by participants. There was no significant difference between *GR-T* and *GR-S* in terms of Ring Observing Frequency.

5 Discussion

In this section, we discuss the results for validating the hypotheses.

H_1: *These GazeRing techniques have higher efficiency and usability than using Gaze and GS in object manipulation.*

Our results partially support this hypothesis. In terms of speed, when eye tracking is relatively accurate and objects are unoccluded, GazeRing techniques do not demonstrate significant advantages (see Fig. 9 (a)). However, when objects become occluded or eye tracking accuracy decreases, the advantages of GazeRing techniques emerge, achieving faster interaction speeds (see Fig. 9 (b)(c) and Fig. 8 (a)). Regarding usability, we found that GazeRing was easy to use, especially when eye tracking was inaccurate. Participant P9 stated, "*GR-S* can correct errors and compensate for inaccurate eye tracking." P7 and P15 expressed similar opinions. P14 mentioned, "With *GR-T*, the influence of inaccurate eye tracking was hardly noticeable." P9 reported that "When eye tracking is inaccurate, *Gaze* poses challenges, making it difficult to execute commands precisely." P12 commented, "*GS* allows for adjustments in selection and translation through voice commands in the insufficient eye tracking conditions; however, the delay in speech recognition is relatively high, making fine-tuning difficult." In summary, GazeRing interaction provides higher usability, particularly when eye tracking accuracy is compromised.

H_2: *These GazeRing techniques are more attractive than using Gaze and GS in object manipulation.*

Our results support this hypothesis. In terms of user preference, users primarily ranked *GR-T* first and *GR-S* second, as shown in Fig. 11. The responses to the open-ended questions revealed that most participants found *GR-T* and *GR-S* to be enjoyable. For example, P15 stated, "After a certain period of adaptation, the control method using the pressure ring is relatively precise and provides a good operating experience." Similarly, P12 mentioned, "The sliding interaction of *GR-S* is quite interesting and easy to understand."

The comparison between *GR-S* and *GR-T*. In the preference ranking, participants considered *GR-T* to be superior to *GR-S*. We believe this might be due to the relatively small sensor area of *GR-S*, which supports 8 sliding directions and requires more precise operations. For users with larger thumbs, this may lead to accidental touches. Some participants mentioned, "For *GR-S*, the sensor area is relatively small, making it difficult for fingers to control, and sliding needs to be done slowly" (P7). P2 reported that "It's quite interesting, but it requires a certain level of understanding because the direction is obtained through sliding, and you need to be constantly aware of where you are sliding towards." This might also explain why the observing frequency of *GR-S* is higher than *GR-T* in Fig. 11 (b). P3 believed that "*GR-T* is easier to control selection compared to *GR-S* and is more flexible." However, some users with smaller thumbs felt that "*GR-S* is a bit more effortless than *GR-T*" (P13), suggesting that this size of FPS might be sufficient for them. To address this issue, we plan to use FPS of different sizes to expand the applicability to a wider range of users.

Private and subtle interaction of GazeRing. This paper does not design experiments to compare the use of GazeRing and traditional hand-eye coordination in public settings. Instead, it explores the potential of GazeRing and designs a set of strategies for the combined operation of eye-gaze tracking and ring-based touch. Furthermore, it investigates the performance of GazeRing in object manipulation under two eye-tracking conditions and two degrees of occlusion. Considering public social acceptance, we believe that GazeRing has clear advantages due to the private and subtle nature of eye-gaze and touch-ring interactions. As demonstrated in our supplementary video, GazeRing allows hands to move outside the camera view, increasing the degrees of freedom for hand gesture movements. This feature enhances the privacy and subtlety of the interaction, making it more suitable for use in public settings.

Limitations and future works. We identify four main aspects for future improvements and research: 1) We observed that participants with larger thumbs experienced accidental touches due to the ring's small sensor area. To address this, we plan to use FPS of various sizes to accommodate a wider range of users. 2) The current GazeRing supports only object selection and translation. Future iterations will expand interactions to include object rotation and scaling, to enhance the system's versatility. 3) Exploring the combination of eye gaze with two pressure rings to potentially enhance the overall performance and flexibility of the interaction technique, offering users more control and precision in their interactions with virtual objects. 4) We acknowledge a bias towards younger participants. Additionally, we believe the ring design may be more favorable for female users due to typically smaller finger sizes, facilitating easier manipulation of small sensors. Future research will expand the dataset to include individuals of various ages and backgrounds, with a higher proportion of female participants.

6 Conclusion

In this study, we explore GazeRing, a novel multimodal interaction technique that enables private and subtle hand-eye coordination while allowing users' hands complete freedom of movement. We designed a pressure-sensitive ring equipped with FPS, which supports sliding in eight directions for 3D object manipulation. Additionally, we introduced two control modes for the ring. We developed the eye-gaze and pressure-ring combined interaction mechanism and evaluated the efficiency and usability of four interaction techniques. The experimental results demonstrate that GazeRing exhibits superior efficiency in scenarios involving occlusion or inaccurate eye tracking and is more preferred by users. Our work enhances the privacy and convenience of hand-eye coordination, potentially improving user experience in public settings.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant 62372019.

References

[1] Apple. Introducing apple vision pro: Apple's first spatial computer, 2024.04. https://www.apple.com/newsroom/2023/06/introducing-apple-vision-pro/. 1, 2

[2] S. Aziz, D. J. Lohr, L. Friedman, and O. Komogortsev. Evaluation of eye tracking signal quality for virtual reality applications: A case study in the meta quest pro, 2024. 6

[3] A. N. Balaji, C. Kimber, D. Li, S. Wu, R. Du, and D. Kim. Retrosphere: Self-contained passive 3d controller tracking for augmented reality. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 6(4), jan 2023. doi: 10.1145/3569479 1, 2

[4] Y. Bao, J. Wang, Z. Wang, and F. Lu. Exploring 3d interaction with gaze guidance in augmented reality. In *2023 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pp. 22–32, 2023. doi: 10.1109/VR55154.2023.00018 1, 2, 4, 5

[5] S. Bardot, S. Rawat, D. T. Nguyen, S. Rempel, H. Zheng, B. Rey, J. Li, K. Fan, D.-Y. Huang, W. Li, and P. Irani. Aro: Exploring the design of smart-ring interactions for encumbered hands. In *Proceedings of the 23rd International Conference on Mobile Human-Computer Interaction*, MobileHCI '21. Association for Computing Machinery, New York, NY, USA, 2021. doi: 10.1145/3447526.3472037 2, 3

[6] S. Barteit, L. Lanfermann, T. Bärnighausen, F. Neuhann, and C. Beiersmann. Augmented, mixed, and virtual reality-based head-mounted devices for medical education: Systematic review. *JMIR Serious Games*, 9(3):e29080, Jul 2021. doi: 10.2196/29080 1

[7] Z. Cai, Y. Ma, and F. Lu. Robust dual-modal speech keyword spotting for xr headsets. *IEEE Transactions on Visualization and Computer Graphics*, 30(5):2507–2516, 2024. doi: 10.1109/TVCG.2024.3372092 2, 6

[8] N. Chaconas and T. Höllerer. An evaluation of bimanual gestures on the microsoft hololens. In *Proc. IEEE Conf. Virtual Real. 3D User Interfaces*, pp. 33–40. Reutlingen, Germany, Mar. 2018. 1, 2

[9] Y.-C. Chang, N. Gandi, K. Shin, Y.-J. Mun, K. Driggs-Campbell, and J. Kim. Specifying target objects in robot teleoperation using speech and natural eye gaze. In *2023 IEEE-RAS 22nd International Conference on Humanoid Robots (Humanoids)*, pp. 1–7, 2023. doi: 10.1109/Humanoids57100.2023.10375186 6

[10] I. Chatterjee, R. Xiao, and C. Harrison. Gaze+ gesture: Expressive, precise and targeted free-space interactions. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pp. 131–138, 2015. 1

[11] T. Chen, T. Li, X. Yang, and K. Zhu. Efring: Enabling thumb-to-index-finger microgesture interaction through electric field sensing using single smart ring. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(4):1–31, 2023. 2

[12] A. Colaço, A. Kirmani, H. S. Yang, N.-W. Gong, C. Schmandt, and V. K. Goyal. Mime: compact, low power 3d gesture sensing for interaction with head mounted displays. In *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology*, UIST '13, p. 227–236. Association for Computing Machinery, New York, NY, USA, 2013. doi: 10.1145/2501988.2502042 1, 2

[13] M. Elepfandt and M. Grund. Move it there, or not?: the design of voice commands for gaze with speech. In *Proc. Workshop Eye Gaze Intelligent Hum. Mach. Interact., Gaze-In*, pp. 1–3. California, USA, Oct. 2012. 2

[14] A. O. S. Feiner. The flexible pointer: An interaction technique for selection in augmented and virtual reality. In *Proc. ACM Symp. User Interface Softw. Technol.*, pp. 81–82. BC, Canada, Nov. 2003. 1

[15] J. Franco and D. Cabral. Augmented object selection through smart glasses. In *Proceedings of the 18th International Conference on Mobile and Ubiquitous Multimedia*, MUM '19. Association for Computing Machinery, New York, NY, USA, 2019. doi: 10.1145/3365610.3368416 2

[16] L. Freina and M. Ott. A literature review on immersive virtual reality in education: state of the art and perspectives. In *The international scientific conference elearning and software for education*, vol. 1, pp. 10–1007, 2015. 1

[17] Y. Gu, C. Yu, Z. Li, W. Li, S. Xu, X. Wei, and Y. Shi. Accurate and low-latency sensing of touch contact on any surface with finger-worn imu sensor. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*, UIST '19, p. 1059–1070. Association for Computing Machinery, New York, NY, USA, 2019. doi: 10.1145/3332165.3347947 2

[18] S. G. Hart. Nasa-task load index (nasa-tlx); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, vol. 50, pp. 904–908, 2006. doi: 10.1177/154193120605000909 7

[19] Z. He, C. Lutteroth, and K. Perlin. Tapgazer: Text entry with finger tapping and gaze-directed word selection. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22. Association for Computing Machinery, New York, NY, USA, 2022. doi: 10.1145/3491102.3501838 1

[20] J. D. Hincapié-Ramos, X. Guo, P. Moghadasian, and P. Irani. Consumed endurance: a metric to quantify arm fatigue of mid-air interactions. In *Proc. Conf. Human Factors Comput. Syst.*, pp. 1063–1072. Ontario, Canada, Apr. 2014. 1, 2

[21] M. Kaur, M. Tremaine, N. Huang, J. Wilder, Z. Gacovski, F. Flippo, and C. S. Mantravadi. Where is it? event synchronization in gaze-speech input systems. In *Int. Conf. Multimodal Interfaces*, pp. 151–158. BC, Canada, Nov. 2003. 2

[22] J. J. Kim, Y. Wang, H. Wang, S. Lee, T. Yokota, and T. Someya. Skin electronics: next-generation device platform for virtual and augmented reality. *Advanced Functional Materials*, 31(39):2009602, 2021. 3

[23] Y. Kubo, Y. Koguchi, B. Shizuki, S. Takahashi, and O. Hilliges. Audiotouch: Minimally invasive sensing of micro-gestures via active bio-acoustic sensing. In *Proceedings of the 21st international conference on human-computer interaction with mobile devices and services*, pp. 1–13, 2019. 2

[24] M. Kytö, B. Ens, T. Piumsomboon, G. A. Lee, and M. Billinghurst. Pinpointing: Precise head-and eye-based target selection for augmented reality. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, 2018. 1, 2

[25] J.-J. Lee and J.-M. Park. 3d mirrored object selection for occluded objects in virtual environments. *IEEE Access*, 8:200259–200274, 2020. 6

[26] T. Li, Y. Liu, S. Ma, M. Hu, T. Liu, and W. Song. Nailring: An intelligent ring for recognizing micro-gestures in mixed reality. In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 178–186. IEEE, 2022. 2

[27] C. Liang, C. Yu, Y. Qin, Y. Wang, and Y. Shi. Dualring: Enabling subtle and expressive hand interaction with dual imu rings. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 5(3), sep 2021. doi: 10.1145/3478114 2

[28] H. Lim, J. Chung, C. Oh, S. Park, and B. Suh. Octaring: examining pressure-sensitive multi-touch input on a finger ring device. In *Adjunct Proceedings of the 29th Annual ACM Symposium on User Interface Software and Technology*, pp. 223–224, 2016. 2, 3

[29] M. N. Lystbæk, K. Pfeuffer, J. E. S. Grønbæk, and H. Gellersen. Exploring gaze for assisting freehand selection-based text entry in ar. *Proc. ACM Hum.-Comput. Interact.*, 6(ETRA), may 2022. doi: 10.1145/3530882 1

[30] M. N. Lystbæk, P. Rosenberg, K. Pfeuffer, J. E. Grønbæk, and H. Gellersen. Gaze-hand alignment: Combining eye gaze and mid-air pointing for interacting with menus in augmented reality. *Proceedings of the ACM on Human-Computer Interaction*, 6(ETRA):1–18, 2022. 1

[31] H. Martínez, D. Skournetou, J. Hyppölä, S. Laukkanen, and A. Heikkilä. Drivers and bottlenecks in the adoption of augmented reality applications. *Journal of Multimedia Theory and Applications*, 2:27–44, 03 2014. doi: 10.11159/jmta.2014.004 1

[32] Microsoft. Eye tracking on hololens 2, 2024.05. https://learn.microsoft.com/en-us/windows/mixed-reality/design/eye-tracking. 6

[33] P. Mohan, W. B. Goh, C. Fu, and S. Yeung. DualGaze: Addressing the midas touch problem in gaze mediated vr interaction. In *Proc. IEEE Int. Symp. Mix. Augmented Real.*, pp. 79–84. Munich, Germany, Oct. 2018. 1, 2

[34] L. Pandey and A. S. Arif. Design and evaluation of a silent speech-

based selection method for eye-gaze pointing. *Proceedings of the ACM on Human-Computer Interaction*, 6(ISS):328–353, 2022. 6

[35] K. Pfeuffer, J. Alexander, M. K. Chong, and H. Gellersen. Gaze-touch: combining gaze with multi-touch for interaction on the same surface. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*, UIST '14, p. 509–518. Association for Computing Machinery, New York, NY, USA, 2014. doi: 10.1145/2642918.2647397 1

[36] K. Pfeuffer, B. Mayer, D. Mardanbegi, and H. Gellersen. Gaze + pinch interaction in virtual reality. In *Proc. Symp. Spatial User Interact.*, pp. 99–108. Brighton, United kingdom, Oct. 2017. 1, 2

[37] T. Piumsomboon, D. Altimira, H. Kim, A. Clark, G. Lee, and M. Billinghurst. Grasp-shell vs gesture-speech: A comparison of direct and indirect natural interaction techniques in augmented reality. In *2014 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 73–82. IEEE, 2014. 2

[38] T. Piumsomboon, A. J. Clark, M. Billinghurst, and A. Cockburn. User-defined gestures for augmented reality. In *Proc. Conf. Human Factors Comput. Syst.*, pp. 955–960. Paris, France, Apr. 2013. 2

[39] T. Santini, D. C. Niehorster, and E. Kasneci. Get a grip: Slippage-robust and glint-free gaze estimation for real-time pervasive head-mounted eye tracking. In *Proceedings of the 11th ACM symposium on eye tracking research & applications*, pp. 1–10, 2019. 5

[40] K. A. Satriadi, B. Ens, M. Cordeil, B. Jenny, T. Czauderna, and W. Willett. Augmented reality map navigation with freehand gestures. In *2019 IEEE conference on virtual reality and 3D user interfaces (VR)*, pp. 593–603. IEEE, 2019. 2

[41] Y. Shi, H. Zhang, K. Zhao, J. Cao, M. Sun, and S. Nanayakkara. Ready, steady, touch! sensing physical contact with a finger-mounted imu. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 4(2), jun 2020. doi: 10.1145/3397309 2

[42] L. Sidenmark, C. Clarke, J. Newn, M. N. Lystbæk, K. Pfeuffer, and H. Gellersen. Vergence matching: Inferring attention to objects in 3d environments for gaze-assisted selection. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–15, 2023. 6

[43] L. Sidenmark, C. Clarke, X. Zhang, J. Phu, and H. Gellersen. Outline pursuits: Gaze-assisted selection of occluded objects in virtual reality. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, p. 1–13. Association for Computing Machinery, New York, NY, USA, 2020. doi: 10.1145/3313831.3376438 6

[44] M. Soliman, F. Mueller, L. Hegemann, J. S. Roo, C. Theobalt, and J. Steimle. Fingerinput: Capturing expressive single-hand thumb-to-finger microgestures. In *Proceedings of the 2018 ACM International Conference on Interactive Surfaces and Spaces*, pp. 177–187, 2018. 2

[45] Y.-C. Tung, C.-Y. Hsu, H.-Y. Wang, S. Chyou, J.-W. Lin, P.-J. Wu, A. Valstar, and M. Y. Chen. User-defined game input for smart glasses in public space. In *Proceedings of the 33rd Annual ACM conference on human factors in computing systems*, pp. 3327–3336, 2015. 2

[46] D. Ungureanu, F. Bogo, S. Galliani, P. Sama, X. Duan, C. Meekhof, J. Stühmer, T. J. Cashman, B. Tekin, J. L. Schönberger, et al. Hololens 2 research mode as a tool for computer vision research. *arXiv preprint arXiv:2008.11239*, 2020. 2

[47] R.-D. Vatavu and L.-B. Bilius. Gesturing: A web-based tool for designing gesture input with rings, ring-like, and ring-ready devices. In *The 34th Annual ACM Symposium on User Interface Software and Technology*, pp. 710–723, 2021. 2, 3

[48] B. Velichkovsky, A. Sprenger, and P. Unema. Towards gaze-mediated interaction: Collecting solutions of the "midas touch problem". In *Proc. Human-Comput. Interact.*, pp. 509–516, 1997. 1

[49] P. Wang, X. Bai, M. Billinghurst, S. Zhang, W. He, D. Han, Y. Wang, H. Min, W. Lan, and S. Han. Using a head pointer or eye gaze: The effect of gaze on spatial AR remote collaboration for physical tasks. *Interact. Comput.*, 32:153–169, 2020. 2

[50] T. Wang, X. Qian, F. He, X. Hu, Y. Cao, and K. Ramani. Gesturar: An authoring system for creating freehand interactive augmented reality applications. In *The 34th Annual ACM Symposium on User Interface Software and Technology*, pp. 552–567, 2021. 2

[51] Z. Wang, X. Gu, and F. Lu. Deamp: Dominant-eye-aware foveated rendering with multi-parameter optimization. In *2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 632–641, 2023. doi: 10.1109/ISMAR59233.2023.00078 2

[52] Z. Wang, H. Wang, H. Yu, and F. Lu. Interaction with gaze, gesture, and speech in a flexibly configurable augmented reality system. *IEEE Transactions on Human-Machine Systems*, 51(5):524–534, 2021. 5, 6

[53] Z. Wang, H. Yu, H. Wang, Z. Wang, and F. Lu. Comparing single-modal and multimodal interaction in an augmented reality system. In *2020 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pp. 165–166, 2020. doi: 10.1109/ISMAR-Adjunct51615.2020.00052 2

[54] Z. Wang, Y. Zhao, and F. Lu. Control with vergence eye movement in augmented reality see-through vision. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 548–549, 2022. doi: 10.1109/VRW55335.2022.00125 2

[55] Z. Wang, Y. Zhao, and F. Lu. Gaze-vergence-controlled see-through vision in augmented reality. *IEEE Transactions on Visualization and Computer Graphics*, 28(11):3843–3853, 2022. doi: 10.1109/TVCG.2022.3203110 2

[56] M. Whitlock, E. Harnner, J. R. Brubaker, S. K. Kane, and D. A. Szafir. Interacting with distant objects in augmented reality. In *Proc. IEEE Conf. Virtual Real. 3D User Interfaces*, pp. 41–48. Reutlingen, Germany, Mar. 2018. 2

[57] R. Xiao, J. Schwarz, N. Throm, A. D. Wilson, and H. Benko. Mrtouch: Adding touch input to head-mounted mixed reality. *IEEE transactions on visualization and computer graphics*, 24(4):1653–1660, 2018. 1

[58] K. Xu, Y. Lu, and K. Takei. Multifunctional skin-inspired flexible sensor systems for wearable electronics. *Advanced Materials Technologies*, 4(3):1800628, 2019. 3

[59] D. Yu, X. Lu, R. Shi, H. Liang, T. Dingler, E. Velloso, and J. Gonçalves. Gaze-supported 3d object manipulation in virtual reality. In *CHI '21: CHI Conference on Human Factors in Computing Systems*, pp. 734:1–734:13. ACM, 2021. doi: 10.1145/3411764.3445343 1, 2, 5

[60] T. Zhang, Y. Shen, G. Zhao, L. Wang, X. Chen, L. Bai, and Y. Zhou. Swift-eye: Towards anti-blink pupil tracking for precise and robust high-frequency near-eye movement analysis with event cameras. *IEEE Transactions on Visualization and Computer Graphics*, 30(5):2077–2086, 2024. doi: 10.1109/TVCG.2024.3372039 1

[61] M. Zhao, A. M. Pierce, R. Tan, T. Zhang, T. Wang, T. R. Jonker, H. Benko, and A. Gupta. Gaze speedup: Eye gaze assisted gesture typing in virtual reality. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23, p. 595–606. Association for Computing Machinery, New York, NY, USA, 2023. doi: 10.1145/3581641.3584072 1

[62] M. Zhu, Z. Sun, Z. Zhang, Q. Shi, T. He, H. Liu, T. Chen, and C. Lee. Haptic-feedback smart glove as a creative human-machine interface (hmi) for virtual/augmented reality applications. *Science Advances*, 6, 2020. 2