GazePointAR: A Context-Aware Multimodal Voice Assistant for Pronoun Disambiguation in Wearable Augmented Reality

Jaewook Lee
University of Washington
Seattle, WA, USA

Jun Wang
University of Washington
Seattle, WA, USA

Elizabeth Brown
University of Washington
Seattle, WA, USA

Liam Chu
University of Washington
Seattle, WA, USA

Sebastian S. Rodriguez
University of Illinois at Urbana-Champaign
Urbana, IL, USA

Jon E. Froehlich
University of Washington
Seattle, WA, USA

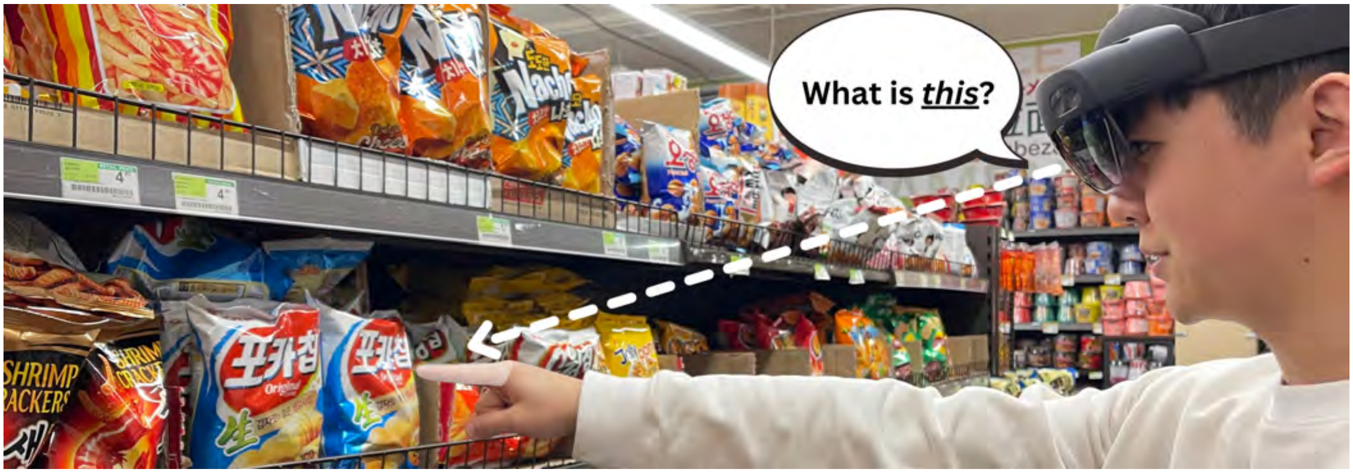


Figure 1: An example interaction with GazePointAR. The user's query *"What is this?"* is automatically resolved by using real-time gaze tracking, pointing gesture recognition, and computer vision to replace "*this*" with "*packaged item with text that says orion pocachip original*," which is then sent to a large language model for processing and the response read by a text-to-speech engine.

ABSTRACT

Voice assistants (VAs) like Siri and Alexa are transforming human-computer interaction; however, they lack awareness of users' spatiotemporal context, resulting in limited performance and unnatural dialogue. We introduce *GazePointAR*, a fully-functional context-aware VA for wearable augmented reality that leverages eye gaze, pointing gestures, and conversation history to disambiguate speech queries. With GazePointAR, users can ask "*what's over there?*" or "*how do I solve this math problem?*" simply by looking and/or pointing. We evaluated GazePointAR in a three-part lab study (N=12): (1) comparing GazePointAR to two commercial systems, (2) examining GazePointAR's pronoun disambiguation across three tasks; (3) and an open-ended phase where participants could suggest and try their own context-sensitive queries. Participants appreciated the naturalness and human-like nature of pronoun-driven queries, although sometimes pronoun use was counter-intuitive. We then iterated on GazePointAR and conducted a first-person diary study examining how GazePointAR performs in-the-wild. We conclude by enumerating limitations and design considerations for future context-aware VAs.

CCS CONCEPTS

• **Human-centered computing** → **Mixed / augmented reality**; **Interaction techniques**; **Natural language interfaces**.

KEYWORDS

augmented reality, multimodal input, voice assistants, gaze tracking, pointing gesture recognition, LLM

ACM Reference Format:
Jaewook Lee, Jun Wang, Elizabeth Brown, Liam Chu, Sebastian S. Rodriguez, and Jon E. Froehlich. 2024. GazePointAR: A Context-Aware Multimodal Voice Assistant for Pronoun Disambiguation in Wearable Augmented Reality. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA*. ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3613904.3642230

CHI '24, May 11–16, 2024, Honolulu, HI, USA
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0330-0/24/05.
https://doi.org/10.1145/3613904.3642230

1 INTRODUCTION

Voice assistants (VAs) are transforming human-computer interaction. In a recent study of 2,000+ people [65], 72% of respondents indicated that they use VAs for tasks such as playing music, setting timers, controlling IoT devices, and managing shopping lists [6, 8, 78]. While widespread and useful, state-of-the-art VAs like Amazon Alexa, Google Assistant, and Apple Siri do not yet consider a user's spatiotemporal context, which can result in unnatural dialogue or unanswerable queries [6]. For example, the query "*What is that?*" requires the VA to understand what "*that*" refers to—a problem known as pronoun disambiguation [18]. Despite their prominence in human speech [23], pronouns are not well supported by current VAs.

To resolve pronoun ambiguity, humans employ a variety of contextual clues, including eye gaze, pointing, and conversation history [23]. For example, a person may physically gesture at an item in a store and ask "*How much is this?*" While straightforward for a human to resolve, current VAs are unable to answer this query precisely because they lack spatiotemporal context. Pronoun disambiguation and multimodal input have a rich history of research in HCI [71, 84]—perhaps best marked by Bolt's visionary "*Put That There*" system in 1980 [9] and beyond [17, 44, 92]. With recent advances in machine learning, speech recognition, and large language models (LLMs), new approaches are now possible. For example, emerging context-aware VA prototypes such as *WorldGaze* [60], *Nimble* [83], and *TouchVA* [47] examine how to use head gaze, pointing, and touch to resolve ambiguous queries. While promising and informative to our own work, these prototypes share similar limitations: they use *Wizard-of-Oz* (WoZ) setups [19], are accompanied by tightly-controlled lab studies *vs.* open-ended queries, employ only one additional modality alongside speech, and are designed for smartphones rather than always-available head-worn displays.

In this paper, we introduce *GazePointAR*, a context-aware VA for wearable augmented reality (AR), which uses eye gaze, pointing gestures, and conversation history to support pronoun disambiguation. If a user's spoken query contains a pronoun, we process the user's field-of-view using real-time computer vision, automatically extract objects and written text in the scene, and generate a new coherent query phrase that is sent to OpenAI's GPT-3 [70] for processing. The response is then verbally read using speech synthesis. Pronouns are replaced using an empirically-tuned heuristic model that incorporates CV results based on gaze and pointing. For example, when asking "*How much is this?*" while looking at a bottle of mango juice (Figure 2), GazePointAR extracts information such as object type, brand name, and flavor name to generate "*How much is a bottle with text that says Naked Mighty Mango 290 Calories?*".

To evaluate GazePointAR and explore the potential of context-aware VAs in wearable AR, we conducted two studies. First, we performed a three-part qualitative laboratory study with 12 participants to compare GazePointAR to two state-of-the-art query systems (*i.e.,* Google Voice Assistant and Google Lens) (Part 1) and examine GazePointAR's usability and performance across various scenarios (Parts 2 & 3). For example, participants searched for the price difference between two salt boxes (*e.g.,* "*Can you compare the price between these two?*"). In Part 3, we invited participants to brainstorm and try their own queries to further assess how context-aware VAs may be used in the future and how well GazePointAR currently supports such uses. Participants primarily used gaze to ask a diverse range of queries, from retrieving object information to foreign language translation, and were impressed by GazePointAR's ability to include their gaze to resolve queries. Participants also noted limitations, such as only capturing gaze data once after a query is spoken, the inability to handle queries with multiple pronouns, lack of AI explainability, and object recognition errors.

Informed by these findings, we created a second GazePointAR prototype with improved object recognition and phrase generation techniques using prompt engineering, and conducted a follow-up first-person diary study [22]. Here, the first author used GazePointAR in their daily life for five days and recorded a written diary of usage, reflections, and observations of both successes and failures. In 20 hours of usage (4 hrs/day), the first author used GazePointAR across various contexts from cafes and restaurants to shopping malls and cinemas, and posed 48 queries, including recommendations for allergy-friendly menu items, ratings of movies, and cheaper alternatives to expensive clothing. Although the first author found GazePointAR to be more natural, instinctual, and robust against complex-to-describe objects in the real world than a traditional VA like Siri, they also encountered similar limitations as the study participants, such as static gaze data and limited object recognition capabilities, as well as privacy concerns with using a speech- and camera-based system in public.

In summary, our contributions include: (1) a fully-functional, context-aware VA for wearable AR that uses real-time computer vision and LLMs for pronoun disambiguation and more natural query dialogue; (2) findings from two user studies, including how users instinctively generate context-sensitive queries, how GazePointAR performs on queries from different scenarios, and limitations such as continuously tracking gaze information and AI explainability; and (3) a discussion on how to design future context-aware VAs that support any natural query a user poses spontaneously.

2 RELATED WORK

We provide background on pronoun usage in speech before enumerating relevant literature in multimodal interaction with a focus on voice assistants and augmented reality.

2.1 Pronoun Usage in Speech

Pronouns are frequently used in human speech, both in conversations between humans and in task-oriented dialogue systems—computational systems that complete tasks described in natural language. Leech *et al.* ranked the frequency of 100 million spoken English words showing that pronouns, including demonstrative pronouns (*e.g.,* "*this*," "*that*," "*these*," "*those*," "*here*," and "*there*") and third-person pronouns (*e.g.,* "*it*," "*he*," "*him*," "*she*," "*her*," "*they*," and "*them*") all ranked in the top 200 [48]. As further evidence, Byron and Allen annotated a corpus of task-oriented dialogues and found that over one-third of 1,068 dialogue turns contained referential occurrences of pronouns "*it*" and "*that*" [13]. Similarly, HCI studies have highlighted the importance of pronouns in human speech as they contribute to enhancing its naturalness and expressivity [9, 42, 47] and that users desire to communicate to

VAs using pronouns [34]. To resolve pronoun ambiguity, humans rely on multimodality such as looking at or pointing at referents while speaking and conversational context [23]. In our work, we investigate real-time gestures, eye gaze, and conversation history to enable pronoun disambiguation in human-VA interaction.

2.2 Multimodal Interaction

The HCI community has long been interested in multimodal interaction, highlighting various benefits such as improved naturalness, robustness, and expressiveness compared with unimodal interaction techniques [71, 84]. For instance, researchers explored gaze as a multimodal input technique in mobile devices to address shortcomings of touch, such as slow interaction speed, limited reach on large screens, and impreciseness on small screens [26, 28, 43, 58, 76]. Additionally, gestures and speech have often been combined with gaze to improve the accuracy of gaze-alone systems [15, 62]. In our work, we rely both on *gaze* and, if identified in the visual frame, *pointing gestures* to resolve speech ambiguities. Many consumer products now support multiple modes of input, which allow users to interact using both touch and speech. Although the field of multimodal input is vast [71, 84], for the purposes of this paper, we focus primarily on its use in voice assistants and augmented reality.

2.2.1 Multimodal Interaction with Voice Assistants. The integration of speech with additional input modalities has long been a topic of interest in HCI. For example, Bolt's foundational "*Put That There*" explored the use of speech and gestures as input [9]. Further research has expanded on this idea by examining other input modalities, such as gaze pointing [92], pen and voice interaction [17, 46], and merging speech, gestures, and eye gaze [44]. More recently, researchers have examined multimodal speech and gaze interactions in the context of hands-free communication between humans and vehicles [4, 64, 82], as well as speech and gestures to support natural interactions with virtual objects in AR [36, 50, 77]. Others have explored AR-based WoZ VA prototypes that support more natural dialogue between users and VAs by employing gaze [60], touch [47], or gestures [83] alongside speech. The importance of multimodality in the design of voice user interfaces is widely acknowledged [1, 23] because it enables flexible, expressive, natural, and contextual human-VA communication [9, 32, 42]. Our work aims to contribute to this literature by implementing and evaluating a fully-functional multimodal VA with ambiguous speech support.

2.2.2 Multimodal Interaction in Augmented Reality. In AR specifically, multimodal interaction is frequently employed to improve object selection and manipulation, typically using hand gestures, gaze, and/or voice [35, 89]. For instance, both Olwal *et al.* and Piumsomboon *et al.* used speech as a supplement to gesture for improved object selection in AR [67, 77]. Additionally, Kytö *et al.* used both head motion and eye gaze to increase the efficiency and accuracy of target selection in AR [45]. Furthermore, Lystbæk *et al.* used eye gaze to assist mid-air gestures with distant object selection in AR [57]. Lastly, Liao *et al.* used gestures and speech to generate and interact with AR presentation augmentations [50]. Similarly, GazePointAR employs hand gestures to support gaze with a goal of enhancing real-world object selection.

Most relevant to our work, recent research has explored multimodal interaction in AR for pronoun disambiguation. More specifically, when a multimodal VA receives an ambiguous query, such as "*When does this store open?*", AR is used to analyze various visual contexts, including objects, texts, gaze, and gestures. For instance, Mayer *et al.* presented WorldGaze, a WoZ smartphone-based multimodal VA that leverages head gaze information to clarify ambiguous queries [60]. Others have explored touch [47] and pointing gestures [83] to resolve ambiguity. Each modality has tradeoffs: head gaze is quick and hands-free but can be inaccurate [60], touch is accurate but slower and not hands-free [47], and gestures fall in between the two modalities [83]. In this work, we employ a combination of gaze supported by pointing gestures to create a efficient, mostly hands-free, and accurate input modality for speech disambiguation. We evaluate this in a fully-functional VA for wearable AR in various contexts.

2.2.3 Other Uses of Gaze, Pointing, and Speech in Wearable AR. We conclude by highlighting recent studies that, while not employing gaze, pointing gestures, and speech as multimodal interaction techniques, present novel applications for each of these input sources in wearable AR. For instance, researchers have used eye gaze to design AR interfaces that adaptively control the display of information based on context, including its timing, placement, and volume [52, 56, 75, 80]. Additionally, hand gestures are often classified using machine learning to enable more natural object and UI manipulation [72, 86, 91]. Furthermore, wearable AR glasses have been used to caption, translate, and augment speech in a non-intrusive way [29, 37–39, 53, 61, 66, 73, 80, 87]. GazePointAR, while multimodal, is influenced by this prior work in wearable AR for enhanced interaction and context.

3 GAZEPOINTAR PROTOTYPE 1

To advance the naturalness and economy of expression in how humans interact with VAs, we designed and built GazePointAR—a fully-functional context-aware VA for AR glasses that uses eye gaze, pointing gestures, and conversation history to support pronoun disambiguation. Below, we describe GazePointAR's design and implementation, starting with a taxonomy of pronoun usage drawn from linguistics literature.

3.1 Taxonomy of Pronoun Use and Resolution

To design GazePointAR, we first examined commonly-spoken pronouns in human speech and referent resolution strategies. We analyzed Leech *et al.*'s ranked frequency list of 100 million spoken English words [48] and filtered to pronouns spoken at least 500 times per one million words. From this process, we extracted thirteen pronouns across three distinct groups of pronouns, all of which GazePointAR supports: nominal demonstrative pronouns: "*this*," "*that*," "*these*," and "*those*", adverbial demonstrative pronouns: "*here*" and "*there*", and third person pronouns: "*it*," "*he*," "*him*," "*she*," "*her*," "*they*," and "*them*".

Demonstrative pronouns are used to point to specific people or things and can be further broken down into *nominal* and *adverbial* [25]. In human conversations, gaze and/or pointing gestures are often used for referent disambiguation [23]. While demonstrative pronouns such as "*this*" and "*that*," "*these*" and "*those*," and "*here*" and "*there*" seem similar, humans naturally employ one based on relative distance from the speaker to the referent [23]. For example, a person may ask "*How much is this?*" when referring to a nearby object and "*How much is that?*" if the object is further away.

For third-person pronouns, "*it*" may function as an *anaphoric*, which refers to a word used previously in a phrase such as "*I have a bicycle. It is red.*"; *pleonastic*, which is the use of more words than needed to express meaning either unintentionally or for emphasis such as "*kick it with your feet.*; or as an *event reference* such as "*He lost his job. It came as a total surprise.*" [54]. When resolving the anaphoric or pleonastic "*it*," humans need prior conversation history, while for event reference, "*it*" can be used interchangeably with "*this*" or "*that*" [33, 54]. For other third person pronouns, humans often refer to entities such as other people or animals with "*he*" or "*her*", for example, but these pronouns must be used cautiously, as they can introduce gender bias [14].

Grounded in this analysis, we designed a taxonomy of frequently-spoken pronouns and how ambiguity from each pronoun can be resolved. When implementing GazePointAR, we adhered closely to this taxonomy, enabling our system to handle all thirteen pronouns and determine their referents based on gaze, pointing gesture, and conversation history.

3.2 System Implementation

We designed and implemented GazePointAR for the Microsoft HoloLens 2 with Unity 2021.3.16f1[1] and Mixed Reality Toolkit (*MRTK*) 2.8.2[2]. While our overarching vision is to develop an always available context-aware VA for lightweight AR displays, the HoloLens 2—despite its bulk—allowed us to rapidly prototype an implementation.

We designed GazePointAR to resemble the user experience of a commercial VA such as Apple Siri or Amazon Alexa. GazePointAR waits for a user to say *"Hey Glass"* and make a verbal query. If the user's query contains one of thirteen pronouns in our taxonomy, it analyzes the user's field-of-view using various machine learning (ML) solutions, constructs a coherent phrase to describe the user's referent, replaces the pronoun with its referent, and sends the modified query to a large language model (OpenAI's GPT-3 [70]). The query response is vocalized to the user using a text-to-speech engine within 10 seconds. See the system diagram in Figure 2. We expand on key components below. As a rough examination of system response time, we asked the query "*How much is this?*" while gazing at a bottle of mango juice (a tutorial task) ten times. GazePointAR responded in 7.51 ± 0.45 seconds. We include sub-component performance times from this same procedure below.

Activating GazePointAR. To activate GazePointAR, the user states "*Hey Glass.*" For this, we implemented a continuously-running background process checking for the trigger phrase. Upon recognition, GazePointAR replies, "*Hi, I'm listening.*" and waits for a spoken query. After the query, GazePointAR performs a substring search to check for pronouns from our taxonomy.

Capturing and analyzing the user's field-of-view. If the query contains a pronoun, GazePointAR prompts the HoloLens to take a 1080p photo of the user's field-of-view. For user and bystander privacy, the captured image is stored temporarily and deleted once a query response is received. This process takes 2.27 ± 0.16 seconds to complete.

Once the user's field-of-view is captured, we begin analyzing the image for objects, texts, and faces. We send the captured image to three ML models through asynchronous POST requests to minimize runtime: *Google Cloud Vision*'s (1) *Object Localization* and (2) *Optical Character Recognition* (OCR) models [16], as well as (3) *Amazon Rekognition's Celebrity Recognition* model [7]. This process takes 3.37 ± 0.23 seconds to complete.

After receiving JSON responses from the ML services, GazePointAR identifies hierarchical relationships between the detected objects, faces, and texts. We treat the object detection and celebrity recognition results as the parent layer. The child layer, comprised of OCR results, is connected to parent bounding boxes that have at least 70% pixel overlap (a threshold tuned empirically). Each parent can have up to five OCR results, ranked by bounding box size. This ensures that GazePointAR prioritizes important textual information, such as product and brand names, which tend to be larger in the user's field-of-view, while ignoring less important, smaller details like promotional blurbs. For example, as shown in Figure 2, when a user asks "*How much is this?*" while holding a bottle of Naked Mighty Mango juice, possible parent layer objects include "*person*" and "*bottle*", with "*bottle*" having child layer objects such as "*Naked*", "*Mighty*", "*Mango*", "*290*", and "*calories*".

Gaze tracking and gesture recognition. To capture the user's eye gaze and pointing gesture, we customized MRTK's built-in gaze and pointer modules. For gaze, we designed a white sphere that follows the user's gaze from a fixed distance (*i.e.,* 2 meters) and is overlaid in their field-of-view. This allows us to retrieve 3D gaze coordinate data and also provides visual feedback to the user about their system-inferred gaze.

For pointing, we implemented a finger-pointing gesture to supplement the base palm-pointing gesture, since extending the arm and index finger is a more typical pointing gesture [23]. Performing a pointing gesture creates a ray that extends away from the user's hand until a collision with an object in the physical world occurs. To achieve this, we integrated MRTK's spatial awareness into GazePointAR to detect collisions between user inputs and spatial meshes generated in real-time.

As the HoloLens captures an image, GazePointAR simultaneously logs the locations of both the user's gaze and pointing gesture. To convert 3D gaze and pointing gesture coordinates to their corresponding pixel locations on the captured image, we use projection.

Query assembly and pronoun replacement. Using the ML-generated results and pixel coordinates of gaze and pointing gesture, GazePointAR assembles a coherent phrase to replace the user-spoken pronoun. To accomplish this, we employ a state diagram, which encompasses the differences in pronouns in our taxonomy.

If a pronoun is singular, GazePointAR computes whether any input coordinates fall within any parent (*i.e.,* object recognition and celebrity recognition results) bounding boxes. If so, GazePointAR takes that parent object's child layer (*i.e.,* OCR results) and creates the following phrase: "*[parent] with text that says [children]*". Otherwise, GazePointAR takes the five nearest child layer texts from each input coordinate, computes a union, orders them by distance, and uses the five closest to build the following phrase: "*[OCR Result*

[1] https://unity.com

[2] https://learn.microsoft.com/en-us/windows/mixed-reality/mrtk-unity/mrtk2

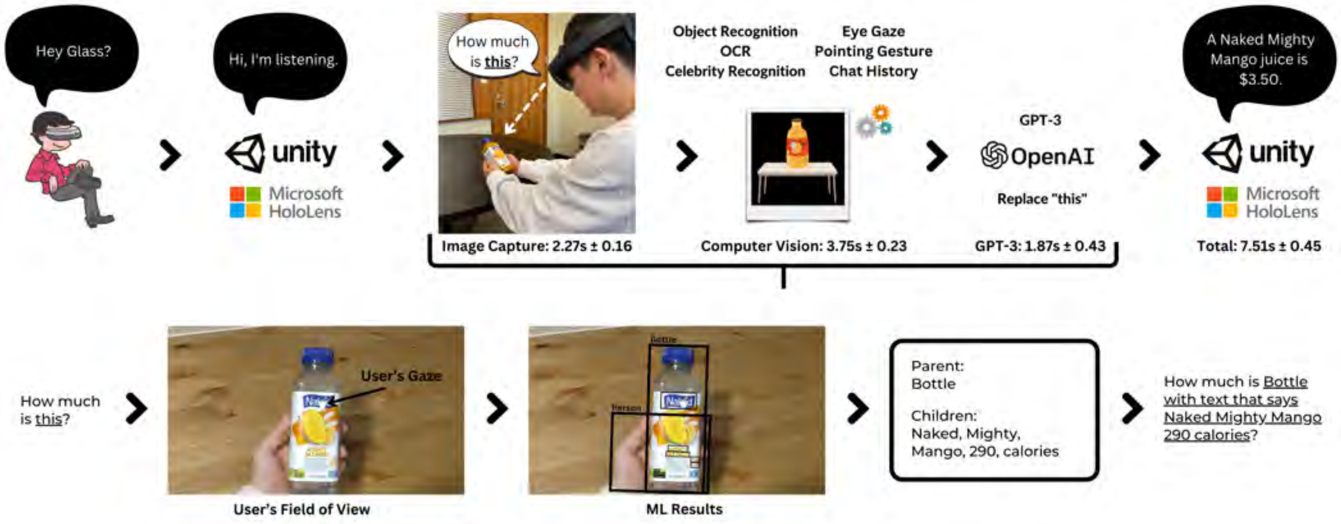


Figure 2: System overview and implementation details of GazePointAR

1] [OCR Result 2] ... [OCR Result 5]." For example, returning to the price of a Naked Mighty Mango juice in Figure 2, a user is looking at the "*Bottle*", meaning GazePointAR generates the phrase "*Bottle with text that says Naked Mighty Mango 290 calories*".

If the pronoun is plural, GazePointAR expands the gaze and pointing gesture pixel coordinates into bounding boxes with width and height equivalent to half of the captured image's width and height. Then, GazePointAR computes whether any input bounding boxes have at least 70% overlap with any parent bounding boxes. The rest of the procedure is the same as with singular pronouns.

Answering the query. GazePointAR assembles the final query by combining the user-spoken query, the ML-generated phrase, and text from the five most recent query-answer pairs. The final result is processed by OpenAI's GPT-3 [70], which takes 1.87 ± 0.43 seconds to complete. The output is displayed as text and read aloud. If there are no ML results or GPT-3 cannot process the modified query, GazePointAR responds "*Sorry, I did not understand your question.*" Users can ask follow-up questions or provide additional information appropriately.

4 STUDY 1: THREE-PART LAB EVALUATION OF GAZEPOINTAR

To evaluate GazePointAR and explore the potential of context-aware VAs in wearable AR, we conducted two studies: (1) a laboratory study to compare GazePointAR to two state-of-the-art query systems and examine how participants generate and use their own context-sensitive queries; and (2) a first-person diary study using GazePointAR in the real world. We report on the first study below.

For the lab study, we sought to address three primary research questions: (1) How do users initially perceive and use a multimodal, context-aware VA for pronoun disambiguation? (2) How does performance compare to traditional VAs? (3) What types of queries do users want to perform with a context-aware VA, and how well does GazePointAR support these queries? As initial work, our primary aim was not to quantitatively examine GazePointAR's performance but rather to observe how participants reacted to and used a fully-functional, context-aware query system for AR glasses.

To address these questions, we conducted a three-part, within-subjects laboratory study with 12 participants. In Part 1, we asked participants to complete a common query task with GazePointAR as well as two state-of-the-art commercial systems: *Google Voice Assistant* (*voice input*) and *Google Lens* (*image+text input*). In Part 2, participants completed three additional context-dependent query tasks with GazePointAR, which were designed to highlight different aspects in our design space (*e.g.,* pronoun use, gaze, gesture, and conversation history). Finally, in Part 3, participants brainstormed and tried their own context-sensitive queries.

4.1 Participants

We recruited 12 participants via mailing lists, social media, and snowball sampling. Participants were screened via a demographic questionnaire, which asked about prior experiences with VAs, AR, and AI chat systems. Given the reliance on gaze and speech in our study, we filtered participants who indicated visual or auditory disabilities, have a history of seizures or epilepsy, or are not fluent in English. All twelve participants indicated at least some previous experience with VAs, including *Amazon Alexa*, *Apple Siri*, *Google Voice Assistant*, *Microsoft Cortana*, and *Samsung Bixby*. Most (9/12) had not previously used AR headsets or glasses—those that did (3/12) mentioned Google Glass, Microsoft HoloLens, and Meta Quest Pro. Finally, all participants indicated at least some familiarity with AI chat systems with six stating that they use them at least once a week (two participants marked never). Most commonly, participants mentioned ChatGPT (8/12) and customer support chatbots (2/12).

4.2 Procedure

The in-person laboratory study took place on a university campus and lasted 60 minutes. Instructions were presented orally with backing slides to improve comprehension. Consent and background forms were emailed in advance; written consent was taken in person. All sessions were video recorded for *post hoc* analysis. Because we were interested in candid reactions, we did not tell participants that we created GazePointAR.

Tutorial. After consenting, participants completed a short tutorial about each VA system: GazePointAR, Google VA, and Google Lens. The tutorial order was counterbalanced but the query task was the same: "*Your task is to find the price of this bottle of Naked Mighty Mango juice*" (Figure 2). During the tutorial, participants could ask questions of the study facilitator and, for GazePointAR, configure the AR headset fit and calibration. The study commenced once each participant was comfortable with all three VA systems.

Part 1: Comparing VAs. Part 1's goal was to examine how participants constructed queries for a common VA scenario: cooking. Specifically, we asked participants to "*find a marinara pasta recipe that uses this jar of Rao's Marinara sauce; the more specific, the better*" (Figure 3) using each of the VA systems—which were again counterbalanced. For each VA system, we encouraged participants to construct the query to best leverage the system's input modality (*e.g.,* taking a picture for Google Lens, gazing or pointing for GazePointAR). The search task was deemed complete when the participant had found, from their perspective, a satisfactory recipe. After using each VA, participants filled out a *System Usability Scale* (SUS) [12, 74] questionnaire and answered interview questions regarding their experience. At the end of Part 1, we asked participants to rank the three systems in terms of perceived intelligence, helpfulness, naturalness, and overall preference. We then asked follow-up questions to justify rankings.

Part 2: Context-sensitive Queries with GazePointAR. While Part 1 examined differences in query behavior depending on modality and technology, Part 2 specifically focused on examining context-sensitive queries with GazePointAR. We asked participants to complete three tasks that, based on our own usage of GazePointAR, benefited from context-dependent queries and pronoun disambiguation: (1) *Write a simple math equation on a sheet of paper and ask GazePointAR if it is mathematically accurate*; (2) *Use GazePointAR to find the cost difference between two items*; (3) *Use GazePointAR to find more information about a person in a magazine article* (Figure 4). Again, at the end of Part 2, we asked participants to remark on their GazePointAR experiences and the additional search tasks.

Part 3: Design Probe and Co-design. Finally, in Part 3, participants helped co-design the future of context-aware VA systems. Using a design probe method similar to Mauriello *et al.* [59], participants first watched five video clips of GazePointAR being used across diverse scenarios: cooking, math, language translation, recycling materials, and asking if there are dangerous items nearby (Figure 5). After viewing and discussing the design probe videos, participants brainstormed and then actually attempted their own context-sensitive queries—a study task that is only possible with a fully-functional prototype like GazePointAR.

4.3 Data and Analysis

We analyzed three sources of data: interview transcripts, observations from the user study sessions, and the post-task questionnaires. For the qualitative data, we used reflexive thematic coding [10, 11]. The first author, who facilitated all user study sessions, created an initial codebook by reviewing study transcripts. The entire team then collaboratively iterated on the codebook while checking for bias and coverage. With a final codebook consisting of 34 codes, the first author coded participants' quotes, after which the team discussed the resulting themes. While this exploratory study focused on participants' reactions to GazePointAR, we also collected quantitative data from Part 1 to compare GazePointAR with existing systems. For SUS scores, we converted survey responses, which are on a scale of 0-40 when summed, to a range between 0-100[3] [12]. We then conducted a Friedman test as an omnibus test with an appropriate number of Wilcoxon signed-rank tests corrected with Holm's sequential Bonferroni procedure for statistical significance. See Figure 6 for a summary of quantitative results.

4.4 Findings

We report key findings, including how VA input modality influenced perceived performance and query formation, the queries participants generated using GazePointAR, and successes and failures of GazePointAR in various scenarios. We denote each participant as P# (*e.g.,* P1 for participant 1). Quotes have been lightly modified for concision and clarity.

4.4.1 Part 1: Comparing VAs.

In Part 1, participants completed an open-ended query task to find a recipe for a specific marinara sauce with the three different VA systems (Figure 3). We first provide overall reactions before analyzing query formations, perceived intelligence, naturalness, and helpfulness, task completion time, and usability.

Overall. Overall, participants preferred using Google VA ($mean_{rank}$ =1.7; *SD*=0.7) and GazePointAR (1.8; *SD*=0.9) over Google Lens (2.6; *SD*=0.7)—lower is better, range is 1-3. For Google Lens, participants emphasized that while taking photos was familiar (3/12) and lessened the specificity of their queries compared to voice-only systems (3/12), manually capturing an image and supplying written text felt tedious (3/12) and unnatural (2/12). As P2 stated, "*I had to take a picture and then add more information... It's like an extra step, right? Is this necessary?*". Similarly, P4 said, "*Google Lens is the most unnatural, because sometimes you have to type extra context, and I feel like that's just another hurdle.*". Finally, the quality of Google Lens' responses influenced opinions: four participants were initially guided to a recipe for *making* marinara sauce rather than *using* Rao's Marinara sauce. Two participants mentioned losing confidence in Google Lens due to poor responses.

For Google VA, participants appreciated the straightforward (6/12), quick (4/12), and hands-free (2/12) nature of the system. Additionally, four participants emphasized that, compared to Google Lens and GazePointAR, it was easier to review query responses, visit different links, and decide on the best answer themselves. As P5 said, "*Google voice assistant displayed a typical Google search result [on the phone], which gives me a lot of options... clicking into them allows you to try until you find the recipe that you're satisfied with.*" Half of the participants also mentioned the familiarity of Google VA and the results interface. For limitations, participants noted that Google VA requires queries to be highly specific (4/12),

[3] $(((Q1 + Q3 + Q5 + Q7 + Q9) - 5) + (25 - (Q2 + Q4 + Q6 + Q8 + Q10))) * 2.5$

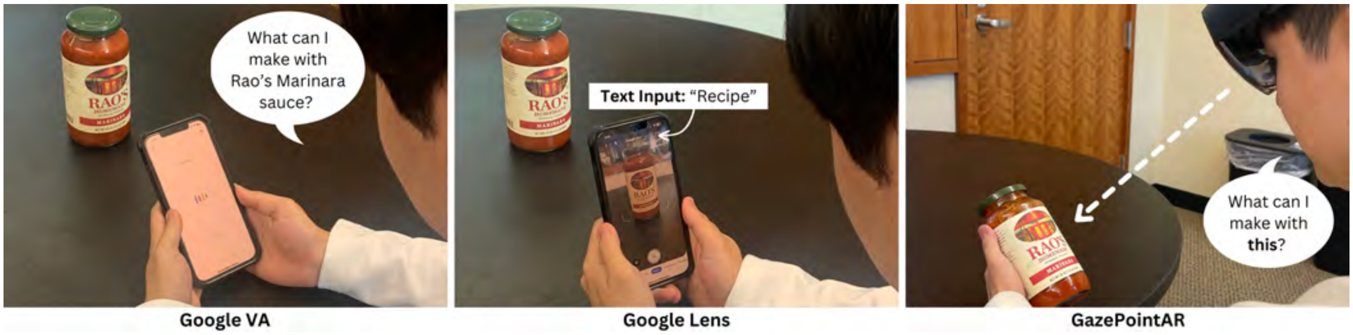

What can I make with Rao's Marinara sauce?
Text Input: "Recipe"
What can I make with **this**?
Google VA
Google Lens
GazePointAR

Figure 3: Cooking scenario and the three VAs used in Part 1 of the study.

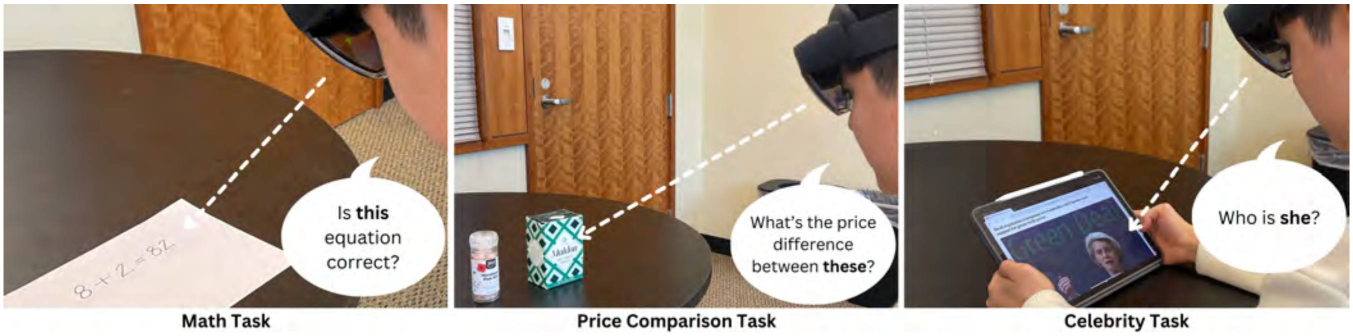

8 + 2 = 82
Is **this** equation correct?
What's the price difference between **these**?
Who is **she**?
Math Task
Price Comparison Task
Celebrity Task

Figure 4: Usage scenarios in Part 2 of the study.

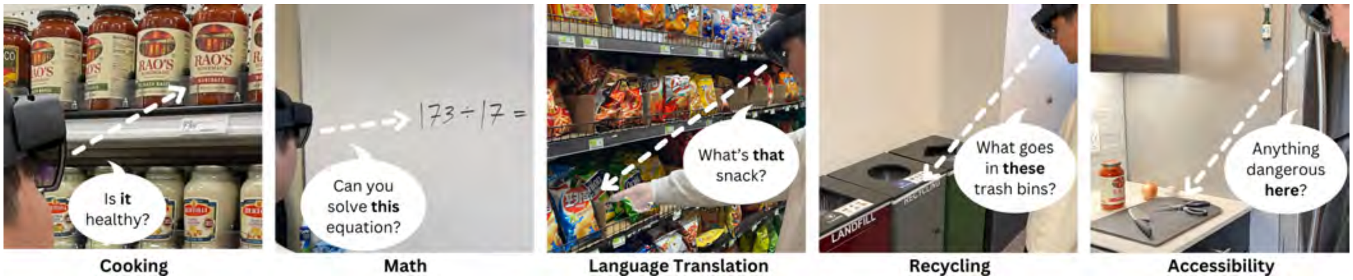
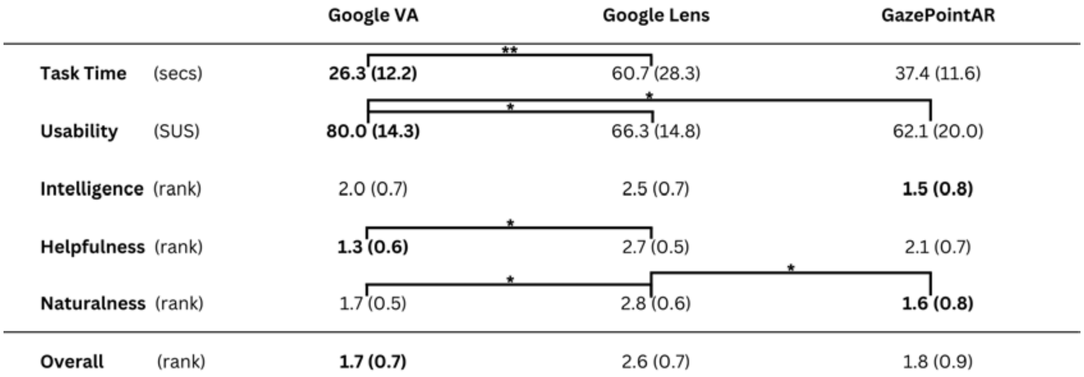


Figure 5: Design probes in Part 3 of the study. See supplementary materials for the videos.

necessitates accurate pronunciation of complex words like "*Rao's*" (4/12), and leads to longer queries, which are laborious to say (3/12). P12 aptly summarized theses issues by stating, "*You have to be more specific and have to say a lot more... I also think that a lot of people might mispronounce Rao's.*" One participant (P3) felt strongly about voice-edit capabilities—as Google VA only allows query iteration through text but not voice-based editing.

Finally, for GazePointAR, participants felt that it was simpler (8/12) and faster (8/12) to interact with as well as more natural (7/12) and human-like (6/12) to speak to than Google VA and Google Lens. In part, this was because participants could reduce the specificity of their queries with GazePointAR's context-awareness features. As P10 said, "*When speaking to GazePointAR, I am giving it a voice input while also interacting with the product that I am talking about. Perceptually, this is the most natural way of speaking, which is why we do this when talking to other people as well.*". Another said: "*When you're talking to someone, you point to or look at something and say 'what is this?' They can see what you're pointing to or looking at, which is exactly what the headset is doing... I was also able to receive an answer quickly without having to look through web pages.*" (P4). However, the most common criticism (8/12) was that GazePointAR provided only a single answer rather than an interactive, explorable list like a traditional search engine. Participants also requested more transparency from the system about their gaze and pointing gestures, the image GazePointAR took for scene processing, desired citations in the query response, and wanted queries to be editable.

Query formations. Beyond overall reactions, we also explored *how* participants formed queries with the three systems. When examining query length, unsurprisingly, the two multimodal systems had shorter queries on average: Google Lens (*avg*=1.3 words



		Google VA	Google Lens	GazePointAR
Task Time	(secs)	**26.3 (12.2)**	60.7 (28.3)	37.4 (11.6)
Usability	(SUS)	**80.0 (14.3)**	66.3 (14.8)	62.1 (20.0)
Intelligence	(rank)	2.0 (0.7)	2.5 (0.7)	**1.5 (0.8)**
Helpfulness	(rank)	**1.3 (0.6)**	2.7 (0.5)	2.1 (0.7)
Naturalness	(rank)	1.7 (0.5)	2.8 (0.6)	**1.6 (0.8)**
Overall	(rank)	**1.7 (0.7)**	2.6 (0.7)	1.8 (0.9)

Figure 6: The mean and standard deviation of task time, usability, perceived intelligence, helpfulness, naturalness, and overall preference. Task Time is in seconds. Usability is 0-100; higher the better. Rankings are 1-3; lower is better. For statistical significance, one asterisk (*) is $p < 0.05$; two asterisks (**) is $p < 0.01$.

long; *SD*=0.5) and GazePointAR (*avg*=6.3; *SD*=1.8) than Google VA (*avg*=8.4; *SD*=2.2). With Google Lens, all participants took a picture of the sauce jar then supplied additional text, including "*recipe*" (9/12) and "*recipe using*" (3/12). With GazePointAR, all participants used the pronoun "*this*" along with gaze but did not use pointing. P2 reasoned that "*If you're pointing at something, you have to use your hand. This implies that you still have use of your hands during some tasks. Also, because the jar is so close, the system shouldn't need pointing to tell what I'm talking about.*" Finally, with Google VA, all participants used proper nouns, including various formations of "*Rao's homemade Marinara sauce*". Full queries are in Appendix 1.

Perceived intelligence, helpfulness, and naturalness. For perceived intelligence, participants ranked GazePointAR the highest with $mean_{rank}$*=1.5* (*SD*=0.8), followed by Google VA (2.0; *SD*=0.7), then Google Lens (2.5; *SD*=0.7). A majority of participants (8/12) reasoned that GazePointAR "*recognized things I am talking about just from my gaze and pointing*" (P3), while for Google VA and Google Lens, "*instead of it figuring things out itself, I have to provide everything*" (P12). For perceived helpfulness, participants ranked Google VA the highest with $mean_{rank}$*=1.3* (*SD*=0.6), followed by GazePointAR (2.1; *SD*=0.7), then Google Lens last (2.7; *SD*=0.5). Half of the participants reasoned that Google VA displays multiple options and images in a familiar UI, which helped them decide on a satisfactory answer.

For perceived naturalness, participants ranked both GazePointAR and Google VA highly with $mean_{rank}$=1.6 (*SD*=0.8) and 1.7 (*SD*=0.5) respectively, followed by Google Lens (2.8; *SD*=0.6). Participants generally equated naturalness to the ease with which the query was constructed (10/12). As P12 said, "*I wish I can say queries with and without pronouns, because whichever comes to mind first, that's the one I want to say.*" Given the simplicity of the search task, P5, P11, and P12 indicated that the high specificity demanded by Google VA is not much of a concern; however, as search queries become more complex, Google VA can quickly fall behind other systems. As one example, three participants were unsure how to pronounce "*Rao's*" so felt more comfortable saying "*this*". While seven participants felt GazePointAR was most natural, P12 emphasized that humans are conversationally adaptable and have learned how to speak to modern VAs: "*GazePointAR was definitely the most human-like if we mean most 'natural' and 'human-like' in terms of speaking to another person; however, if we say 'natural' as in speaking to a machine, then Google Voice Assistant wins*".

Task completion time. While we allowed participants to define their own stoppage mark for determining a satisfactory query answer, task time is still an interesting metric and central to information retrieval [30]. On average, the fastest completion was Google VA (*avg*=26.3 secs; *SD*=12.2) followed by GazePointAR (37.4 secs; *SD*=11.6) then Google Lens (60.7 secs; *SD*=28.3). For both Google VA and Google Lens, participants primarily spent time clicking and viewing links to find a satisfactory recipe while with GazePointAR, participants received a direct answer but were delayed by query and image processing. To form the query, Google Lens took the longest as participants had to input both an image and textual content; for both Google VA and GazePointAR, participants could form queries hands-free, which increased interaction speed.

System usability. Finally, for the SUS questionnaire, participants gave Google VA a higher usability score (*avg*=80.0; *SD*=14.3) than Google Lens (66.3; *SD*=14.8) and GazePointAR (62.1; *SD*=20.0)—higher is better, range is 0-100. Various factors influenced usability, including familiarity with Google suite, autonomy in choosing a satisfactory answer from Google UI, naturalness in coming up with and vocalizing queries, and task completion time.

4.4.2 Part 2: Context-sensitive Queries.

While Part 1 explored differences between VA systems, Part 2 focuses specifically on GazePointAR and three context-sensitive queries: solving a math equation, comparing costs between items, and finding information about a celebrity (Figure 4). We did not guide participants in how to complete the queries, so our findings are based on participants' initial instincts. For all tasks, participants chose to use gaze+speech rather than pointing as participants felt that pointing was unnecessary (7/12) and like extra work (6/12). In a few instances, participants relied on conversation history; for example, P1 asked "*How much do these cost?*", then, after receiving the prices of two items, they asked “*What's the cost difference?*”. Below, we report on participants' query formations and their overall reactions across tasks.

Solving a math equation. Interestingly, all participants constructed this query similarly: using the pronoun “*this*”, which felt most natural (9/12). As P10 said, “*the equation I wrote is right there, but I don't want to say the whole thing out loud... being able to just look and say 'this' and have it read the equation is pretty useful.*” All but one participant preferred using a context-sensitive query and pronouns compared to vocalizing the whole equation. Some participants (5/12) mentioned feeling unsure where to look to properly capture the equation during their query: “*having to keep my gaze on the equation is more difficult than a jar, since I know I have to fix my gaze, but I am not sure where I should look*” (P3).

Comparing costs between two items. Unlike the math equation, participants constructed this query using two different pronouns: ten used the pronoun “*these*” and two used “*them*”. Currently, GazePointAR only supports one pronoun per query. Five participants felt that constructing a comparative query with multiple pronouns would have felt more natural such as, “*Compare the cost of this to that.*” As P1 stated, “*when there are exactly two objects, I feel like I will more likely say 'this or that' rather than 'these'*”. Similar to the math task, participants were unsure where to look to communicate intent (*i.e.,* multiple object referents) with GazePointAR. Participants also reiterated wanting more system transparency to understand what GazePointAR was capturing for the context-sensitive query: “*It is impressive that it can figure out multiple objects, but it will likely be more incorrect when trying to guess multiple objects I am talking about, so I really want to know what it thought I meant*” (P5).

Finding information about a celebrity. For this task, the query construction was most varied: five used the pronoun “*this*' (*e.g., “Who is this?”*), four “*her*” (*e.g., “Tell me about her.”*), and three “*she*” (*e.g., “Who is she?”*. Seven participants specifically mentioned how helpful pronouns were with this task: “*if you are looking at something you don't know, like a photo of a person, the only way to ask a question is by saying 'who is she' or 'who is he'*” (P11).

4.4.3 Part 3: Design Probe and Co-design.

Finally, for Part 3, we showed five video clips of GazePointAR and then invited participants to co-brainstorm and try their own context-sensitive queries (Figure 5). Below, we first report on reactions to our design probe and then describe participant-generated queries and how well GazePointAR performed.

Reactions to design probes. Overall, participants believed that GazePointAR has many uses, as many referents are difficult to describe in words. As P10 said: “*although I use voice assistants almost every day to play music or something, I now realize that many things I look at are difficult to clearly describe in text... since with this people can now input their environment easily, I think it will make speaking to voice assistants easier in many everyday activities.*” P3 was surprised with the range of supported queries. Additionally, participants expressed a particular interest in the societal impact examples, such as the hazardous object clip (7/12), which shows an accessibility example where a user is asking “*Anything dangerous here?*” while looking ahead, and the recycling clip (4/12), which shows a user asking “*What goes in these trash bins?*”. After viewing the hazardous object probe, P5 said, “*all you have to do is use pronouns and it can process objects in a person's field-of-view... that's great for blind people, which I really like.*” Participants summarized that when a visual referent is either unknown or difficult to vocalize, pronouns become especially useful.

Brainstorming and trying queries. For the co-design task, participants generated a total of 32 queries—see Appendix 3—and used gaze (32/32), pointing (6/32), and conversation history (1/32). Queries in which participants used pointing gestures had pronouns “that” (4/6), “there” (1/6), and “they” (1/6), which were all referring to objects faraway from the user. Conversation history was used when asking follow-up questions to find more information about a celebrity. Most queries (23/32) were aimed at deriving information about an object or person, including an object's name and price, a location's distance, and a person's name and accomplishments. Other queries included foreign language translation (4/32) like “*How do you pronounce this?*”, object comparison (3/32), and to confirm the correctness of a user's action (*e.g.,* “*Can I put this [trash] in here [recycling trash bin]?*”) (2/32). In analyzing pronoun usage, participants most commonly used “*this*” (16 occurrences), followed by “*that*” (8), “*s/he*” or “*him/her*” (5), and “*they*” (1).

GazePointAR provided a satisfactory answer for 13 of the 32 queries, including “*Who is s/he [person]? What is his/hers [musician] top hit?*” and “*What's happening over there?*”. Many of the unanswerable queries were due to lack of information, such as limitations in object recognition (*e.g.,* while a object localization model can recognize a car, it does not know the make and model of the car) and missing access to information online (*e.g.,* a price of an item may vary and GPT does not have access to store-specific information).

Other unanswerable queries were due to GazePointAR's inability to handle multiple pronouns in a single query (*e.g.,* “*Tell me the price difference between this and that.*”) or past referents (*e.g.,* “*Who was s/he again?*”). Participants suggested that GazePointAR should capture gaze over time. P3 added that this will remove the need for dwelling on a referent, which will allow users to gaze more naturally and improve the system's overall usability. While P10 was in favor of this feature, they also expressed privacy concerns. P5 went even further and said GazePointAR should record objects nearby gaze to support scenarios where gaze target is not the object in question (*e.g.,* “*What is the object next to that chair*”).

4.5 Study 1 Summary

Participants appreciated GazePointAR for its simplicity, naturalness, and human-likeness. When using GazePointAR, participants mostly relied on gaze to keep the interaction hands-free and efficient, while occasionally using pointing gestures and conversation history. Participants preferred to speak pronouns, especially when referents had difficult-to-pronounce, long, or unknown names. In some cases, including pronouns in a query felt less natural (*e.g.,* “*What can I make with this?*” vs. “*What can I make with Rao's Marinara sauce?*”). In terms of limitations, we found that GazePointAR should support multiple pronouns, provide more answer options and explanations when answering queries, use more robust ML models, and that users

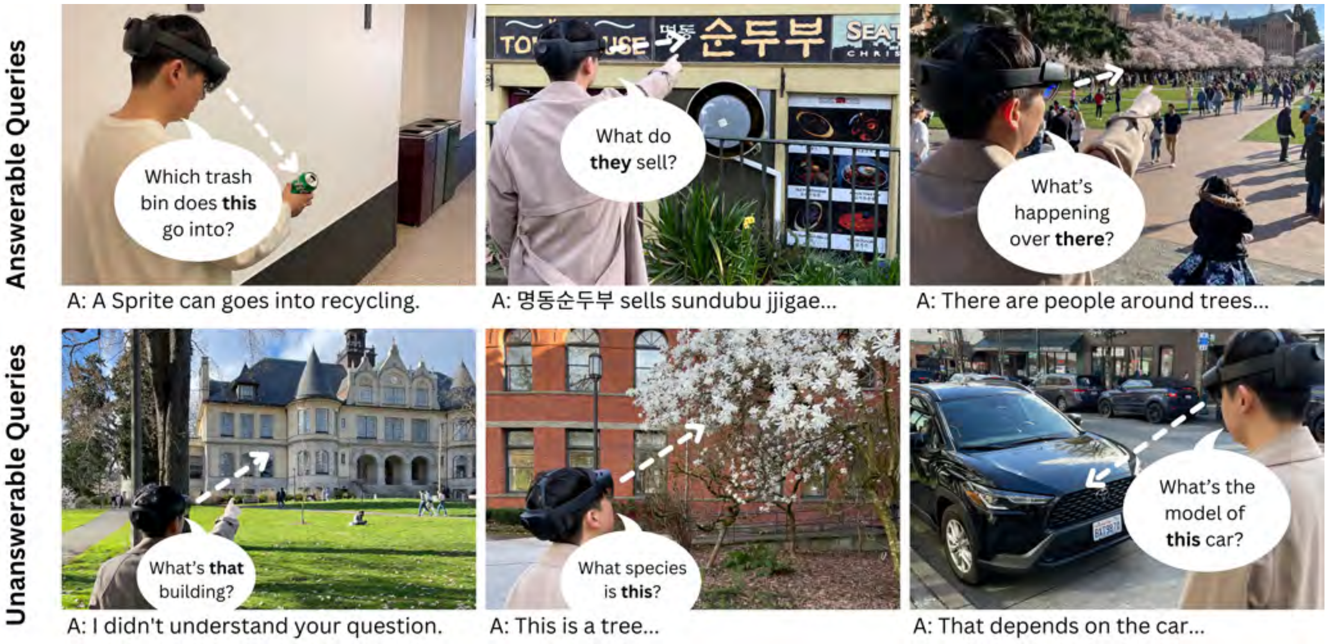

Answerable Queries
Which trash bin does **this** go into?
A: A Sprite can goes into recycling.
What do **they** sell?
A: 명동순두부 sells sundubu jjigae...
What's happening over **there**?
A: There are people around trees...
Unanswerable Queries
What's **that** building?
A: I didn't understand your question.
What species is **this**?
A: This is a tree...
What's the model of **this** car?
A: That depends on the car...

Figure 7: A subset of the scenarios participants came up with during Part 3 of Study 1. The top row shows recreations of answerable queries while the bottom rows highlights example queries that returned unsatisfactory responses.

could tire due to explicit gazing. Participants suggested several features for improvement: capturing gaze information over time, communicating to the user about captured images, gaze, pointing, and citations used in deriving answers, and displaying an explorable search result similar to Google.

5 GAZEPOINTAR PROTOTYPE 2

Informed by Study 1 findings and our own experiences using GazePointAR, we created a second GazePointAR prototype with three advancements: first, we replaced Google Cloud Vision's Object Localization model with *YOLOv8* [41]; second, we redesigned the multimodal contextual phrase generator using prompt engineering [79]; third, and finally, we updated the chat completions API to leverage GPT-3.5. We describe these advancements below and then discuss our five-day first-person diary study using GazePointAR version 2 "*in the wild*" [81].

Updating GazePointAR's object recognizer. For the initial GazePointAR prototype, we chose Google Cloud Vision's Object Localization model, as it enabled rapid prototyping. However, a key limitation of this model is that it categorizes an object as "*packaged goods*" if it cannot precisely identify the object, which confused both GPT-3 and our participants. In this iteration of GazePointAR, we instead employed a state-of-the-art YOLOv8 model trained on the *MS COCO* dataset [51] by building a local API server using *FastAPI*[4] and *Docker*[5], and tunneling the local API using *Localtunnel*[6]. This increased the ML services' runtime to 3.75 ± 0.31 seconds (+11.28%) and the overall runtime to 7.94 ± 0.38 seconds (+5.73%).

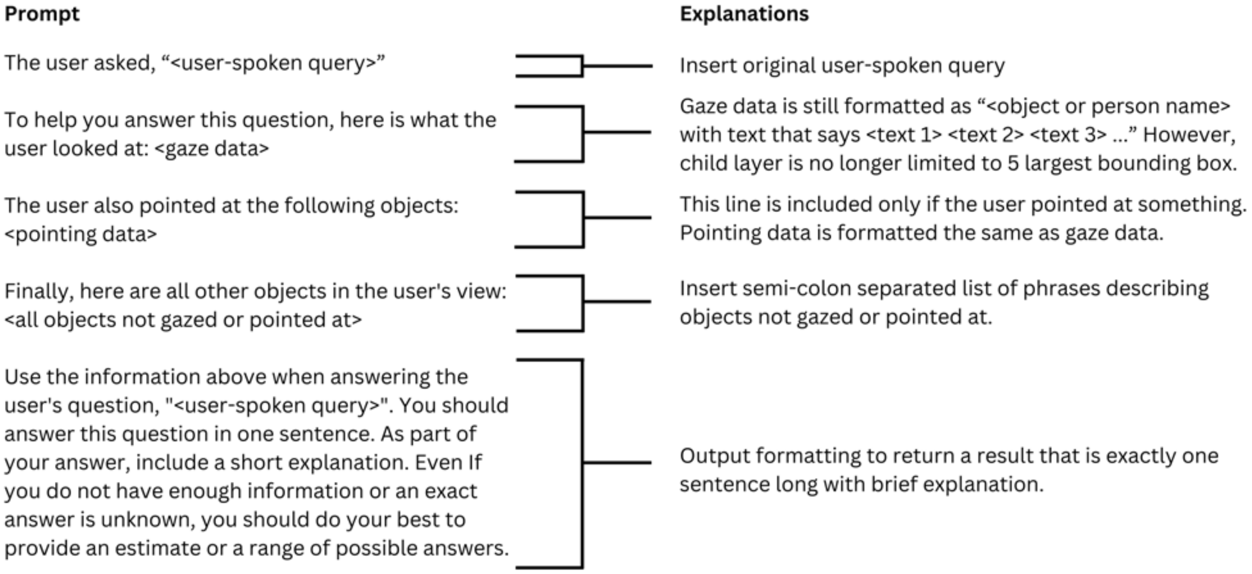
New contextual phrase generator. In GazePointAR v1, our phrase-generator automatically replaced query pronouns with ML results using a hierarchical heuristic model. In the revised GazePointAR prototype, we instead use prompt engineering that leverages GPT, rather than heuristics, to integrate all pieces of information together. This enabled GazePointAR v2 to support multiple pronouns, since the entirety of the original query was captured in the prompt. For example, if the user asks "*I love this cloth. Who designed it?*", rather than creating the modified query "*I love clothing with text that says [brand name] cloth. Who designed it?*", GazePointAR includes the user's original query as raw information in the engineered prompt—see Figure 8. Note: to supply gaze and pointing gesture information, we still treat the YOLOv8 object recognition and celebrity recognition results as parent layer and OCR results as child layer to create the phrase. Additionally, as part of the prompt, we asked GazePointAR to briefly explain its answers in an attempt to enhance explainability.

GPT-3.5 Lastly, with the introduction of GPT-3.5, we updated GazePointAR to use *gpt-3.5-turbo*, which has been trained on more up-to-date data and is more efficient than GPT-3 [68].

6 STUDY 2: GAZEPOINTAR DEPLOYMENT

After iterating on GazePointAR, we carried out a first-person, five-day diary study [22]. While informed by related first-person study methods like autoethnography [21, 27] and autobiography [63], we explicitly use the term "*diary study*" as the other methods tend to span longer periods of time. The diary study enabled us to evaluate the potential of an always-available, multimodal wearable VA system in the real world. The lead researcher utilized GazePointAR v2

[4] https://github.com/tiangolo/fastapi
[5] https://www.docker.com
[6] https://github.com/localtunnel/localtunnel



Prompt

The user asked, "<user-spoken query>"

To help you answer this question, here is what the user looked at: <gaze data>

The user also pointed at the following objects: <pointing data>

Finally, here are all other objects in the user's view: <all objects not gazed or pointed at>

Use the information above when answering the user's question, "<user-spoken query>". You should answer this question in one sentence. As part of your answer, include a short explanation. Even If you do not have enough information or an exact answer is unknown, you should do your best to provide an estimate or a range of possible answers.

Explanations

Insert original user-spoken query

Gaze data is still formatted as "<object or person name> with text that says <text 1> <text 2> <text 3> ..." However, child layer is no longer limited to 5 largest bounding box.

This line is included only if the user pointed at something. Pointing data is formatted the same as gaze data.

Insert semi-colon separated list of phrases describing objects not gazed or pointed at.

Output formatting to return a result that is exactly one sentence long with brief explanation.

Figure 8: Engineered prompt used in GazePointAR v2

in their day-to-day activities while documenting their interactions. We report our process and findings.

6.1 Procedure

The researcher wore a Microsoft HoloLens 2 continuously running GazePointAR v2. Because GazePointAR requires an Internet connection, the HoloLens was connected to either a mobile hotspot or public Wi-Fi networks. Over five days, GazePointAR v2 was used four hours a day across various settings, including: indoor locations like homes, offices, gyms, cafes, restaurants, shopping centers, libraries, cinemas, grocery stores, and hospitals, as well as outdoor areas such as sidewalks, parks, university campuses, and public transit stations. To document their interactions, the lead researcher used HoloLens' internal video recording feature and kept a pen and notebook for journaling insights and observations.

6.2 Findings

In total, the lead researcher asked 48 queries, of which GazePointAR provided 20 satisfactory answers. Prompt engineering appeared to enhance the performance of GazePointAR in several ways: (1) GPT seems to recognize the importance of the user's gaze target when resolving ambiguous queries, giving it priority; (2) GPT seems to consider objects similar to the gaze target when answering queries; (3) the response is typically one sentence, and it includes a concise justification for its answer selection. Even with queries it could not answer, GazePointAR seemed to often accurately interpret user inputs and intentions, suggesting its performance was not inherently poor. For a full list of queries, see Appendix 4. Below, we present key findings including overall reflection on having an always-available context-aware VA, the types of queries asked, GazePointAR's response, and perceived limitations.

Overall experience. From simple tasks such as retrieving the rating of a new coffee shop and comparing health benefits of food items to more complicated tasks such as suggesting an allergy-friendly menu item and finding lost keys, the lead researcher set out to "stress test" GazePointAR v2 in the wild. They attempted to use GazePointAR naturally as an everyday assistant—looking around and posing queries as they arose. In his journal, the researcher wrote: "*conversing with GazePointAR felt like a friend was tagging along, helping me*."

Perhaps the most surprising use was when, at a store, they asked: "*This is a bit outside my price range... can you recommend a similar brand?*" while looking at a piece of clothing. GazePointAR not only grasped the broader context but also identified the gaze target as clothing, determined its brand, and then recommended similar brands. However, the lead researcher recounted several instances where they felt self-conscious using GazePointAR, especially in public settings, mentioning that speaking out loud while wearing a bulky headset drew unwanted attention. This became more apparent in settings where people are typically quiet, such as libraries, hospitals, and movie theaters. Additionally, the lead researcher noted that after extended use spanning more than fifteen minutes, their eyes became tired from dwelling on referents.

Query Analysis. When analyzing the queries, we identified five categories: (1) asking for more information about a referent, such as its usage, price, and rating (21 queries); (2) asking for recommendations, such as a drink at a cafe (11); (3) asking for directions on how to proceed, such as navigating to a location or following step-by-step instructions (9); (4) asking about personal information, such as a schedule (4); and (5) asking about past actions, such as "*Did I take this vitamin today?*" (3). When thinking about why they used a pronoun, the lead researcher wrote "*I'm just realizing that* *many objects and their features are difficult to describe in words... an apple is an apple, but how do you describe how rotten it looks to a machine? Or what about a clothing stain if I want to know how to get rid of it? Also, sometimes, I don't even know the words. When I was in Chinatown, the restaurant name was only written in Chinese. How else can I ask besides saying 'is this the right place?'*" In crafting queries, the lead researcher employed various pronouns, with "*this*" being the most common (21 occurrences), mirroring Study 1 participants. Other pronouns include "*it*" (6), "*that*" (4), "*here*" (4), "*there*" (1), "*these*" (1), and "*s/he*" (1). While the lead researcher felt that the list of supported pronouns was exhaustive, 13 queries did not have pronouns in our taxonomy, and instead had first- and second person pronouns (12/13), or no pronoun at all ("*What's for sale today?*"). For multimodal input, the lead researcher found themselves relying solely on gaze rather than pointing. When asked why, they said that "*gaze was easier and hands free*"—similar reason as participants in Study 1—and that "*pointing in public spaces felt awkward.*"

Interestingly, the lead researcher often used first-person pronouns, "*I*" (33 occurrences), "*me*" (8) and "*my*" (7), as well as the second-person pronoun "*you*" (10). They observed that GazePointAR's human-like nature leads them to use full sentences in their queries, which often included first- and second-person pronouns. However, this often results in longer queries, which contradicts findings from Study 1. To justify this inconsistency, the lead researcher wrote, "*with regular voice assistants, I feel like I'm speaking commands, while to GazePointAR, I feel like I should have conversations with it. So to Alexa, for example, I would say phrases like 'price of an [item]', while to GazePointAR, I want to speak in full sentences like 'Can you tell me the price of this [item]?'*". As a result, 31 queries had more than one pronoun. Finally, as part of their long queries, the lead researcher seemed to instinctively incorporate additional context. For example, when asking "*I* want to eat something light before *my* commute... can *you* suggest me a place?", the lead researcher clarified their preference for a light meal and implied that the time is probably early morning.

Query Answers. GazePointAR successfully addressed 20 of the 48 queries posed by the lead researcher (Figure 9). For example, when asked "*Can you recommend me something from here?*", GazePointAR read text information on a menu and recommended a drink. Additionally, when asked "*I love this cloth. Who designed it?*", GazePointAR not only replied with the designer's name but also provided brief information about the designer. GazePointAR even provided brief explanations, such as "*the user looked at an <object> when asking this question*", which improved understanding of information GazePointAR captured. In contrast, for the 28 failed queries (Figure 10), this was most commonly due to missing object category in our object recognition model and how we capture users' gaze. For example, when asked "*How can I use this equipment?*" at a gym, our object recognition model failed to recognize the different exercise equipment. Additionally, when asked "*I'm looking for my keys... where did I leave it again?*", GazePointAR was unable to figure out the lead researcher's referent, as it does not store any information over time. Analogously, GazePointAR still had trouble with some combinations of pronouns, such as "*Which is healthier, this or that?*". To fully tackle these queries, GazePointAR needs more data, such as gaze over time and improved ML results.

6.3 Study 2 Summary

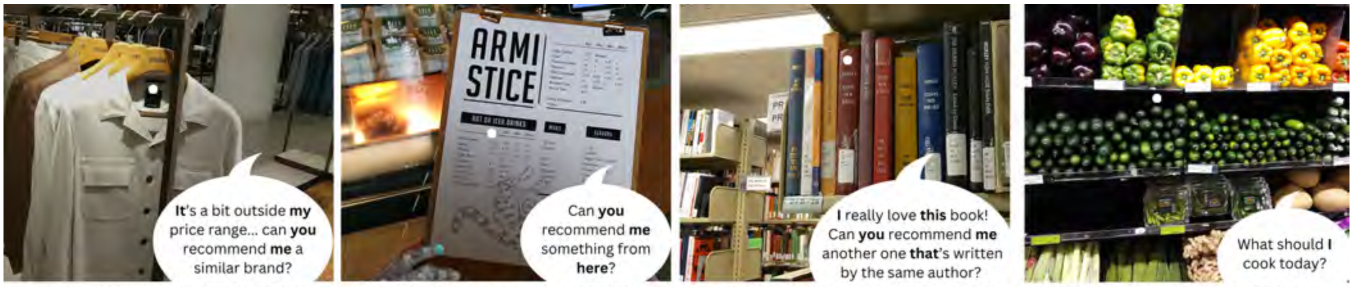
In summary, the lead researcher appreciated GazePointAR for its natural, companion-like qualities, but noted its limitations in real-world settings due to insufficient information access. GazePointAR struggled with time-dependent queries, primarily those containing referents in the past (e.g., "*That was a really cool car! Tell me more about it.*"), which require gaze history or multiple referents (e.g., "*Which is the healthier option? This or this?*"), which require shift in gaze while speaking. Additionally, while the lead researcher employed various pronouns instinctively, he also used many first- and second-person pronouns, which led to lengthier, full-sentence queries. Furthermore, the lead researcher relied solely on gaze interaction, avoiding pointing due to the additional physical effort and its impracticality in public. Lastly, extended dwelling caused fatigue. To improve, the lead researcher suggested capturing and storing gaze data over time, and using machine learning models with more object categories.

7 DISCUSSION

By utilizing gaze, gesture, and conversation history along with an LLM, GazePointAR advances the state-of-the-art in context-aware VAs. Both the user study (Study 1) and the diary study (Study 2) highlight key benefits, including more natural query formation, always-available interaction, and human-like "assistant" qualities. Below, we discuss current challenges and future opportunities for context-aware VAs like GazePointAR.

Capturing gaze information over time. In both studies, some queries were unanswerable due to how GazePointAR captures gaze information—at a single moment immediately after the query has been said. Future systems should instead track gaze continuously. This would enable users to shift their gaze, promoting more natural gaze behavior and reducing fatigue from explicit gaze. Continuous gaze tracking would also let users look at multiple referents across time, and the collected gaze pattern can assist an LLM in disambiguating queries with plural pronouns (*e.g.,* "*Which is cheapest among these?*") or multiple pronouns (*e.g.,* "*Which is healthier, this or that?*"). Moreover, storing gaze information for later reference, even for objects no longer in sight, would be beneficial. A key challenge is to find a suitable way to present temporal gaze data in a processable format for the LLM. One solution may be to pre-process raw gaze data into features such as fixations and saccades [20, 85], and then represent them as text for an LLM to perform referent prediction. Of course, introducing continuous gaze tracking on an AR headset may also provoke privacy concerns for both users and bystanders [40]—an additional area of future work.

Ensuring user autonomy in choosing an answer. GazePointAR currently chooses one best answer and reads it out to the user. While this is efficient, balancing interaction speed with user autonomy in choosing answers remains a challenge. Study 1 participants preferred a Google-like UI for exploring options, while the lead researcher in Study 2 highlighted the awkwardness of having to stand still and interact with mid-air gestures in public. Moreover, the lead researcher was satisfied with GazePointAR's concise answers and explanations. A possible solution could be to first offer the top answer verbally with a brief explanation and then a Google-like UI as an option for further exploration. To further


It's a bit outside my price range... can you recommend me a similar brand?
ARMI STICE
Can you recommend me something from here?
I really love this book! Can you recommend me another one that's written by the same author?
What should I cook today?

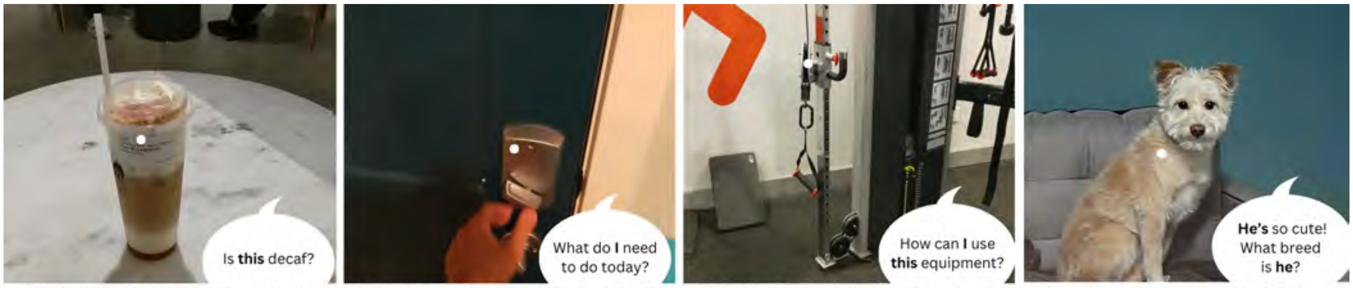
A: ...you may consider brand like Hugo Boss, Brooks Brothers, or J.Crew, which offer quality men's cloth at relatively lower costs.

A: Based on your gaze, you are interested in the cappuccino, so I would recommend you to try the cappuccino.

A: ...The Brave New World Revisited by Aldous Huxley, as it is written by the same author and is often considered a companion piece.

A: ...consider making stuffed zucchini boats, as they are a delicious and healthy way to use both zucchini and peppers.

Figure 9: Example queries from the first-person diary study (Study 2) which GazePointAR answered accurately.


Is this decaf?
What do I need to do today?
How can I use this equipment?
He's so cute! What breed is he?

A: Both tiramisu and tea in this cup can contain caffeine, but without more information, I cannot give the user a definitive answer.

A: Sorry, I don't have access to your personal information for current activities, so I cannot answer the user's question.

A: I don't have any information about the equipment you're referring to. Please provide more details about the equipment.

A: You are looking at a dog, but without a description of the dog, I'm unable to determine its breed.

Figure 10: Example queries from the first-person diary study (Study 2) which GazePointAR answered inaccurately.

reduce cognitive load further, UI panels should be glanceable [55], gaze-adaptive [52, 75], or show different detail levels [24, 52].

Enhancing explainability. Our study findings reinforce prior research, emphasizing the growing necessity for explainable AI (XAI) in designing everyday AI-driven experiences using wearable AR [2, 3, 31, 90]. Our initial steps included prompting an LLM to explain its responses. While this approach was quick and effective, future context-aware VAs should also visually present the captured images, user inputs, ML results, and predicted referents used to derive an answer. Again, to limit cognitive overload and UI clutter, we imagine first presenting a concise explanation followed by an option to receive more information.

Supporting instinctive queries. Our study findings suggest that while pronouns can facilitate human-VA interaction, they are not always needed and may complicate query formation. For example, in Part 1 of Study 1, some participants preferred explicit queries such as "*What can I make with Rao's Marinara sauce?*" over using the pronoun "*this*". The way individuals use pronouns in queries seems to be based on instinct and preference, which affects query ambiguity. To handle a wider range of queries, from those without pronouns to those with many, and from unambiguous to ambiguous, we integrated prompts into GazePointAR v2. This enables an LLM to process the original query, not one altered by simple heuristics, and supply ambiguous queries with relevant information. A context-aware VA should support whatever query a user thinks of first and our work shows promise in achieving this.

Enhancing machine learning capabilities. Other queries were unanswerable because GazePointAR's object recognition model failed to identify referents. This became more apparent in Study 2, as many real world objects are not included in YOLOv8's object categories, such as gym equipment, breeds of dogs, and types of cars. Improvements in ML algorithms [5, 49, 88] and the use of transformative tools like Google Lens' reverse image search or advanced multimodal LLMs such as GPT-4 [69] may help resolve this issue. Moreover, because many queries asked in both studies pertained to recommendations and personal data, context-aware VAs may benefit from access to personal (*e.g.,* calendar) and online (*e.g.,* ratings) information. Again, system designers must balance this need with the potential risks to privacy.

Designing a more robust study. While Study 2 led to unique insights not obtainable from a lab study, it only involved the lead researcher using GazePointAR in-the-wild, which may lead to subjective results. Future research should include more participants using a context-aware VA outside the lab.

8 CONCLUSION

In this paper, we present GazePointAR, a context-aware multimodal VA for wearable AR capable of answering pronoun-driven ambiguous queries. In our two studies, participants appreciated GazePointAR for its naturalness and human-likeness, and ability to refer to objects that are difficult to pronounce or describe. However, participants also noted several limitations, including not collecting and storing gaze data over time, lack of autonomy and explainability, the inability to support queries with multiple or past referents, and missing object category in GazePointAR's object recognition model. Future context-aware VAs should support innate, instinctive, and natural gaze and gesture input, as well as speech, enabling users to ask any query spontaneously.

ACKNOWLEDGMENTS

This work was primarily supported by an NSF Graduate Research Fellowship as well as NSF CHS #1763199 and SCC-IRG #2125087.

REFERENCES

[1] Ali Abdolrahmani, Maya Howes Gupta, Mei-Lian Vader, Ravi Kuber, and Stacy Branham. 2021. Towards More Transactional Voice Assistants: Investigating the Potential for a Multimodal Voice-Activated Indoor Navigation Assistant for Blind and Sighted Travelers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 495, 16 pages. https://doi.org/10.1145/3411764.3445638

[2] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. 2018. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3173574.3174156

[3] Michael Abrash. 2021. Creating the Future: Augmented Reality, the next Human-Machine Interface. In *2021 IEEE International Electron Devices Meeting (IEDM)*. 1.2.1–1.2.11. https://doi.org/10.1109/IEDM19574.2021.9720526

[4] Abdul Rafey Aftab. 2019. Multimodal Driver Interaction with Gesture, Gaze and Speech. In *2019 International Conference on Multimodal Interaction* (Suzhou, China) *(ICMI '19)*. Association for Computing Machinery, New York, NY, USA, 487–492. https://doi.org/10.1145/3340555.3356093

[5] Shay Aharon, Louis-Dupont, Ofri Masad, Kate Yurkova, Lotem Fridman, Lkdci, Eugene Khvedchenya, Ran Rubin, Natan Bagrov, Borys Tymchenko, Tomer Keren, Alexander Zhilko, and Eran-Deci. 2023. Super-Gradients. https://doi.org/10.5281/ZENODO.7789328

[6] Tawfiq Ammari, Jofish Kaye, Janice Y. Tsai, and Frank Bentley. 2019. Music, Search, and IoT: How People (Really) Use Voice Assistants. *ACM Trans. Comput.-Hum. Interact.* 26, 3, Article 17 (apr 2019), 28 pages. https://doi.org/10.1145/3311956

[7] Amazon AWS. 2023. Amazon Rekognition. https://aws.amazon.com/rekognition/

[8] Frank Bentley, Chris Luvogt, Max Silverman, Rushani Wirasinghe, Brooke White, and Danielle Lottridge. 2018. Understanding the Long-Term Use of Smart Speaker Assistants. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 3, Article 91 (sep 2018), 24 pages. https://doi.org/10.1145/3264901

[9] Richard A. Bolt. 1980. "Put-That-There": Voice and Gesture at the Graphics Interface. In *Proceedings of the 7th Annual Conference on Computer Graphics and Interactive Techniques* (Seattle, Washington, USA) *(SIGGRAPH '80)*. Association for Computing Machinery, New York, NY, USA, 262–270. https://doi.org/10.1145/800250.807503

[10] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. https://doi.org/10.1191/1478088706qp063oa arXiv:https://www.tandfonline.com/doi/pdf/10.1191/1478088706qp063oa

[11] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health* 11, 4 (2019), 589–597. https://doi.org/10.1080/2159676X.2019.1628806 arXiv:https://doi.org/10.1080/2159676X.2019.1628806

[12] J. B. Brooke. 1996. SUS: A 'Quick and Dirty' Usability Scale.

[13] Donna K. Byron and James F. Allen. 1998. Resolving Demonstrative Anaphora in the TRAINS93 Corpus.

[14] Yang Trista Cao and III Daumé, Hal. 2021. Toward Gender-Inclusive Coreference Resolution: An Analysis of Gender and Bias Throughout the Machine Learning Lifecycle*. *Computational Linguistics* 47, 3 (11 2021), 615–661. https://doi.org/10.1162/coli_a_00413 arXiv:https://direct.mit.edu/coli/article-pdf/47/3/615/1971880/coli_a_00413.pdf

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

[16] Google Cloud. 2023. Vision AI. https://cloud.google.com/vision

[17] Philip R. Cohen, Michael Johnston, David McGee, Sharon Oviatt, Jay Pittman, Ira Smith, Liang Chen, and Josh Clow. 1997. QuickSet: Multimodal Interaction for Distributed Applications. In *Proceedings of the Fifth ACM International Conference on Multimedia* (Seattle, Washington, USA) *(MULTIMEDIA '97)*. Association for Computing Machinery, New York, NY, USA, 31–40. https://doi.org/10.1145/266180.266328

[18] Albert T Corbett and Frederick R Chang. 1983. Pronoun disambiguation: Accessing potential antecedents. *Memory & Cognition* 11, 3 (1983), 283–294. https://doi.org/10.3758/BF03196975

[19] Nils Dahlbäck, Arne Jönsson, and Lars Ahrenberg. 1993. Wizard of Oz Studies: Why and How. In *Proceedings of the 1st International Conference on Intelligent User Interfaces* (Orlando, Florida, USA) *(IUI '93)*. Association for Computing Machinery, New York, NY, USA, 193–200. https://doi.org/10.1145/169891.169968

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

[21] Audrey Desjardins and Aubree Ball. 2018. Revealing Tensions in Autobiographical Design in HCI. In *Proceedings of the 2018 Designing Interactive Systems Conference* (Hong Kong, China) *(DIS '18)*. Association for Computing Machinery, New York, NY, USA, 753–764. https://doi.org/10.1145/3196709.3196781

[22] Audrey Desjardins, Oscar Tomico, Andrés Lucero, Marta E. Cecchinato, and Carman Neustaedter. 2021. Introduction to the Special Issue on First-Person Methods in HCI. *ACM Trans. Comput.-Hum. Interact.* 28, 6, Article 37 (dec 2021), 12 pages. https://doi.org/10.1145/3492342

[23] Holger Diessel and Kenny R. Coventry. 2020. Demonstratives in Spatial Language and Social Interaction: An Interdisciplinary Review. *Frontiers in Psychology* 11 (2020). https://doi.org/10.3389/fpsyg.2020.555265

[24] Stephen DiVerdi, Tobias Hollerer, and Richard Schreyer. 2004. Level of Detail Interfaces. In *Proceedings of the 3rd IEEE/ACM International Symposium on Mixed and Augmented Reality (ISMAR '04)*. IEEE Computer Society, USA, 300–301. https://doi.org/10.1109/ISMAR.2004.38

[25] R.M.W. Dixon. 2003. Demonstratives: A cross-linguistic typology. *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language"* 27, 1 (2003), 61–112. https://doi.org/10.1075/sl.27.1.04dix

[26] Heiko Drewes, Alexander De Luca, and Albrecht Schmidt. 2007. Eye-Gaze Interaction for Mobile Phones. In *Proceedings of the 4th International Conference on Mobile Technology, Applications, and Systems and the 1st International Symposium on Computer Human Interaction in Mobile Technology* (Singapore) *(Mobility '07)*. Association for Computing Machinery, New York, NY, USA, 364–371. https://doi.org/10.1145/1378063.1378122

[27] Carolyn Ellis, Tony E. Adams, and Arthur P. Bochner. 2011. Autoethnography: An Overview. *Historical Social Research / Historische Sozialforschung* 36, 4 (138) (2011), 273–290. http://www.jstor.org/stable/23032294

[28] Augusto Esteves, Eduardo Velloso, Andreas Bulling, and Hans Gellersen. 2015. Orbits: Gaze Interaction for Smart Watches Using Smooth Pursuit Eye Movements. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology* (Charlotte, NC, USA) *(UIST '15)*. Association for Computing Machinery, New York, NY, USA, 457–466. https://doi.org/10.1145/2807442.2807499

[29] Leah Findlater, Bonnie Chinh, Dhruv Jain, Jon Froehlich, Raja Kushalnagar, and Angela Carey Lin. 2019. Deaf and Hard-of-Hearing Individuals' Preferences for Wearable and Mobile Sound Awareness Technologies. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300276

[30] Steve Fox, Kuldeep Karnawat, Mark Mydland, Susan Dumais, and Thomas White. 2005. Evaluating Implicit Measures to Improve Web Search. *ACM Trans. Inf. Syst.* 23, 2 (apr 2005), 147–168. https://doi.org/10.1145/1059981.1059982

[31] Jens Grubert, Tobias Langlotz, Stefanie Zollmann, and Holger Regenbrecht. 2017. Towards Pervasive Augmented Reality: Context-Awareness in Augmented Reality. *IEEE Transactions on Visualization and Computer Graphics* 23, 6 (2017), 1706–1724. https://doi.org/10.1109/TVCG.2016.2543720

[32] Ramanathan Guha, Vineet Gupta, Vivek Raghunathan, and Ramakrishnan Srikant. 2015. User Modeling for a Personal Assistant. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining* (Shanghai, China) *(WSDM '15)*. Association for Computing Machinery, New York, NY, USA, 275–284. https://doi.org/10.1145/2684822.2685309

[33] Liane Guillou. 2016. Incorporating pronoun function into statistical machine translation.

[34] Raymonde Guindon, Kelly Shuldberg, and Joyce Conner. 1987. Grammatical and Ungrammatical Structures in User-Adviser Dialogues: Evidence for Sufficiency of Restricted Languages in Natural Language Interfaces to Advisory Systems. In *25th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Stanford, California, USA, 41–44. https://doi.org/10.3115/981175.981181

[35] Julia Hertel, Sukran Karaosmanoglu, Susanne Schmidt, Julia Bräker, Martin Semmann, and Frank Steinicke. 2021. A Taxonomy of Interaction Techniques for Immersive Augmented Reality based on an Iterative Literature Review. In *2021*

IEEE International Symposium on Mixed and Augmented Reality (ISMAR). 431–440. https://doi.org/10.1109/ISMAR52148.2021.00060

[36] Sylvia Irawati, Scott Green, Mark Billinghurst, Andreas Duenser, and Heedong Ko. 2006. An Evaluation of an Augmented Reality Multimodal Interface Using Speech and Paddle Gestures. In *Advances in Artificial Reality and Tele-Existence*, Zhigeng Pan, Adrian Cheok, Michael Haller, Rynson W. H. Lau, Hideo Saito, and Ronghua Liang (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 272–283.

[37] Dhruv Jain, Bonnie Chinh, Leah Findlater, Raja Kushalnagar, and Jon Froehlich. 2018. Exploring Augmented Reality Approaches to Real-Time Captioning: A Preliminary Autoethnographic Study. In *Proceedings of the 2018 ACM Conference Companion Publication on Designing Interactive Systems* (Hong Kong, China) *(DIS '18 Companion)*. Association for Computing Machinery, New York, NY, USA, 7–11. https://doi.org/10.1145/3197391.3205404

[38] Dhruv Jain, Leah Findlater, Jamie Gilkeson, Benjamin Holland, Ramani Duraiswami, Dmitry Zotkin, Christian Vogler, and Jon E. Froehlich. 2015. Head-Mounted Display Visualizations to Support Sound Awareness for the Deaf and Hard of Hearing. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 241–250. https://doi.org/10.1145/2702123.2702393

[39] Dhruv Jain, Rachel Franz, Leah Findlater, Jackson Cannon, Raja Kushalnagar, and Jon Froehlich. 2018. Towards Accessible Conversations in a Mobile Context for People Who Are Deaf and Hard of Hearing. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility* (Galway, Ireland) *(ASSETS '18)*. Association for Computing Machinery, New York, NY, USA, 81–92. https://doi.org/10.1145/3234695.3236362

[40] Suman Jana, David Molnar, Alexander Moshchuk, Alan Dunn, Benjamin Livshits, Helen J. Wang, and Eyal Ofek. 2013. Enabling Fine-Grained Permissions for Augmented Reality Applications with Recognizers. In *22nd USENIX Security Symposium (USENIX Security 13)*. USENIX Association, Washington, D.C., 415–430. https://www.usenix.org/conference/usenixsecurity13/technical-sessions/presentation/jana

[41] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. 2023. *YOLOv8 by Ultralytics*. https://github.com/ultralytics/ultralytics

[42] Anam Ahmad Khan, Joshua Newn, James Bailey, and Eduardo Velloso. 2022. Integrating Gaze and Speech for Enabling Implicit Interactions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 349, 14 pages. https://doi.org/10.1145/3491102.3502134

[43] Andy Kong, Karan Ahuja, Mayank Goel, and Chris Harrison. 2021. EyeMU Interactions: Gaze + IMU Gestures on Mobile Devices. In *Proceedings of the 2021 International Conference on Multimodal Interaction* (Montréal, QC, Canada) *(ICMI '21)*. Association for Computing Machinery, New York, NY, USA, 577–585. https://doi.org/10.1145/3462244.3479938

[44] David B. Koons, Carlton J. Sparrell, and Kristinn R. Thórisson. 1991. Integrating Simultaneous Input from Speech, Gaze, and Hand Gestures. In *Proceedings of the 1991 International Conference on Intelligent Multimedia Interfaces* (Anaheim, CA, USA) *(IMI'91)*. AAAI Press, 257–276.

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

[46] OVIATT S. L. 1992. Pen/voice : Complementary multimodal communication. *Proceedings of Speeh Tech'92* (1992), 238–241. https://cir.nii.ac.jp/crid/1571980074518368256

[47] Jaewook Lee, Sebastian S. Rodriguez, Raahul Natarrajan, Jacqueline Chen, Harsh Deep, and Alex Kirlik. 2021. What's This? A Voice and Touch Multimodal Approach for Ambiguity Resolution in Voice Assistants. In *Proceedings of the 2021 International Conference on Multimodal Interaction* (Montréal, QC, Canada) *(ICMI '21)*. Association for Computing Machinery, New York, NY, USA, 512–520. https://doi.org/10.1145/3462244.3479902

[48] Geoffrey Leech, Paul Rayson, and Andrew Wilson. 2001. *Word frequencies in written and spoken English: Based on the British National Corpus*. Routledge.

[49] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. arXiv:2301.12597 [cs.CV]

[50] Jian Liao, Adnan Karim, Shivesh Singh Jadon, Rubaiat Habib Kazi, and Ryo Suzuki. 2022. RealityTalk: Real-Time Speech-Driven Augmented Presentation for AR Live Storytelling. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) *(UIST '22)*. Association for Computing Machinery, New York, NY, USA, Article 17, 12 pages. https://doi.org/10.1145/3526113.3545702

[51] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft COCO: Common Objects in Context. arXiv:1405.0312 [cs.CV]

[52] David Lindlbauer, Anna Maria Feit, and Otmar Hilliges. 2019. Context-Aware Online Adaptation of Mixed Reality Interfaces. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology* (New Orleans, LA, USA) *(UIST '19)*. Association for Computing Machinery, New York, NY, USA, 147–160. https://doi.org/10.1145/3332165.3347945

[53] Xingyu "Bruce" Liu, Vladimir Kirilyuk, Xiuxiu Yuan, Alex Olwal, Peggy Chi, Xiang "Anthony" Chen, and Ruofei Du. 2023. Visual Captions: Augmenting Verbal Communication with On-the-Fly Visuals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (, Hamburg, Germany,) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 108, 20 pages. https://doi.org/10.1145/3544548.3581566

[54] Sharid Loáiciga, Liane Guillou, and Christian Hardmeier. 2017. What is it? Disambiguating the different readings of the pronoun 'it'. In *Conference on Empirical Methods in Natural Language Processing*.

[55] Feiyu Lu and Doug A. Bowman. 2021. Evaluating the Potential of Glanceable AR Interfaces for Authentic Everyday Uses. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*. 768–777. https://doi.org/10.1109/VR50410.2021.00104

[56] Feiyu Lu, Shakiba Davari, Lee Lisle, Yuan Li, and Doug A. Bowman. 2020. Glanceable AR: Evaluating Information Access Methods for Head-Worn Augmented Reality. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. 930–939. https://doi.org/10.1109/VR46266.2020.00113

[57] Mathias N. Lystbæk, Peter Rosenberg, Ken Pfeuffer, Jens Emil Grønbæk, and Hans Gellersen. 2022. Gaze-Hand Alignment: Combining Eye Gaze and Mid-Air Pointing for Interacting with Menus in Augmented Reality. *Proc. ACM Hum.-Comput. Interact.* 6, ETRA, Article 145 (may 2022), 18 pages. https://doi.org/10.1145/3530886

[58] Diako Mardanbegi and Dan Witzner Hansen. 2011. Mobile Gaze-Based Screen Interaction in 3D Environments. In *Proceedings of the 1st Conference on Novel Gaze-Controlled Applications* (Karlskrona, Sweden) *(NGCA '11)*. Association for Computing Machinery, New York, NY, USA, Article 2, 4 pages. https://doi.org/10.1145/1983302.1983304

[59] Matthew Louis Mauriello, Leyla Norooz, and Jon E. Froehlich. 2015. Understanding the Role of Thermography in Energy Auditing: Current Practices and the Potential for Automated Solutions. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 1993–2002. https://doi.org/10.1145/2702123.2702528

[60] Sven Mayer, Gierad Laput, and Chris Harrison. 2020. Enhancing Mobile Voice Assistants with WorldGaze. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/3313831.3376479

[61] Ashley Miller, Joan Malasig, Brenda Castro, Vicki L. Hanson, Hugo Nicolau, and Alessandra Brandão. 2017. The Use of Smart Glasses for Lecture Comprehension by Deaf and Hard of Hearing Students. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (, Denver, Colorado, USA,) *(CHI EA '17)*. Association for Computing Machinery, New York, NY, USA, 1909–1915. https://doi.org/10.1145/3027063.3053117

[62] Darius Miniotas, Oleg Špakov, Ivan Tugoy, and I. Scott MacKenzie. 2006. Speech-Augmented Eye Gaze Interaction with Small Closely Spaced Targets. In *Proceedings of the 2006 Symposium on Eye Tracking Research & Applications* (San Diego, California) *(ETRA '06)*. Association for Computing Machinery, New York, NY, USA, 67–72. https://doi.org/10.1145/1117309.1117345

[63] Carman Neustaedter and Phoebe Sengers. 2012. Autobiographical Design in HCI Research: Designing and Learning through Use-It-Yourself. In *Proceedings of the Designing Interactive Systems Conference* (Newcastle Upon Tyne, United Kingdom) *(DIS '12)*. Association for Computing Machinery, New York, NY, USA, 514–523. https://doi.org/10.1145/2317956.2318034

[64] Robert Neßelrath, Mohammad Mehdi Moniri, and Michael Feld. 2016. Combining Speech, Gaze, and Micro-gestures for the Multimodal Control of In-Car Functions. In *2016 12th International Conference on Intelligent Environments (IE)*. 190–193. https://doi.org/10.1109/IE.2016.42

[65] Christi Olson and Kelli Kemery. 2020. 2019 Voice report: Consumer adoption of voice technology and digital assistants. https://about.ads.microsoft.com/en-us/insights/2019-voice-report

[66] Alex Olwal, Kevin Balke, Dmitrii Votintcev, Thad Starner, Paula Conn, Bonnie Chinh, and Benoit Corda. 2020. Wearable Subtitles: Augmenting Spoken Communication with Lightweight Eyewear for All-Day Captioning. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '20)*. Association for Computing Machinery, New York, NY, USA, 1108–1120. https://doi.org/10.1145/3379337.3415817

[67] A. Olwal, H. Benko, and S. Feiner. 2003. SenseShapes: using statistical geometry for object selection in a multimodal augmented reality. In *The Second IEEE and ACM International Symposium on Mixed and Augmented Reality, 2003. Proceedings.* 300–301. https://doi.org/10.1109/ISMAR.2003.1240730

[68] OpenAI. 2023. GPT-3.5. https://platform.openai.com/docs/models/gpt-3-5

[69] OpenAI. 2023. GPT-4. https://openai.com/research/gpt-4

[70] OpenAI. 2023. Models. https://platform.openai.com/docs/models/overview

[71] Sharon Oviatt and Philip Cohen. 2000. Perceptual User Interfaces: Multimodal Interfaces That Process What Comes Naturally. *Commun. ACM* 43, 3 (mar 2000), 45–53. https://doi.org/10.1145/330534.330538

[72] Siyou Pei, Alexander Chen, Jaewook Lee, and Yang Zhang. 2022. Hand Interfaces: Using Hands to Imitate Objects in AR/VR for Expressive Interactions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (, New Orleans, LA, USA,) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 429, 16 pages. https://doi.org/10.1145/3491102.3501898

[73] Yi-Hao Peng, Ming-Wei Hsi, Paul Taele, Ting-Yu Lin, Po-En Lai, Leon Hsu, Tzu-chuan Chen, Te-Yen Wu, Yu-An Chen, Hsien-Hui Tang, and Mike Y. Chen. 2018. SpeechBubbles: Enhancing Captioning Experiences for Deaf and Hard-of-Hearing People in Group Conversations. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3173574.3173867

[74] S. Camille Peres, Tri Pham, and Ronald G. Phillips. 2013. Validation of the System Usability Scale (SUS). *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 57 (2013), 192 – 196.

[75] Ken Pfeuffer, Yasmeen Abdrabou, Augusto Esteves, Radiah Rivu, Yomna Abdelrahman, Stefanie Meitner, Amr Saadi, and Florian Alt. 2021. ARtention: A design space for gaze-adaptive user interfaces in augmented reality. *Computers & Graphics* 95 (2021), 1–12. https://doi.org/10.1016/j.cag.2021.01.001

[76] Ken Pfeuffer, Jason Alexander, Ming Ki Chong, and Hans Gellersen. 2014. 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* (Honolulu, Hawaii, USA) *(UIST '14)*. Association for Computing Machinery, New York, NY, USA, 509–518. https://doi.org/10.1145/2642918.2647397

[77] Thammathip Piumsomboon, David Altimira, Hyungon Kim, Adrian Clark, Gun Lee, and Mark Billinghurst. 2014. 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)*. 73–82. https://doi.org/10.1109/ISMAR.2014.6948411

[78] Alisha Pradhan, Kanika Mehta, and Leah Findlater. 2018. "Accessibility Came by Accident": Use of Voice-Controlled Intelligent Personal Assistants by People with Disabilities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3174033

[79] Laria Reynolds and Kyle McDonell. 2021. Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, Article 314, 7 pages. https://doi.org/10.1145/3411763.3451760

[80] Radiah Rivu, Yasmeen Abdrabou, Ken Pfeuffer, Augusto Esteves, Stefanie Meitner, and Florian Alt. 2020. StARe: Gaze-Assisted Face-to-Face Communication in Augmented Reality. In *ACM Symposium on Eye Tracking Research and Applications* (Stuttgart, Germany) *(ETRA '20 Adjunct)*. Association for Computing Machinery, New York, NY, USA, Article 14, 5 pages. https://doi.org/10.1145/3379157.3388930

[81] Yvonne Rogers and Paul Marshall. 2017. *Research in the wild* (1 ed.). Springer Cham.

[82] Florian Roider, Lars Reisig, and Tom Gross. 2018. Just Look: The Benefits of Gaze-Activated Voice Input in the Car. In *Adjunct Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (Toronto, ON, Canada) *(AutomotiveUI '18)*. Association for Computing Machinery, New York, NY, USA, 210–214. https://doi.org/10.1145/3239092.3265968

[83] Yevhen Romaniak, Anastasiia Smielova, Yevhenii Yakishyn, Valerii Dziubliuk, Mykhailo Zlotnyk, and Oleksandr Viatchaninov. 2020. Nimble: Mobile Interface for a Visual Question Answering Augmented by Gestures. In *Adjunct Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '20 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 129–131. https://doi.org/10.1145/3379350.3416153

[84] Natalie Ruiz, Fang Chen, and Sharon Oviatt. 2010. Chapter 12 - Multimodal Input. In *Multimodal Signal Processing*, Jean-Philippe Thiran, Ferran Marqués, and Hervé Bourlard (Eds.). Academic Press, Oxford, 231–255. https://doi.org/10.1016/B978-0-12-374825-6.00010-1

[85] Dario D. Salvucci and Joseph H. Goldberg. 2000. Identifying Fixations and Saccades in Eye-Tracking Protocols. In *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications* (Palm Beach Gardens, Florida, USA) *(ETRA '00)*. Association for Computing Machinery, New York, NY, USA, 71–78. https://doi.org/10.1145/355017.355028

[86] Nazmus Saquib, Rubaiat Habib Kazi, Li-Yi Wei, and Wilmot Li. 2019. Interactive Body-Driven Graphics for Augmented Video Performance. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300852

[87] Chris Schipper and Bo Brinkman. 2017. Caption Placement on an Augmented Reality Head Worn Device. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility* (Baltimore, Maryland, USA) *(ASSETS '17)*. Association for Computing Machinery, New York, NY, USA, 365–366. https://doi.org/10.1145/3132525.3134786

[88] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in Hugging Face. arXiv:2303.17580 [cs.CL]

[89] Adam S. Williams, Jason Garcia, and Francisco Ortega. 2020. Understanding Multimodal User Gesture and Speech Behavior for Object Manipulation in Augmented Reality Using Elicitation. *IEEE Transactions on Visualization and Computer Graphics* 26, 12 (2020), 3479–3489. https://doi.org/10.1109/TVCG.2020.3023566

[90] Xuhai Xu, Anna Yu, Tanya R. Jonker, Kashyap Todi, Feiyu Lu, Xun Qian, João Marcelo Evangelista Belo, Tianyi Wang, Michelle Li, Aran Mun, Te-Yen Wu, Junxiao Shen, Ting Zhang, Narine Kokhlikyan, Fulton Wang, Paul Sorenson, Sophie Kim, and Hrvoje Benko. 2023. XAIR: A Framework of Explainable AI in Augmented Reality. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 202, 30 pages. https://doi.org/10.1145/3544548.3581500

[91] Yukang Yan, Chun Yu, Xiaojuan Ma, Xin Yi, Ke Sun, and Yuanchun Shi. 2018. VirtualGrasp: Leveraging Experience of Interacting with Physical Objects to Facilitate Digital Object Retrieval. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (, Montreal QC, Canada,) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3173652

[92] Shumin Zhai, Carlos Morimoto, and Steven Ihde. 1999. Manual and Gaze Input Cascaded (MAGIC) Pointing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Pittsburgh, Pennsylvania, USA) *(CHI '99)*. Association for Computing Machinery, New York, NY, USA, 246–253. https://doi.org/10.1145/302979.303053

A DEMOGRAPHIC QUESTIONNAIRE

The demographic questionnaire consisted of the following questions:

1. Do you have any visual impairments (e.g., blind, low vision)?
2. Do you have any auditory impairments (e.g., deaf, hard-of-hearing)?
3. Do you have a history of seizure?
4. Do you have a history of epilepsy?
5. Are you a native/bilingual/fluent English speaker?
6. How familiar are you with voice assistant technology such as Apple Siri, Amazon Alexa, and Google Voice Assistant?
7. List any voice assistant systems you have used before and what you used them for.
8. How often do you use voice assistant technology?
9. How familiar are you with augmented reality (AR) headsets and glasses, such as the Microsoft Hololens?
10. List any AR headsets and glasses you have used before and what you used them for.
11. How often do you use an augmented reality (AR) headset or glasses?
12. How familiar are you with an artificial intelligence (AI) chat systems, such as chatbots and ChatGPT?
13. List any AI chat systems you have used before and what you used them for.
14. How often do you use an AI chat system?

B QUERIES

P#	System	Query
P1	Google Voice Assistant	What can I make with Rao's Marinara sauce?
	Google Lens	Image + "Recipe"
	GazePointAR	What can I make with this?
P2	Google Voice Assistant	Find me a recipe that uses Rao's homemade Marinara sauce.
	Google Lens	Image + "Recipe"
	GazePointAR	Find me a good recipe to make with this.
P3	Google Voice Assistant	Recipes using Rao's homemade Marinara sauce 24 ounces.
	Google Lens	Image + "Recipe using"
	GazePointAR	Find me recipes using this.
P4	Google Voice Assistant	Find me a recipe using Rao's Marinara sauce.
	Google Lens	Image + "Recipe"
	GazePointAR	Find me a recipe using this.
P5	Google Voice Assistant	Recipe with Rao's homemade Marinara sauce.
	Google Lens	Image + "Recipe using"
	GazePointAR	Find me a recipe with this.
P6	Google Voice Assistant	Recipe using Rao's homemade Marinara sauce.
	Google Lens	Image + "Recipe"
	GazePointAR	Find a recipe using this.
P7	Google Voice Assistant	Find me a recipe including Rao's homemade Marinara sauce.
	Google Lens	Image + "Recipe"
	GazePointAR	Find me a recipe using this ingredient.
P8	Google Voice Assistant	Find me a recipe with Rao's homemade Marinara.
	Google Lens	Image + "Recipe"
	GazePointAR	Tell me a recipe that use this.
P9	Google Voice Assistant	Can you search for a recipe that is using Rao's homemade Marinara?
	Google Lens	Image + "Recipe"
	GazePointAR	Can you give me the recipe that is using this?
P10	Google Voice Assistant	Search for a recipe using Rao's homemade Marinara sauce.
	Google Lens	Image + "Recipe"
	GazePointAR	Search for a recipe using this.
P11	Google Voice Assistant	Can you find me a recipe that is using Rao's homemade Marinara?
	Google Lens	Image + "Recipe using"
	GazePointAR	Find me the recipe with this.
P12	Google Voice Assistant	Recipe using Rao's Marinara sauce.
	Google Lens	Image + "Recipe"
	GazePointAR	Recipe using this.

Table 1: User-spoken Queries in Part 1 of the Study

Task	P#	Query
Math Task	P1	Is this equation correct?
	P2	Did I do this equation right?
	P3	Is this correct?
	P4	Is this equation correct?
	P5	What's the answer to this equation?
	P6	Is this equation correct?
	P7	Is this correct?
	P8	Is this correct?
	P9	Is this equation correct?
	P10	Is this correct?
	P11	Tell me if this is correct.
	P12	Is this correct?
Price Difference Task	P1	How much do these cost?
	P2	Which of these is more expensive, and by how much?
	P3	Can you compare the price between these two?
	P4	What's the price difference between these two items?
	P5	Find me the difference in costs between these two items.
	P6	What's the price difference between these?
	P7	What is the difference in price between these?
	P8	How much is the price difference between these?
	P9	What's the price difference between them?
	P10	What's the price difference between these?
	P11	Tell me the price difference between them.
	P12	What's the price difference between these?
Celebrity Task	P1	Who is she?
	P2	Can you tell me more information about her?
	P3	Who is this person?
	P4	Who is this person?
	P5	Find me more information about this person.
	P6	Tell me about her.
	P7	Who is this?
	P8	Find me more information about her.
	P9	Who is she?
	P10	Who is she?
	P11	Tell me more about her.
	P12	Who is this?

Table 2: User-spoken Queries in Part 2 of the Study

P#	Query	Satisfactory?
P1	Tell *me* the price difference between *this* and *that*.	No
P2	Did *I* solve *this* [complex calculus problem] correctly?	No
	How far away am *I* from *my* house?	No
P3	Which trash bin does *this* [trash] go into?	Yes
	Can *I* put *this* [trash] in any of *these* [trash bins]?	Yes
	Can *I* put *this* [trash] in *here* [recyling trash bin]?	No
P4	Can *you* explain *that* [diagram in a classroom]?	No
	What can *I* use to clean *this* [stain on a surface]?	No
	Can *you* translate what *she* [foreigner] is saying?	No
P5	A child constantly asking "What's *this*?"	Yes
	A blind person asking "What did *s/he* [speaker] point to?"	No
	Tell *me* the price difference between *this* and *this*.	No
	What is in the box sitting on top of *that* chair?	No
P6	Tell *me* more about *that* building.	No
	Who made *that* [car]?	No
	What species is *this* [plant]?	No
	A blind person can ask "Who is *that* on TV?"	Yes
	Who wrote *it* [book]? Tell *me* more about the author.	Yes
	Who are *those* people?	No
	What's happening over *there*?	Yes
	What is the object next to *that* [an object I know]?	No
P7	How do *you* pronounce *this* [foreign or complex word]?	Yes
	How do *you* translate *this* [foreign or complex word]?	Yes
P9	Who is *s/he* [person]?	Yes
	What is *his/hers* [musician] top hit?	Yes
P10	Who is left of *him/her*?	No
	What do *they* sell?	Yes
P11	What does *this* mean [foreign language]?	Yes
P12	Can *you* tell me more about *this* [unknown objects]?	Yes
	Who was *s/he* [celebrity] again?	No
	Compare *this* to *that* thing from before [an object I saw a few seconds ago].	No
	When did *I* do *this* [activity]?	No

Table 3: User-spoken Queries in Part 3 of the Study

Location	Query	Satisfactory?
Home	What do *I* need to do today?	No
	My plant seems to be dying. What can *I* do for *it*?	Yes
	I’m done with *this*. Where can *I* buy another one?	Yes
	What should *I* eat today?	Yes
	I want to grab coffee on the way... suggest *me* a cafe nearby.	No
	What’s the best settings on *this* [coffee] machine?	No
	Does *this* look spoiled?	No
	I’m looking for *my* keys... where did *I* leave *it* again?	No
	Can *you* let *me* know when *you* find *my* keys?	No
	I stained *this* clothing... how can *I* remove *it*?	No
	Did *I* take *this* vitamin today?	No
	Did *I* turn off the stove?	No
Work	What’s *my* agenda for today?	No
	I want to eat something light before *my* commute... can *you* suggest *me* a place?	No
Gym	How can *I* use *this* equipment?	No
	I’m working on legs today... can *you* recommend a workout plan?	Yes
	What should *I* eat post workout?	Yes
Cafe	How well rated is *this* coffee shop?	Yes
	Can *you* recommend *me* something from *here*?	Yes
	After *I*’m done, where should *I* toss *this*?	No
	Is *this* decaf?	No
Restaurant	Can *you* recommend something from *here*?	Yes
	I’m allergic to *that*... can *you* recommend something else?	Yes
	I love *this* dish! How can *I* make *this* from home?	No
Shopping Mall	What’s *this* store known for?	Yes
	Which of *these* stores should *I* visit?	Yes
	I love *this* cloth. Who designed *it*?	Yes
	It’s a bit outside *my* price range... can *you* recommend *me* a similar brand?	Yes
Library	*I* really love *this* book. Can *you* recommend another book by the same author?	Yes
	What’s one latest book *you* can recommend that *I* read?	Yes
Movie Theater	Is *this* a good movie?	Yes
	Tell *me* the history behind *this* scene	No
Grocery Store	What should *I* cook today?	Yes
	What’s for sale today?	No
	Which is the healthier option? *This* or *this*?	Yes
	Anything *I*’m missing *here*?	No
Hospital	Pull up *my* appointment details.	No
	Do *I* have to be anywhere after *this*?	No
Park	What [dog] breed is *s/he*?	No
	Can *you* tell *me* more about *that* plant?	No
	Can *I* buy *this* plant from somewhere?	No
University Campus	*I*’m supposed to meet a friend from [location]. How do *I* get *there*?	No
	When was *this* building built?	Yes
Public Transit Station	*I*’m trying to get to [location]. Is *this* the bus *I* should take?	No
	Where should *I* go from *here*?	No
Sidewalk	*That* was a really cool car! Tell *me* more about *it.*	No
	When does *that* store close?	Yes
	Is *this* the right place [store I am trying to reach]?	No

Table 4: User-spoken Queries in the First-Person Diary Study