Order Up! Multimodal Interaction Techniques for Notifications in Augmented Reality

Lucas Plabst, Florian Niebling, Sebastian Oberdörfer, and Francisco Ortega

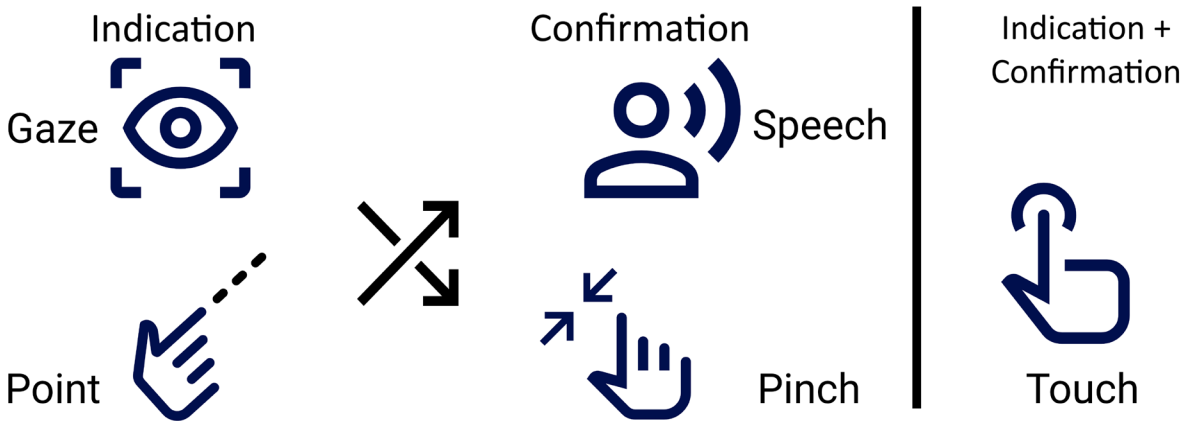

Fig. 1: Modalities used for indication and confirmation of notifications in the experiment. Modalities consist of Gaze, Hand-Pointing, Speech, and Touch. The resulting interactions are Gaze and Pinch, Gaze and Speech, Point and Pinch, Point and Speech and Touch.

Abstract—As augmented reality (AR) headsets become increasingly integrated into professional and social settings, a critical challenge emerges: how can users effectively manage and interact with the frequent notifications they receive? With adults receiving nearly 200 notifications daily on their smartphones, which serve as primary computing devices for many, translating this interaction to AR systems is paramount. Unlike traditional devices, AR systems augment the physical world, requiring interaction techniques that blend seamlessly with real-world behaviors. This study explores the complexities of multimodal interaction with notifications in AR. We investigated user preferences, usability, workload, and performance during a virtual cooking task, where participants managed customer orders while interacting with notifications. Various interaction techniques were tested: Point and Pinch, Gaze and Pinch, Point and Voice, Gaze and Voice, and Touch. Our findings reveal significant impacts on workload, performance, and usability based on the interaction method used. We identify key issues in multimodal interaction and offer guidance for optimizing these techniques in AR environments.

Index Terms—augmented reality, multimodal interaction, eye gaze, speech commands, notifications, gestures, 3D user interfaces

1 Introduction

Notifications are a unique aspect within user interfaces (UI), defined by their dynamic and temporary nature [24]. Notifications possess a unique state, capable of conveying critical information and presenting trivial updates. Their significance stems from their real-time connection to events, messages, or updates, rendering them inherently time-sensitive. Yet, their temporary nature sets them apart, as they can quickly lose relevance if the notification-producing event is no longer relevant. Whether serving as ignorable reminders or vital alerts, notifications try to grab the user's attention. Unless they are ignored until they expire, most notifications are interacted with [52], even if it is just a dismissal, as they do not provide relevant information at that time. Due to their invasive nature, the sudden appearance of a notification and the subsequent interaction with it could cause distraction or frustration. With a smartphone, ignoring a notification while the phone is not in use is as simple as leaving the phone in your pocket until you want to attend it [26]. But what if the display is no longer in your pocket, but is instead worn on your head? In a future where AR glasses could become our means of interacting with the digital and the physical world, finding a way to deal with these distractors is essential. AR can, for example, be a useful tool in surgery [18]. So imagine a future surgeon operating on a patient while wearing an AR head-mounted display (HMD). Patient alarms or other time-sensitive alerts could be directly displayed in the headset without the surgeon needing to turn away from the patient. However, these alerts cannot be detrimental to the current task, and it is crucial that the surgeon can interact with the notifications, such as acknowledging a patient alarm, to ensure the information has been processed and the situation addressed. Even simple urban navigation could be impacted by poorly designed pop-ups, since it has been shown that notifications are highly interruptive, even if ignored [65]. To systematically explore the complexities of AR notifications and their impact on user performance and experience, we first turned to a more general task: a simulated cooking environment. This setting allowed us to control variables and observe the fundamental interactions between users and AR notifications in a broadly applicable way, before extending insights to more specialized and high-stakes scenarios. Our primary contributions in this work are as follows:

- *Lucas Plabst is with Computer Science & NUILAB at Colorado State University*
  E-mail: lplabst@colostate.edu
- *Florian Niebling is with Advanced Media Institute at TH Köln*
  E-mail: florian.niebling@th-koeln.de
- *Sebastian Oberdörfer is with HCI Group at University of Würzburg*
  E-mail: sebastian.oberdoerfer@uni-wuerzburg.de
- *Francisco Ortega is with Computer Science & NUILAB at Colorado State University*
  E-mail: fortega@colostate.edu

Received 18 September 2024; revised 13 January 2025; accepted 13 January 2025. Date of publication 7 March 2025; date of current version 31 March 2025. Digital Object Identifier no. 10.1109/TVCG.2025.3549186

- We provide an analysis of various interaction modalities for AR notifications, demonstrating how their combinations influence user perception and performance.
- Our findings show that compared to using each modality on its own, integrating multiple interaction modalities yields superior results.
- We identify the ongoing challenges and limitations faced by multimodal interactions in AR environments.
- Based on our findings, we offer suggestions for designing future AR systems that effectively leverage multimodal interactions.

2 Related Work

2.1 Notifications in Mixed Reality

Notifications have been studied in desktop environments [14, 15] and more recently in the smartphone field, given their ubiquitous use, with an estimated 90% of American adults owning and using smartphones daily [50]. Working adults receive an average of between 45 to 80 notifications per day, with some groups like college students receiving over 400 [1, 60, 39]. Research by Pielot et al. [52] revealed that disabling notifications for a day led to participants feeling less distracted and more productive, though they also expressed concern about missing important information and feeling disconnected from their social networks. With Mixed Reality (MR) anticipated to grow at a rate of 30% annually until 2032 [10], and more companies releasing MR-HMDs, a future where these headsets become our primary means of computing is plausible. AR-HMDs have a great potential to become pervasive in day-to-day life, as they do not isolate the user from their surroundings and can supplement the real world with helpful information. In this work, we refer to AR as seeing the real-world with virtual objects superimposed onto or composited with, following Azuma et al. [6]. We use MR as an umbrella term referring to a part of the Reality-Virtuality Continuum [63], containing both AR and VR. Janaka et al. [26] have demonstrated that messaging using an AR-HMD showed better multitasking capabilities and faster response times than using a smartphone, and suggest that even now the usage of an HMD as a supplemental device to the phone can improve messaging during multitasking. Research on notification modalities in VR by Ghosh et al. [19] and in AR by Lazaro et al. [35] found that combining visual and auditory notifications improved preference and performance. Participants expressed not only wanting to receive notifications but also to interact with them.

Another output modality that has been researched is the position of notifications in AR. Plabst et al. [53] looked at notification placement in AR and found a bottom-center position in the user's field of view to be optimal in general use cases, world placement in stationary tasks, and body placement in tasks with a lot of digital content. Similar conclusions were reached in a study about VR notification placement [59], as well as in an AR experiment about dual-task scenarios [13]. Imamov et al. [23] also researched glanceable information in simulated AR and found that a bottom-center position close to the primary task was more comfortable and faster than other placements. It is however worth noting that because of the inherent differences between AR and VR, it is unclear how the results of VR experiments apply to AR. Rzayev et al. [58] also showed the effect notification position in AR can have in a social interaction setting. They found that most participants preferred receiving notifications on an AR headset rather than a smartphone and that the position affected the person wearing the headset and the person not wearing the headset differently. Lee et al. [36] found that the position of notifications in AR affected response time, as well as inversely affecting walking speed, leading to the conclusion that the priority of the notification content needs to be considered when deciding where to place them. Another position for glanceable information in AR was researched by Satkowski et al. [62] who found that the ceiling or the floor of a physical space are well suited to display secondary non-critical information in AR without being obtrusive. Looking more at presentation as well as location, Lucero et al. [41] proposed a minimal notification interface in AR that did not distract participants from walking in public while still giving them updates. Notifications were presented as butterflies moving across the field of view, that could be opened and dismissed using swipe gestures on a finger-mounted touchpad.

2.2 Augmented Reality Interaction

In their study on general interactions in AR, Wang et al. [66] explored the fusion of various input modalities. They discovered that a combination of Gaze, Gesture, and Speech yielded the most efficient results, although users preferred the combination of Gesture and Speech. Conversely, unimodal interactions using solely Gaze or Gestures generally underperformed compared to the multimodal options. Lee et al. [38] similarly investigated multimodal versus unimodal inputs in AR, noting that while participants favored multimodal input over sole Speech or Gaze, they did not demonstrate objective performance improvements with it. For accessing generic information on an AR headset, Lu et al. [40] devised a system enabling users to summon information like weather updates or sports scores through either head movement or gaze. They observed that users preferred and performed best with gaze-based interactions.

Kosch et al. [31] evaluated three AR notification selection techniques while cycling: gaze with a dwell time, gestures, and gaze for indication with a physical button press for confirmation. They found that gaze with dwell time resulted in the lowest error rates, but required more attention and focus, whereas the gaze with button approach was preferred by users and had the lowest task completion time. Also using a dedicated input device for AR, Cai et al. [11] used a circular UI on an HMD that was controlled with a ring mouse worn on the hand, which they called *ParaGlassMenu*. In that experiment, Cai et al. compared this with more linear displays, voice assistants, and a smartphone in a conversational setting. They found that the *ParaGlassMenu* outperformed the other interactions in performance as well as subjective measures. Plabst et al. [54] investigated various display types for multiple notifications in AR, along with unimodal interaction techniques. Their findings indicated that Touch outperformed Voice or Gaze, and participants favored having a notification list attached to their hand rather than in the surrounding environment.

3 Experimental Setup: Kitchen Environment

In this work, we researched how we can interact with notifications in AR using multimodal interaction. Especially because notifications are usually short-lived, interaction needs to be quick, easy, and not distracting. For this reason, we set out to answer two research questions:

RQ 1: Is the user's perceived usability of notifications influenced by the interaction technique?

RQ 2: Is the primary task performance influenced by the interaction technique of notifications?

Usability has been defined by the International Organization for Standardization (ISO) as "the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use" [25]. We built upon the open-source research environment proposed by Raikwar et al. [57] to conduct our experiment. While in the cooking environment, participants engaged with various notifications using the Hololens 2, an optical see-through (OST) AR-HMD. The OST design of the headset not only grants users real-time visibility of their hands and bodies but also enhances depth perception, which distinguishes it from VR headsets [29], while also being more accurate for indication and confirmation tasks [33]. In comparison with video see-through AR, OST was found to have better text legibility [16] and better depth perception [2]. The HMD's eye-tracking functionality demonstrated precision within a 1.5° visual angle around the intended target, a finding by Kapp et al. [30]. Hand-interactions, eye-tracking, and voice recognition were implemented using the Mixed-Reality-Toolkit 2.8.3 [46]. Performance logs show that the environment runs at a consistent 60 frames per second, with infrequent drops to around 40.

3.1 Task

In this simulation environment, participants were tasked with fulfilling customer food orders, representing different real-life situations. These

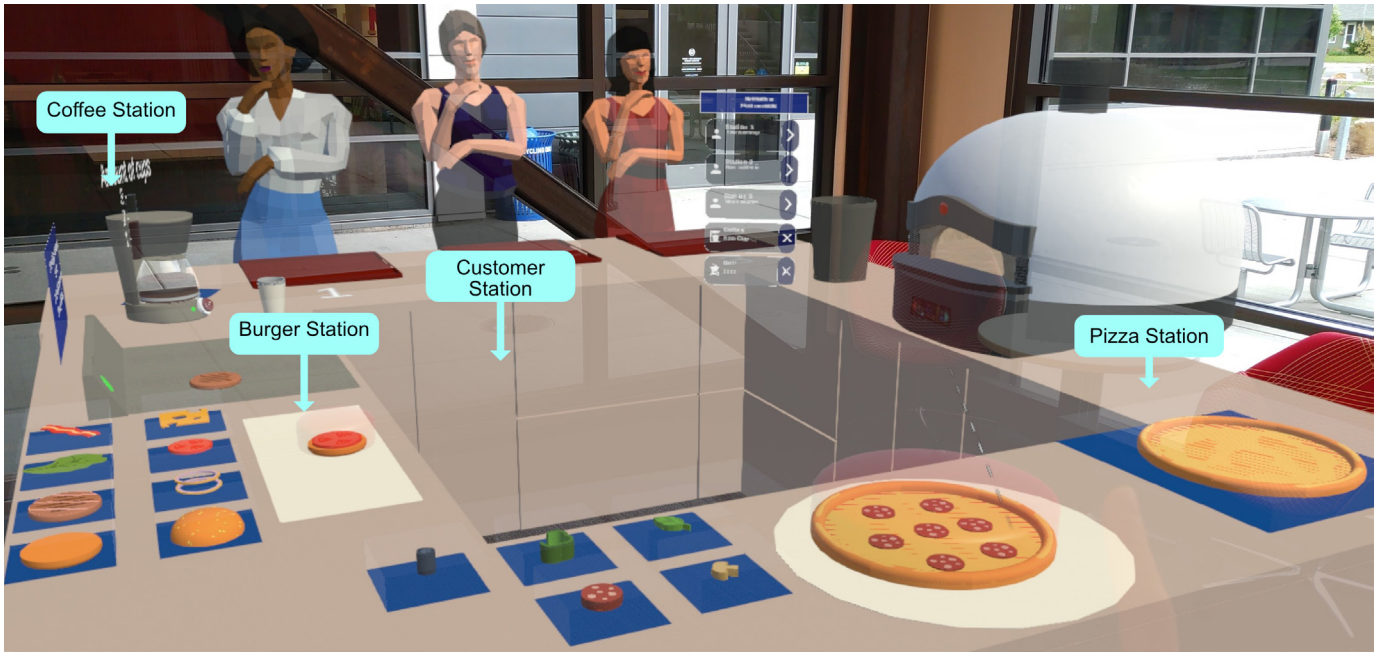

Fig. 2: Virtual cooking environment displayed in a large room in Augmented Reality

scenarios can range from high-stress to low-stress environments, allowing users to handle multiple tasks at once or focus on one thing at a time, moving around or staying in one spot, depending on how the kitchen was set up. Stress can be regulated by several factors, including but not limited to the number of customers, the amount of ingredients on meals, the waiting time for customers, or the time until food is burnt and needs to be recooked. For our work the kitchen task is used as a general use case that people could encounter in their daily life, trying to perform multiple things at once with some time sensitivity. The use of a virtual cooking environment in our study serves as a model for understanding interaction techniques with AR notifications. This environment mirrors several essential elements found in professional settings, allowing us to draw broader conclusions while also exploring specific nuances relevant to high-stakes scenarios. Both cooking and professional environments involve complex task management, where users juggle multiple tasks simultaneously, requiring attention to detail, and prioritization. The cooking environment also demands real-time decision-making, similar to professional settings. A cook must respond promptly to a timer, much like a mechanic reacts to machinery changes. This need for timely responses to notifications is common across different contexts. The choice of a kitchen reflects the need for a versatile, general-use scenario that is easy for participants to understand and engage with, while still being representative of the cognitive and physical challenges encountered in complex, time-sensitive work environments.

The kitchen (see Figure 2) was around 3 meters by 3 meters in size and had four stations. Participants could use both hands to handle food and ingredients and were instructed to grab the virtual objects just like they would with physical ones. At the **Customer Station**, up to three customers could order food. Each customer was represented by a simple human avatar with slight movements. A red tray indicated where the prepared food had to be served. When an order was accepted by the user by interacting with a notification, a two-minute timer appeared, showing how much time was left to finish. No direct interaction with the avatars occurred. The **Coffee Station** had a coffee machine with a pot. Users had to press a button to start the machine, and when enough for a cup was made, a notification alerted the user. They could then pour coffee into a cup, and a lid appeared when it was full, signaling it was ready. The **Pizza station** and **Burger station** had ingredients to assemble the meal and a cooking device. The patty or pizza dough had to be cooked for at least 10 seconds to be done, upon which a notification was sent. After 40 seconds, the food would burn and would not be accepted by the customers. Next to the Pizza station, participants could find a trash can to discard incorrectly prepared food items. A movable list was next to the trash can, where notifications would move after being in the field of view for 8 seconds.

3.2 Notifications

Participants received several notifications throughout their task. Every notification featured a title indicating its source, with the content of the notification underneath. Next to this text was an icon corresponding with the notification's title for quicker information access [27]. We limited notifications to task-related ones for privacy reasons, to control the notification amount, and to ensure participants paid attention to them, as they directly related to their task.

The notification text was set within the comfortable range for text size [44], making it easy to read and allowing all notification texts to fit inside the panel without cutting off the text or decreasing font size to fit. Text was displayed with a white font and a dark gray background for optimal readability in most situations [28], as the changing background in AR environments can lead to contrast issues depending on color choice. The moment a notification was delivered, a sound was played to alert the user of its presence, as this leads to better recognition and performance [35], and ensures that attention is drawn towards the notifications, so that missing notifications is unlikely. Based on previous findings [53, 37, 13], the notifications were positioned in the bottom center of the device's field of view. This position was found to be ideal as it does not distract the user from their task too much while still being noticeable, while also allowing the user to move around without losing sight of the notification. Notifications were placed at a distance of 75 - 90 cm, keeping them within the acceptable range recommended in the development guidelines for the headset [44]. We chose this distance because placing the notifications closer than 75cm could increase eye-fatigue due to the vergence-accommodation conflict [32], whereas placing them too far away would make touch interaction impossible, and could introduce text legibility issues.

Notifications had a blue button on the right side that was used for interaction. Interacting with the button caused an action, depending on whether it was an order or food notification. Looking at user experience design patterns [4, 3], they describe two types of notifications: actionable notifications, which the user has to react to, or informational notifications, whose purpose is to relay information to the user. The main difference between the two types is the expectation of interac-

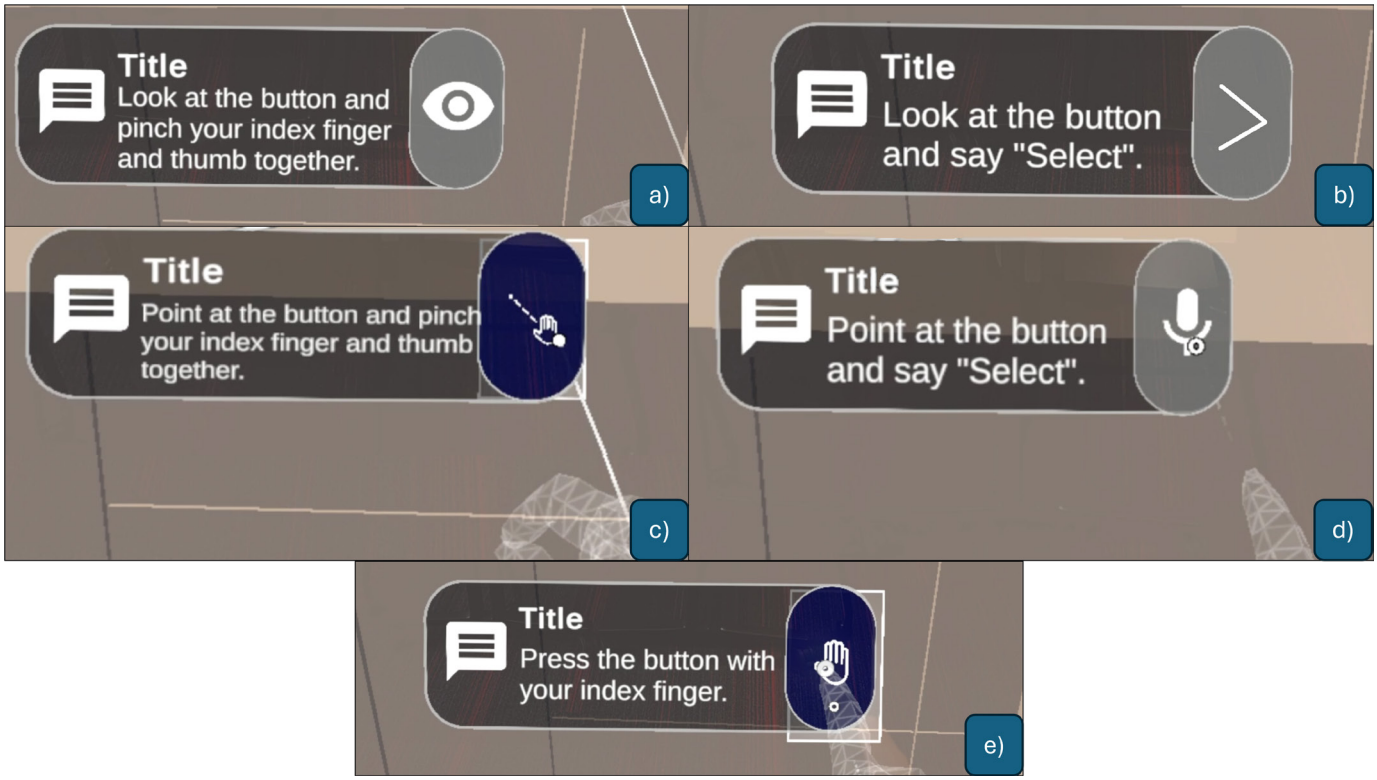

Fig. 3: Interaction techniques used in the experiment: a) Gaze and Pinch; b) Gaze and Speech; c) Point and Pinch; d) Point and Speech; e) Touch

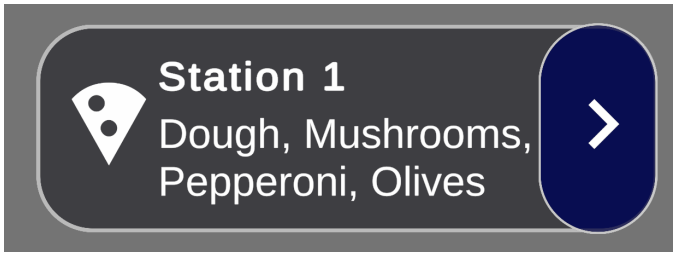

Fig. 4: A notification for a pizza order.

tion. Whereas informational notifications do not require a response, actionable notifications at least expect to be attended. For example, a notification informing a user about today's weather does not expect a response, whereas a notification about an incoming phone call or direct message does expect a response. Since users deal with notifications differently depending on their context [52], it was important to not only have a single type of notification in this experiment. The system we used deploys two types of notifications: order notifications and food notifications.

Order notifications: An order notification was sent for each new customer. This notification had to be interacted with to see what the order was, therefore making it an actionable notification. Upon accepting the order, the notification content would change to show the customer's order (see Figure 4). This would also start a visible two-minute countdown, showing the time remaining to prepare the food, after which the customer. When the meal was placed on the tray before the customer, the notification had to be interacted with again to confirm the order. If the ingredients matched, a "success" tone would play and the notification would display "correct order" before disappearing. If the food was incorrectly prepared, a "failure" tone would play and the notification would display "incorrect order". Activating the button again would cause the "incorrect order" text to disappear and the notification would display the ingredients for the food ordered again.

Food notifications: Every station produced notifications when the desired cooking state was reached. Activating the button dismissed the notifications, making them informational notifications since no direct action was required by the user as a result of the notification, and their purpose was strictly to convey information. Interaction with this notification was not necessary for completing the task, but it could be dismissed for better organization.

3.3 Interaction Techniques

Plabst et al. [54] researched notification interaction. However, their work is limited by the fact that they only looked at unimodal interaction techniques. In our work, we wanted to research multimodal interaction for notifications, as they might offer more flexibility and performance improvements, as well as being preferred by users over unimodal interaction techniques [61, 12]. In the work by Plabst et al. [54], touch interaction was unilaterally found to be the optimal technique, so we used it as a control condition to compare against multimodal interaction techniques (see Figure 3). They found several problems with Gaze and Voice, such as participants' frustration with the artificial slow-down in Gaze interaction due to the dwell-time, or issues with selecting the correct notification in Voice interaction. Our multimodal approach has the potential to address these issues by allowing users to combine complementary input methods, eliminating the need for dwell-time or labels for specific voice commands.

The interaction techniques we used are split into two blocks: Indication and Confirmation. Bowman et al. [9] classify Indication and Confirmation as two steps in a 3D manipulation task. The last step in their classification is Feedback, which is provided in our system by visual and auditory confirmation of the selection. One modality is used to indicate the notification the user wishes to interact with, whereas the other confirms the interaction, in our case the pressing of a button. The two modalities are used *complementarity* [42], meaning that they are processed individually but are merged for a single interaction, leading to faster interaction. We found that the use of hands, eyes, and voice are all extensively studied techniques for MR devices [49]. These modalities are also commonly found and used in state-of-the-art MR

headsets like the Hololens 2, Meta Quest Pro, or Apple Vision Pro. For this study, we therefore chose the following techniques:

Point and Pinch (PP): When pointing at something in the environment, a ray appears extending from the wrist of the user's hand. Notifications could be interacted with by pointing at the button and then pinching the index finger and thumb together. Correctly pointing at the button caused it to be highlighted, confirming the target. Notifications using the pointing interaction were placed at a distance of 90 cm instead of 75 cm like the others, as the shorter distance would require the arm to be held in an uncomfortable position. This technique is used as the primary means of interaction by the Hololens 2 [45], as well as the Meta Horizon OS (Meta Quest) [43].

Point and Speech (PS): Utilizing the same ray, participants could point at the notification button and say the command "Select" out loud while still pointing. Through informal piloting and based on previous literature showing a preference for short voice commands in multimodal interaction [48], we chose this keyword, as it was short, easily recognizable by the Hololens voice recognition, and easy for the participants to remember. This interaction technique, introduced by Bolt et al [8], has since been used extensively in 3D environments.

Gaze and Speech (GS): Subjects could use the device's eye-tracking capabilities to select the button by looking at it, causing it to turn gray to reflect the state of being looked at. They could then say the "Select" keyword to activate the button. Using any confirming action with Gaze also alleviates the Midas touch problem, where systems cannot differentiate between basic eye functions like looking from deliberate interaction.

Gaze and Pinch (GP): Like GS, the button could be highlighted by looking at it and then activated by pinching the thumb and index finger together on one hand without needing to point using the hands. This interaction technique is also used as the primary interaction with headsets like the recently released Apple Vision Pro [5].

Touch: Each notification button was selected and subsequently activated by pressing it with the index finger of either hand. Pushing the button was animated to give the user the appropriate feedback since mid-air touches cannot provide haptic feedback.

As participants needed to grab objects in the environment using their hands, there was a potential mismatch between modalities in the main task and the interaction with the notifications. Our task is a simulation of a real-world task that requires the use of hands, and in the real world, the only interaction with digital content would be the notifications. Also, switching modalities between the UI and the main task is a common practice in 3D applications like video games, where the game's controls and the UI's controls often differ [55, 56]. For example, a VR video game might use the controllers as virtual hands to interact with the environment but use a ray for UI interaction.

4 Experiment

We used a within-subjects design with one independent variable: *interaction technique* (GS, GP, PP, PS, and Touch), yielding five total conditions. Latin-square counterbalancing was applied to set the order of conditions. Overall, 30 participants were recruited from a university campus; however, due to technical issues three of them did not complete the experiment. We analyzed data of the remaining 27 participants (17 male, 9 female, 1 gender-fluid) aged between 19 and 34 (M = 24.4, SD = 3.7). They were given a $30 US-dollar equivalent gift card as compensation in local currency. All participants self-reported normal or corrected to normal vision and English proficiency. Only two participants reported using a Hololens 2 (or any OST HMD) before. The experiment was carried out under the supervision of the Institutional Review Board (IRB) of Colorado State University.

4.1 Procedure

The experiment occurred in a spacious lecture room, approximately ~10x6 meters in size. Blinds were lowered to minimize sun-blinding, and lights were turned on to ensure consistent lighting for all participants. The cooking environment was positioned in the room, with a marker for consistent placement among participants. To ensure consistency for comparison, we replicated the procedure used by Plabst et al. [54].

Upon arrival, participants filled out a demographics questionnaire and gave their informed consent after reading the consent form. They were then instructed on how to use and wear the Hololens 2. After calibration for their eyes, ensuring correct display placement, and eye-tracking setup, participants entered the virtual cooking environment. Here, they underwent a tutorial explaining system functions, followed by a training segment covering all possible notification interactions. This made sure that the systems worked with the participant, and that the participant understood the interactions. Subsequently, participants engaged in a task where they prepared various food items and served customers at their own pace. Once they felt ready, they informed the experimenter to start the experiment. Participants had to fulfill six orders per trial, each comprising a single food type with randomized ingredients for burgers and pizzas. They had to prepare each food type twice in a random sequence. Upon completing all orders or letting them expire, participants were prompted to fill out questionnaires on a tablet. This process repeated for all five conditions before participants could remove the headset and complete the post-questionnaire.

4.2 Measures

When an experiment trial was completed, the participants were asked to fill out questionnaires about their experience with the notifications. A Raw NASA-Task-Load-Index (TLX) questionnaire [22, 21] was used to understand the perceived task load. Besides this, we deployed the System Usability Scale (SUS) questionnaire [20] to understand the system's overall usability. When all trials finished, participants were asked to rank the interaction techniques according to their usage preference and explain their ranking. For performance measurements, the device logged and measured the following metrics during the experiment: For general task performance, we measured the **total time per trial** and the **time** needed **per customer**. To understand the accuracy of task completion, we measured the amount of **incorrect orders** and **expired orders**. **Time until an order notification was accepted** was measured to understand the multitasking performance.

5 Results

For all measures, we calculated a repeated-measures ANOVA for the independent variable *interaction technique* and conducted post-hoc analysis by running Tukey pairwise tests. Almost all measures met assumptions for ANOVA and were tested using Levenes tests (homogeneity of variance) and Shapiro-Wilk tests (normal distribution). Time until order was accepted was found to be not normally distributed, so we used a Kruskal-Wallis Test for main effect and Wilcoxon rank sum tests for pairwise comparisons. Overall, we found participants preferred using Gaze and Speech (GS) less than the other techniques, significant differences in physical workload as well as usability, significant differences in trial time, and that Touch performed better than Point and Speech (PP) in the time it took to start an order.

5.1 Subjective Measures

Task-Load-Index: We found no significant effect on the overall TLX score ($F(4,104) = 1.964$, $p = 0.122$). Looking at the individual subscales, we found a significant effect on the physical demand ($F(4,104) = 3.082$, $p = 0.019$). In the post-hoc analysis, we found no significant differences between individual groups. We found no significant differences for mental demand ($F(4,104) = 2.343$, $p = 0.06$), frustration ($F(4,104) = 2.364$, $p = 0.058$), effort ($F(4,104) = 1.361$, $p = 0.253$), and performance ($F(4,104) = 1.488$, $p = 0.211$).

System Usability Scale: We found a significant effect on the SUS score ($F(4,104) = 5.643$, $p < 0.001$). When conducting post-hoc analysis, we found no significant differences between individual groups. When analyzing the SUS, a value of 68 would place a score into the 50th percentile, with any value above being considered good usability. Looking at the mean values in Figure 2 shows us that only Touch falls into this range of good.

Preference Ranking: We assigned points for each rank using the Borda-Count method [17]. If a technique was ranked as the most

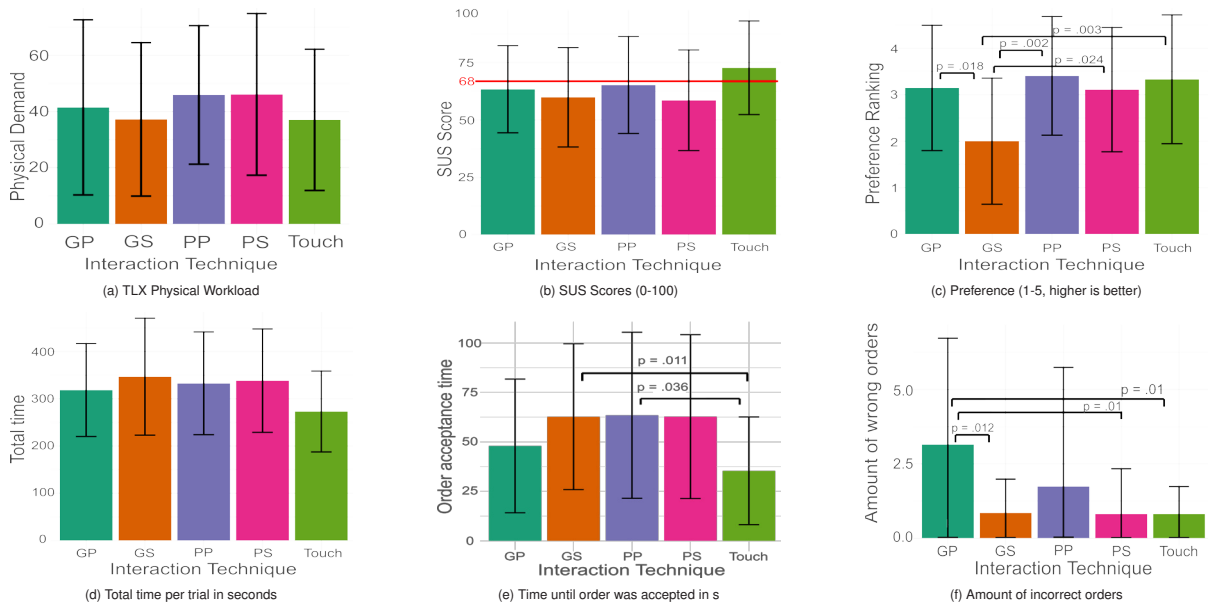

Fig. 5: Performance and subjective measures, bars show standard error

Table 1: TLX Scores with subscales (0-100), SUS Scores, and Preference Rankings (values are Mean, *SD*, `Median`).

Condition	TLX Score			Phys. Demand			Frustration			Ment. Demand			SUS Score			Pref. Rank		
	Mean	*SD*	Med.	Mean	*SD*	Med.	Mean	*SD*	Med.	Mean	*SD*	Med.	Mean	*SD*	Med.	Mean	*SD*	Med.
GP	43.09	*23.93*	45.0	41.48	*31.16*	40	38.52	*28.21*	35	42.59	*28.02*	40.0	63.4	*19.0*	67.5	3.15	*1.35*	3.0
GS	41.67	*23.01*	45.0	37.22	*27.29*	30	32.96	*25.47*	30	41.30	*27.82*	40.0	59.9	*21.7*	65.0	2.00	*1.36*	2.0
PP	41.98	*21.70*	45.0	45.93	*24.65*	55	32.22	*29.59*	25	44.44	*24.94*	45.0	65.3	*21.2*	70.0	3.41	*1.28*	3.0
PS	47.13	*24.76*	50.0	46.11	*28.77*	60	45.19	*31.55*	40	48.33	*27.87*	50.0	58.6	*22.0*	55.0	3.11	*1.34*	3.0
Touch	37.65	*20.25*	41.6	37.04	*25.13*	35	29.44	*25.73*	20	37.22	*24.03*	35.0	72.8	*20.4*	72.5	3.33	*1.39*	3.0

preferable, it received 5 points, whereas the least preferable ranking was 1 point. We found a significant effect of *interaction technique* on the preference ranking score ($F(4,104) = 3.924$, $p = 0.005$). When conducting post-hoc analysis for pairwise comparisons, we found significant differences between GS (M = 2.00, SD = 1.36) and Touch (M = 3.33, SD = 1.39) p = 0.003, GP (M = 3.15, SD = 1.35) p = 0.018, PP (M = 3.41, SD = 1.28) p = 0.002, and PS (M = 3.11, SD = 1.34) p = 0.024.

5.2 Performance

Trial Time: We found a significant effect on the time per trial ($F(4,104) = 3.617$, $p = 0.008$). In the post-hoc analysis, we found no significant differences between individual groups.

Time until order accepted: This measurement was started when the notification was initially sent and stopped when the participant acknowledged the order. We found a significant effect on the time until the order was accepted (Chi square = 14.457, p = 0.006, df = 4). In the post-hoc analysis, we found significant differences between Touch (M = 35.40, SD = 27.20) and PP (M = 63.46, SD = 41.96) p = 0.036 and GS (M = 62.7, SD = 36.9) p = 0.011.

Time per customer: The time per customer began to be measured when the participant accepted the order. We found no significant effect on the time spent per customer ($F(4,104) = 0.602$, $p = 0.662$).

Incorrect Orders: We found a significant effect on the wrong order number ($F(2.28,59.18) = 5.575$, $p = 0.004$). In the post-hoc analysis, we found significant differences between GP (M = 3.15, SD = 3.58) and Touch (M = 0.81, SD = 0.92) p = 0.01, GS (M = 0.85, SD = 1.13) p = 0.012, and PS (M = 0.81, SD = 1.52) p = 0.01.

Expired Orders: We found no significant effect on the number of expired orders ($F(4,104) = 0.673$, $p = 0.612$).

5.3 Preference Responses

Participants were asked to explain their preference ranking. These interview responses were then collected into an Affinity-Diagram [7], where key points were identified from the responses.

Generally, participants had problems with the reliability of some modalities. Starting with eye-tracking, participants found it to be sometimes unreliable and expressed that it took the system too long to recognize that they were looking at the target. P1 said "I did not find the eye tracking as easy to use because often the headset didn't recognize that I was looking at the button," with P13 expressing that "eye-tracking is also hard as it needed time and was hard on my eyes to wait until the system recognizes me."

Participants uttered the same issues about the voice recognition, as they felt that the system did not understand them consistently, with P16 saying, "Sometimes the speech recognition did not recognize my command well, and it made the cooking experiment a bit frustrating." In addition to the unreliable detection, participants were sometimes frustrated by needing to keep saying the keyword repeatedly, which was amplified by needing to repeat a word when it was not recognized. P1 said that they "also didn't like the voice commands that much, because it seemed kind of annoying to have to repeat the word " with P13 adding that "Voice command made me self aware." Most of the negative comments surrounding voice have to do directly with the system not understanding them well enough. However, this did not affect all users, as PS did not score lower on the preference rankings, with P7 stating that they "felt like they had the most control" when using PS. This highlights that detection accuracy is a key issue holding back speech interaction.

Lastly, reliability was also a common theme in the pinching (PP and

Table 2: Performance Results (values are Mean, *SD*, Median).

Condition	Trial time in seconds			Time until order was accepted (s)			Time per customer (s)			Expired orders			Incorrect orders		
	Mean	*SD*	Med.	Mean	*SD*	Med.	Mean	*SD*	Med.	Mean	*SD*	Med.	Mean	*SD*	Med.
GP	318.5	*98.3*	303.93	48.0	*33.8*	40.46	66.1	*19.2*	61.58	1.63	*1.78*	1	3.15	*3.58*	2
GS	346.7	*123.8*	327.51	62.7	*36.9*	55.20	68.4	*16.0*	64.72	1.25	*1.51*	1	0.85	*1.13*	0
PP	332.7	*108.7*	313.34	63.5	*42.0*	52.19	63.1	*13.7*	60.56	1.15	*1.26*	1	1.74	*4.01*	1
PS	338.3	*109.4*	296.57	62.8	*41.4*	50.80	69.5	*23.1*	62.89	1.26	*1.81*	0	0.81	*1.52*	0
Touch	273.0	*85.6*	254.17	35.4	*27.2*	30.83	66.5	*17.1*	63.21	1.11	*1.57*	0	0.81	*0.92*	1

GP) conditions, as participants had issues getting the headset to detect the pinching gesture reliably. P27 attributed this to the headset's small detection field, saying, "The pinch felt inconsistent because I really had to make sure my hand was in front of me," and P5 said that it was "always hard to register the pinch." The registration of the gesture seems to be a trade-off between accuracy and false positives. In their study evaluating Gaze and Pinch interaction, Pfeuffer et al. [51] observed many issues with the system falsely registering small hand movements as a pinch and suggest a more explicit pinch gesture like the one the Hololens uses. However as we saw in our experiment, if the gesture is too explicit, the user frustration is still there, as instead of being too easy to perform, the gesture is now too hard to perform.

We can now also identify an inherent problem of the multimodal interaction techniques used in this experiment: the latency between indication and confirmation. Since both recognition systems take time to register, the interaction relies on both systems registering simultaneously so as not to cause a mismatch. If there are underlying delays or reliability issues with both indication and confirmation, the resulting interaction can cause frustration. This highlights that even if the interaction itself is well-liked and appropriate, a perceived lack of reliability can dramatically decrease the user's experience with the system.

Point and Pinch (PP): Participants felt that PP was effective overall but that it was sometimes hard for them to get the action correct, "as there are some chances of pointing something else while pinching" (P9).

Point and Speech (PS): Participants mentioned that while they found PS convenient to use, they also thought it was uncomfortable and "annoying, with the disadvantage of having to talk and physically point" (P10).

Gaze and Pinch (GP): Using GP, participants stated that they felt there was a disconnect between the indication and confirmation, as "it felt very unnatural to look at the option and select it with my hand" (P15) and "I needed to concentrate on both while managing both, it felt a little bit difficult" (P23).

Gaze and Speech (GS): Comments about GS mostly focused on the difficulty of maintaining one's gaze on the target while saying the keyword. P9 stated: "Eye tracking and voice command [are] very difficult for me because while talking my eye pointing changes and it makes me concentrate more on the selection", with P26 echoing the complaint: "I found it difficult to look in the specific spot for the button while telling it to select."

Touch: Over half of the participants directly expressed liking Touch, as they felt it was the most intuitive and easy interaction method. P14 said that "We touch things in real life, so it is the most interactive and realistic." However, some participants had problems pressing the button when it popped up initially: "Touch didn't work well for me. Sometimes it seemed that buttons were too far" (P2), or when they wanted to interact with the notifications and the list was not within reach: "Touch is easiest to use but when there is a list it is difficult to reach it" (P13).

6 Discussion

Comparing our findings to the results of Plabst et al. [54], we can see that multimodal interaction techniques mostly close the gap to the unimodal "winner," Touch, when compared to unimodal interaction techniques. In their study [54], touch was rated significantly higher in preference than the unimodal techniques, while also performing better and being rated higher in usability. In our study, however, Touch was not rated significantly higher than the multimodal interactions in preference, nor did it have higher ratings or performance benefits besides outperforming PP and GS in order of acceptance times. A challenge we and our participants encountered with touch interaction is the balance required in positioning user interface elements. If placed too close, users may experience discomfort, partly due to the vergence-accommodation conflict, which remains a concern in MR devices [32]. Given that the Hololens' focal distance is approximately two meters [44], only distant content is unaffected by this issue. While positioning elements farther away could enhance comfort and mitigate the conflict's impact, it would render direct touch interaction impossible. Mentioning input legacy bias is also still important [34, 67]. Touch interaction is a natural part of human behavior, spanning both physical and digital environments. From how we communicate and connect with others through physical touch to how we interact with touchscreen devices, it's a fundamental aspect of our daily lives. Whether shaking hands or swiping on a smartphone, touch plays a central role in our experiences, making this interaction more intuitive and efficient. While the task in the experiment was performed using hands, most general daily activities also require the use of hands, making this a well-suited metaphor for daily activities. The other techniques we investigated in this experiment are novel and not commonly found or used in other computing tasks. This is highlighted by P1: "My favorite options were just the standard touch or point and pinch, probably because that is more like what I'm used to doing in real life and in my past experiences with VR/AR systems." Exposing subjects to these multi-modal interaction techniques for a longer time might enhance the user's abilities with them.

To answer RQ1 ("Is the user's perceived usability of notifications influenced by the interaction technique?"), our findings show that GS was rated as being significantly less preferable than the others. We also found effects on SUS and physical workload overall but were not able to find any differences between individual groups. When explaining their rankings, participants often didn't directly refer to the multimodal technique but rather to the individual modalities. More often than not, they directly mentioned issues with the overall modalities, causing them to rank techniques containing that modality as less preferable. Participants frequently encountered challenges in coordinating two modalities. They reported looking or pointing away before completing actions like pinching or speaking, resulting in misalignment. This underscores a challenge in multimodal systems, where latency or recognition errors can significantly impact interaction quality, leading to frustration and errors. Despite utilizing state-of-the-art eye tracking in the Hololens 2 [30] in well-controlled environments, reliability issues persisted, indicating potential challenges in real-world scenarios with less-than-ideal conditions. Prioritizing the reliability of recognition is essential in designing effective interactions. Implementing mechanisms like predicting the selected target based on dwell or pointing time before executing a manipulation command, or smoothing out pointing movements, may help alleviate these challenges, but further research is needed into optimizing complimentary modalities.

Lastly, while both hands could be used in the experiment, the task could be completed entirely with one hand only. Entirely hands-free interaction like GS could be more appropriate in situations where both hands are busy constantly and might also influence the preference

rating of this specific interaction, as participants always had a hand-free to interact with. However, tasks requiring both hands could also demand the user's full attention, locking their gaze onto the task and necessitating a break in focus to interact, making the benefits of hands-free interaction less clear. Since the preference rating did not differ greatly besides this significantly lower preference for GS, allowing the user to choose their preferred interaction is something we recommend.

Answering RQ2 ("Is the primary task performance influenced by the interaction technique of notifications?"), we can see that the number of incorrect orders, representing the accuracy of the task completion, was significantly higher with GP than with GS, PS, and Touch. The higher number can either be attributed to the wrong execution of the food order or a misinput if there was no food on the tray, where the user wanted to choose a different notification or none at all but still confirmed the completion of an order. This means that in more general multitasking scenarios, the GP method is either a more unreliable interaction method, or it distracts the user enough to cause a negative performance impact on whatever they are doing primarily. We also found that the time until an order was accepted was lower with Touch than with PP and to an extent PS and GS, telling us that users using Touch were quicker to accept orders, either because they were multitasking more than in the other conditions, or because the general interaction was quicker. Participants using touch interaction were also 46 seconds faster than the 2nd fastest (GP, 318.5s) and 73 seconds than the slowest (GS, 346.7s) in overall trial time (Mean 320s), which supports this reasoning.

6.1 Design Implications

Based on our findings we want to give some design considerations for multimodal notification interactions in AR:

- Prioritize touch interaction for tasks requiring speed and efficiency, while allowing user customization of distance and size of UI elements.
- Give the user options to choose their preferred interaction technique while giving them a quick way to switch depending on context.
- Focus on reliability of individual modalities as well as the fusion of multiple to minimize errors.
- Provide safeguards to ensure successful action completion, such as briefly maintaining the indication of a target.

6.2 Limitations & Future Work

As participants expressed issues with the reliability of some of the interactions, a limitation of this experiment is, consequently, the systems we used. Even though we and probably many other developers used implementations provided by the manufacturer that were optimized for the device, results may still be impacted by the limitations of current AR systems. For instance, the HoloLens 2's limited camera angle may have impacted hand-tracking accuracy and required exaggerated hand movements. Although these device-specific constraints may have influenced the perceived reliability of some techniques vs. others, the TLX-score and the frustration sub-scale did not indicate differences between techniques. The observed trade-offs in multimodal interaction still show important usability considerations, which we have highlighted in our implications. For a satisfactory experience with multimodal interaction, latency must be minimized and perceived reliability prioritized. Future research on devices with improved capabilities could help isolate the impact of hardware limitations and validate the generalizability of these results.

Another limitation of our task was that notifications had only one action that the user could trigger. In operating systems like Windows or Android, which run on over 70% of all electronic devices connected to the internet [64], notifications have at minimum two interactions depending on program and version: One to dismiss the notification and one to open the corresponding program. However, our notifications only had one action, the dismissal or in case of the order-notifications (which could not be dismissed), the completion of the order. Future work should therefore look at notification interaction with more complex actions like quick-replies or image previews, which would increase the number of interaction targets on each notification. Having more targets could lead to differences in interaction technique performance.

Another possible future direction is to conduct an experiment where uni- and multimodal notification interaction techniques can be used at the same time, to get a better understanding of the context in which each interaction technique is used, since users might not want to always interact multimodally all the time [47]. Another future direction is to look at notification problems in-the-wild.

7 Conclusion

We explored different interaction techniques (four multi-modal: Gaze and Speech, Gaze and Pinch, Point and Pinch, and Point and Speech; one uni-modal: Touch) to interact with notifications in a virtual cooking environment. Our findings show that participants did not like using Gaze and Speech, but had no clear preferences between the other interactions. We also found differences in usability and physical demand, as well as worse task accuracy with the Gaze and Pinch method compared to the other techniques. Based on our findings, we recommend supporting multiple multimodal interaction techniques or allowing the user to choose their preferred interaction if the situation allows it.

Acknowledgments

This work was partially supported by the Office of Naval Research N00014-24-1-2214, Office of Naval Research N00014-21-1-2580, National Science Foundation 2327569, and National Science Foundation 2238313.

References

[1] U. Acer, A. Mashhadi, C. Forlivesi, and F. Kawsar. *Energy Efficient Scheduling for Mobile Push Notifications*, vol. 2. Aug. 2015. Journal Abbreviation: EAI Endorsed Transactions on Energy Web Publication Title: EAI Endorsed Transactions on Energy Web. doi: 10.4108/eai.22-7-2015.2260067

[2] H. Adams, J. Stefanucci, S. Creem-Regehr, and B. Bodenheimer. Depth Perception in Augmented Reality: The Effects of Display, Shadow, and Position. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 792–801, Mar. 2022. ISSN: 2642-5254. doi: 10.1109/VR51125.2022.00101

[3] alison@uxmag.com. Designing Notifications for Apps, Mar. 2020.

[4] Apple. Declaring your actionable notification types, 2023.

[5] Apple. Learn basic gestures and controls on Apple Vision Pro, 2024.

[6] R. T. Azuma. A Survey of augmented reality. *Presence: Teleoperators and Virtual Environments*, pp. 355–385, 1997.

[7] H. Beyer and K. Holtzblatt. Contextual design. *Interactions*, 6(1):32–42, Jan. 1999. doi: 10.1145/291224.291229

[8] R. A. Bolt. "Put-that-there": Voice and gesture at the graphics interface. *SIGGRAPH Comput. Graph.*, 14(3):262–270, July 1980. doi: 10.1145/965105.807503

[9] D. A. Bowman, E. Kruijff, J. J. LaViola, and I. Poupyrev. *3D User Interfaces: Theory and Practice*. Addison Wesley Longman Publishing Co., Inc., USA, June 2004.

[10] F. Business Insights. Virtual Reality [VR] Market Size, Growth, Share by 2032, 2024.

[11] R. Cai, N. N. P. K. Janaka, S. Zhao, and M. Sun. ParaGlassMenu: Towards Social-Friendly Subtle Interactions in Conversations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, pp. 1–21. Association for Computing Machinery, New York, NY, USA, Apr. 2023. doi: 10.1145/3544548.3581065

[12] P. Chojecki, D. Strazdas, D. Przewozny, N. Gard, D. Runde, N. Hoerner, A. Al-Hamadi, P. Eisert, and S. Bosse. Assessing the Value of Multimodal Interfaces: A Study on Human–Machine Interaction in Weld Inspection Workstations. *Sensors*, 23(11):5043, May 2023. doi: 10.3390/s23115043

[13] S. H. Chua, S. T. Perrault, D. J. C. Matthies, and S. Zhao. Positioning Glass: Investigating Display Positions of Monocular Optical See-Through Head-Mounted Display. In *Proceedings of the Fourth International Symposium on Chinese CHI - ChineseCHI2016*, pp. 1–6. ACM Press, San Jose, USA, 2016. doi: 10.1145/2948708.2948713

[14] E. Cutrell, M. Czerwinski, and E. Horvitz. Notification, Disruption, and Memory: Effects of Messaging Interruptions on Memory and Performance. 2001.

[15] M. Czerwinski, E. Cutrell, and E. Horvitz. Instant Messaging and Interruption: Influence of Task Type on Performance.

[16] S. Debernardis, M. Fiorentino, M. Gattullo, G. Monno, and A. Uva. Text Readability in Head-Worn Displays: Color and Style Optimization in Video vs. Optical See-Through Devices. *IEEE transactions on visualization and computer graphics*, 20:125–39, Jan. 2014. doi: 10.1109/TVCG.2013.86

[17] P. Emerson. The original Borda count and partial voting. *Social Choice and Welfare*, 40(2):353–358, Feb. 2013. doi: 10.1007/s00355-011-0603-9

[18] B. Fida, F. Cutolo, G. di Franco, M. Ferrari, and V. Ferrari. Augmented reality in open surgery. *Updates in Surgery*, 70(3):389–400, Sept. 2018. doi: 10.1007/s13304-018-0567-8

[19] S. Ghosh, L. Winston, N. Panchal, P. Kimura-Thollander, J. Hotnog, D. Cheong, G. Reyes, and G. D. Abowd. NotifiVR: Exploring Interruptions and Notifications in Virtual Reality. *IEEE Transactions on Visualization and Computer Graphics*, 24(4):1447–1456, Apr. 2018. Conference Name: IEEE Transactions on Visualization and Computer Graphics. doi: 10.1109/TVCG.2018.2793698

[20] R. A. Grier, A. Bangor, P. Kortum, and S. C. Peres. The System Usability Scale: Beyond Standard Usability Testing. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 57(1):187–191, Sept. 2013. Publisher: SAGE Publications Inc. doi: 10.1177/1541931213571042

[21] S. G. Hart. Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 50(9):904–908, Oct. 2006. Publisher: SAGE Publications Inc. doi: 10.1177/154193120605000909

[22] S. G. Hart and L. E. Staveland. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In P. A. Hancock and N. Meshkati, eds., *Human Mental Workload*, vol. 52 of *Advances in Psychology*, pp. 139–183. North-Holland, 1988. ISSN: 0166-4115. doi: 10.1016/S0166-4115(08)62386-9

[23] S. Imamov, D. Monzel, and W. S. Lages. Where to display? How Interface Position Affects Comfort and Task Switching Time on Glanceable Interfaces. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 851–858, Mar. 2020. ISSN: 2642-5254. doi: 10.1109/VR46266.2020.00110

[24] S. T. Iqbal and B. P. Bailey. Oasis: A framework for linking notification delivery to the perceptual structure of goal-directed tasks. *ACM Transactions on Computer-Human Interaction*, 17(4):1–28, Dec. 2010. doi: 10.1145/1879831.1879833

[25] D. Iso. 9241–210: 2010: ergonomics of human-system interaction—part 210: human-centred design for interactive systems (formerly known as 13407). *Switzerland: International Standards Organization*, 2010.

[26] N. Janaka, J. Gao, L. Zhu, S. Zhao, L. Lyu, P. Xu, M. Nabokow, S. Wang, and Y. Ong. GlassMessaging: Towards Ubiquitous Messaging Using OHMDs. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 7(3):1–32, Sept. 2023. doi: 10.1145/3610931

[27] N. Janaka, S. Zhao, and S. Sapkota. Can Icons Outperform Text? Understanding the Role of Pictograms in OHMD Notifications. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–23. ACM, Hamburg Germany, Apr. 2023. doi: 10.1145/3544548.3580891

[28] J. Jankowski, K. Samp, I. Irzynska, M. Jozwowicz, and S. Decker. Integrating Text with Video and 3D Graphics: The Effects of Text Drawing Styles on Text Readability. p. 10, 2010.

[29] J. A. Jones, J. E. Swan, G. Singh, E. Kolstad, and S. R. Ellis. The effects of virtual reality, augmented reality, and motion parallax on egocentric depth perception. In *Proceedings of the 5th symposium on Applied perception in graphics and visualization*, vol. 7, pp. 9–14. ACM, Los Angeles California, Aug. 2008. doi: 10.1145/1394281.1394283

[30] S. Kapp, M. Barz, S. Mukhametov, D. Sonntag, and J. Kuhn. ARETT: Augmented Reality Eye Tracking Toolkit for Head Mounted Displays. *Sensors*, 21(6):2234, Mar. 2021. doi: 10.3390/s21062234

[31] T. Kosch, A. Matviienko, F. Müller, J. Bersch, C. Katins, D. Schön, and M. Mühlhäuser. NotiBike: Assessing Target Selection Techniques for Cyclist Notifications in Augmented Reality. *Proceedings of the ACM on Human-Computer Interaction*, 6(MHCI):1–24, Sept. 2022. doi: 10.1145/3546732

[32] G. Kramida. Resolving the Vergence-Accommodation Conflict in Head-Mounted Displays. *IEEE Transactions on Visualization and Computer Graphics*, 22(7):1912–1931, July 2016. doi: 10.1109/TVCG.2015.2473855

[33] M. Krichenbauer, G. Yamamoto, T. Taketom, C. Sandor, and H. Kato. Augmented Reality versus Virtual Reality for 3D Object Manipulation. *IEEE Transactions on Visualization and Computer Graphics*, 24(2):1038–1048, Feb. 2018. doi: 10.1109/TVCG.2017.2658570

[34] A. Köpsel and N. Bubalo. Benefiting from legacy bias. *Interactions*, 22(5):44–47, Aug. 2015. doi: 10.1145/2803169

[35] M. J. Lazaro, S. Kim, J. Lee, J. Chun, and M.-H. Yun. Interaction Modalities for Notification Signals in Augmented Reality. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, pp. 470–477. ACM, Montréal QC Canada, Oct. 2021. doi: 10.1145/3462244.3479898

[36] H. Lee and W. Woo. Exploring the Effects of Augmented Reality Notification Type and Placement in AR HMD while Walking. In *2023 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pp. 519–529. IEEE, Shanghai, China, Mar. 2023. doi: 10.1109/VR55154.2023.00067

[37] H. Lee and W. Woo. Exploring the Effects of Augmented Reality Notification Type and Placement in AR HMD while Walking. In *2023 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pp. 519–529, Mar. 2023. ISSN: 2642-5254. doi: 10.1109/VR55154.2023.00067

[38] M. Lee, M. Billinghurst, W. Baek, R. Green, and W. Woo. A usability study of multimodal input in an augmented reality environment. *Virtual Reality*, 17(4):293–305, Nov. 2013. doi: 10.1007/s10055-013-0230-0

[39] U. Lee, J. Lee, M. Ko, C. Lee, Y. Kim, S. Yang, K. Yatani, G. Gweon, K.-M. Chung, and J. Song. Hooked on smartphones: an exploratory study on smartphone overuse among college students. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 2327–2336. ACM, Toronto Ontario Canada, Apr. 2014. doi: 10.1145/2556288.2557366

[40] F. Lu, S. Davari, L. Lisle, Y. Li, and D. A. Bowman. Glanceable AR: Evaluating Information Access Methods for Head-Worn Augmented Reality. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 930–939. IEEE, Atlanta, GA, USA, Mar. 2020. doi: 10.1109/VR46266.2020.00113

[41] A. Lucero and A. Vetek. NotifEye: using interactive glasses to deal with notifications while walking in public. In *Proceedings of the 11th Conference on Advances in Computer Entertainment Technology*, pp. 1–10. ACM, Funchal Portugal, Nov. 2014. doi: 10.1145/2663806.2663824

[42] J.-C. Martin. TYCOON: Theoretical Framework and Software Tools for Multimodal Interfaces. 1997.

[43] Meta. Getting started with Hand Tracking on Meta Quest headsets, 2024.

[44] Microsoft. Comfort - Mixed Reality | Microsoft Learn, Oct. 2021.

[45] Microsoft. Getting around HoloLens 2, Aug. 2021. GettingHoloLens2021.

[46] Microsoft. Releases · microsoft/MixedRealityToolkit-Unity, 2022.

[47] S. Oviatt. Ten myths of multimodal interaction. *Communications of the ACM*, 42(11):74–81, Nov. 1999. Publisher: Association for Computing Machinery (ACM). doi: 10.1145/319382.319398

[48] S. Oviatt, A. DeAngeli, and K. Kuhn. Integration and synchronization of input modes during multimodal human-computer interaction. In *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems*. ACM, Atlanta Georgia USA, Mar. 1997. doi: 10.1145/258549.258821

[49] T. Papadopoulos, K. Evangelidis, T. H. Kaskalis, G. Evangelidis, and S. Sylaiou. Interactions in Augmented and Mixed Reality: An Overview. *Applied Sciences*, 11(18):8752, Sept. 2021. doi: 10.3390/app11188752

[50] R. Pew Research Center. Americans' Use of Mobile Technology and Home Broadband, Jan. 2024.

[51] K. Pfeuffer, B. Mayer, D. Mardanbegi, and H. Gellersen. Gaze + pinch interaction in virtual reality. In *Proceedings of the 5th Symposium on Spatial User Interaction*, pp. 99–108. ACM, Brighton United Kingdom, Oct. 2017. doi: 10.1145/3131277.3132180

[52] M. Pielot, A. Vradi, and S. Park. Dismissed!: a detailed exploration of how mobile phone users handle push notifications. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services*, pp. 1–11. ACM, Barcelona Spain, Sept. 2018. doi: 10.1145/3229434.3229445

[53] L. Plabst, S. Oberdörfer, F. R. Ortega, and F. Niebling. Push the Red Button: Comparing Notification Placement with Augmented and Non-Augmented Tasks in AR. In *Proceedings of the 2022 ACM Symposium on Spatial User Interaction*, pp. 1–11. ACM, Online CA USA, Dec. 2022. doi: 10.1145/3565970.3567701

[54] L. Plabst, A. Raikwar, S. Oberdörfer, F. R. Ortega, and F. Niebling. Exploring Unimodal Notification Interaction and Display Methods in Augmented Reality. In *29th ACM Symposium on Virtual Reality Software and Technology*, pp. 1–11. ACM, Christchurch New Zealand, Oct. 2023. doi: 10.

1145/3611659.3615683
[55] C. Power, P. Cairns, and T. DeHaven. Mapping Virtual Reality Controls to Inform Design of Accessible User Experiences. In J. Abdelnour Nocera, M. Kristín Lárusdóttir, H. Petrie, A. Piccinno, and M. Winckler, eds., *Human-Computer Interaction – INTERACT 2023*, pp. 89–100. Springer Nature Switzerland, Cham, 2023. doi: 10.1007/978-3-031-42280-5_5
[56] K. Raaen and H. Sørum. Survey of interactions in popular vr experiences. In *Norsk Informatikkonferanse*, 2019.
[57] A. Raikwar, L. Plabst, and F. R. Ortega. ARtisan Bistro: A Cooking Task Environment to Conduct Studies in Augmented Reality. In *2022 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pp. 909–910, Oct. 2022. ISSN: 2771-1110. doi: 10.1109/ISMAR-Adjunct57072.2022.00200
[58] R. Rzayev, S. Korbely, M. Maul, A. Schark, V. Schwind, and N. Henze. Effects of Position and Alignment of Notifications on AR Glasses during Social Interaction. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*, pp. 1–11. ACM, Tallinn Estonia, Oct. 2020. doi: 10.1145/3419249.3420095
[59] R. Rzayev, S. Mayer, C. Krauter, and N. Henze. Notification in VR: The Effect of Notification Placement, Task and Environment. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*, pp. 199–211. ACM, Barcelona Spain, Oct. 2019. doi: 10.1145/3311350.3347190
[60] A. Sahami Shirazi, N. Henze, T. Dingler, M. Pielot, D. Weber, and A. Schmidt. Large-scale assessment of mobile notifications. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 3055–3064. ACM, Toronto Ontario Canada, Apr. 2014. doi: 10.1145/2556288.2557189
[61] A. Saktheeswaran, A. Srinivasan, and J. Stasko. Touch? Speech? or Touch and Speech? Investigating Multimodal Interaction for Visual Network Exploration and Analysis. *IEEE Transactions on Visualization and Computer Graphics*, 26(6):2168–2179, June 2020. doi: 10.1109/TVCG.2020.2970512
[62] M. Satkowski, R. Rzayev, E. Goebel, and R. Dachselt. ABOVE & BELOW: Investigating Ceiling and Floor for Augmented Reality Content Placement. In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 518–527, Oct. 2022. ISSN: 1554-7868. doi: 10.1109/ISMAR55827.2022.00068
[63] R. Skarbez, M. Smith, and M. C. Whitton. Revisiting Milgram and Kishino's Reality-Virtuality Continuum. *Frontiers in Virtual Reality*, 2:647997, Mar. 2021. doi: 10.3389/frvir.2021.647997
[64] S. G. Stats. Operating System Market Share Worldwide, 2024.
[65] C. Stothart, A. Mitchum, and C. Yehnert. The attentional cost of receiving a cell phone notification. *Journal of Experimental Psychology: Human Perception and Performance*, 41(4):893–897, Aug. 2015. doi: 10.1037/xhp0000100
[66] 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, Oct. 2021. Conference Name: IEEE Transactions on Human-Machine Systems. doi: 10.1109/THMS.2021.3097973
[67] A. S. Williams, J. Garcia, F. De Zayas, F. Hernandez, J. Sharp, and F. R. Ortega. The Cost of Production in Elicitation Studies and the Legacy Bias-Consensus Trade off. *Multimodal Technologies and Interaction*, 4(4):88, Dec. 2020. doi: 10.3390/mti4040088