Walkie-Talkie: Exploring Longitudinal Natural Gaze, LLMs, and VLMs for Query Disambiguation in XR

Jaewook Lee
University of Washington
Seattle, Washington, USA
jaewook4@cs.washington.edu

Tianyi Wang
Reality Labs Research
Meta Inc.
Redmond, Washington, USA
tianyiwang@meta.com

Jacqui Fashimpaur
Reality Labs Research
Meta Inc.
Redmond, Washington, USA
jacquiwithaq@meta.com

Naveen Sendhilnathan
Reality Labs Research
Meta Inc.
Redmond, Washington, USA
naveensn@meta.com

Tanya R. Jonker
Reality Labs Research
Meta Inc.
Redmond, Washington, USA
tanya.jonker@meta.com

Figure 1: An example interaction with Walkie-Talkie v1 in VR. A user asks "What's the healthiest among these?" as they shift their gaze to look at different items on the shelf, including beer, Coca-Cola, water, and Sprite. An LLM receives a 30-seconds gaze log, word utterance timing, and the user's original query as part of a hard-prompt. The answer is read aloud.

Abstract

Everyday conversations are often ambiguous, which we resolve using nonverbal cues like gaze and pointing. To enable such low-effort interactions with voice assistants (VAs), we explore how large language models (LLMs) and vision-language models (VLMs) can leverage longitudinal natural gaze signals. We introduce *Walkie-Talkie*, a multimodal VA for extended reality (XR) that uses gaze dynamics to disambiguate queries. Through iterative design, our system transforms gaze data—capturing targets, duration, and spatial relationships—into text and/or image for LLM and VLM processing alongside spoken word timing. In a controlled VR study (N=12), Walkie-Talkie outperformed a baseline requiring explicit gaze input, and subsequent AR evaluations explored its real-world feasibility. Our findings highlight the potential of multimodal foundation models for natural, accurate gaze and speech interactions in wearable XR. We conclude by discussing future directions for designing always-available, context-aware AI agents in XR glasses.

CHI EA '25, Yokohama, Japan

ACM ISBN 979-8-4007-1395-8/25/04
https://doi.org/10.1145/3706599.3720236

CCS Concepts

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

Keywords

extended reality, natural gaze input, multimodal input, voice assistants, LLM, VLM

ACM Reference Format:
Jaewook Lee, Tianyi Wang, Jacqui Fashimpaur, Naveen Sendhilnathan, and Tanya R. Jonker. 2025. Walkie-Talkie: Exploring Longitudinal Natural Gaze, LLMs, and VLMs for Query Disambiguation in XR. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3706599.3720236

1 Introduction & Background

Everyday conversations are filled with ambiguities—like saying to a friend, "*I really like that, where and how much did you buy it for?*"—which we resolve naturally using nonverbal cues such as eye gaze and pointing [2, 6, 10]. To replicate these interactions in machines, recent work has leveraged advances in extended reality (XR), computer vision (CV), and multimodal foundation models to develop context-aware voice assistant (VA) prototypes, including *WorldGaze* [24], *Nimble* [33], *TouchVA* [18], *GazePointAR* [19], and

G-VOILA [37]. While promising, these systems remain proof-of-concept prototypes, relying on controlled evaluations and explicit gaze dwell, limiting their use in natural, spontaneous interactions.

Explicit gaze input, while effective for tasks like object targeting [31], poses challenges for question-answering, including the midas touch problem and user fatigue, as users must consciously control their gaze against natural tendencies. In contrast, natural gaze dynamics offer significant benefits, as gaze behavior reflects cognitive processes like attention [30] and decision-making [13]. Additionally, gaze over time enables ambiguous queries referencing objects or events in the recent past (*e.g.,* "*Which is healthier, this or that?*" or "*What was that?*") and even the distant past (*e.g.,* "*Where did I leave my keys?*"). While prior systems have focused on explicit gaze input and present referents (*e.g.,* "*What is this?*"), we emphasize supporting queries tied to both the present and recent past through low-effort, natural gaze interactions.

We present *Walkie-Talkie*, a context-aware VA that leverages *longitudinal natural eye gaze*, large language models (LLMs), and vision-language models (VLMs) for query disambiguation in XR environments. Unlike prior research, Walkie-Talkie captures gaze signals over time, recognizing intent without requiring users to alter their typical gaze behavior. The first iteration, *Walkie-Talkie v1*, integrates a GPT-4 [27] LLM, transforming raw gaze and speech data into structured text by encoding 30 seconds of gaze history (*e.g.,* "*gazed at apple for 0.4 seconds (nearby objects: melon, kiwi)*"), spoken word timing (*e.g.,* "*said 'this' 0.2 seconds ago*"), and the query itself into a hard prompt for processing (Figure 2). With Walkie-Talkie v1, our goal is to explore LLMs' potential to interpret lower-level multimodal human data.

To evaluate Walkie-Talkie, we first deployed it on a *Meta Quest Pro* virtual reality (VR) headset to assess its natural interactions and query disambiguation accuracy compared to explicit gaze input systems, without confounds such as CV performance or system latency. In a within-subjects, two-part lab study (N=12), participants completed free-form queries (Part 1) and researcher-defined tasks (Part 2), comparing Walkie-Talkie to a custom baseline system modeled after prior work that relied on explicit gaze input. Results showed a strong preference for Walkie-Talkie due to its fluid gaze interaction (*e.g.,* participants exhibited fewer fixations and more saccades) and effective query disambiguation, particularly for complex queries involving multiple, small, or distant referents.

Expanding beyond controlled VR environments, we adapted Walkie-Talkie for the *Microsoft HoloLens 2* augmented reality (AR) glasses to assess its real-world performance. This included two iterations: *Walkie-Talkie v2*, which replicated the VR setup using *YOLO-World* [8], an open-vocabulary object detector, alongside a *GPT-4o* LLM [28]; and *Walkie-Talkie v3*, which integrated *RepViT-SAM* [36], a fast *Segment Anything Model (SAM)* [16], with point input (*i.e.,* projected gaze coordinates) and a GPT-4o VLM [28]. After identifying CV limitations in Walkie-Talkie v2, we improved Walkie-Talkie v3 by incorporating a mosaic of stitched gaze-driven crops for better gaze target classification. Then, five participants tested it across three real-world environments (*i.e.,* grocery store, library, and home), providing qualitative feedback. We found that this approach enables accurate, lightweight referent identification, though some misclassifications remained. We conclude by envisioning future always-available AI agents in wearable XR.

In summary, our key contributions include:

(1) An exploration of multimodal foundation models for processing longitudinal gaze data, showcasing prompt-based and image-based approaches to understanding gaze dynamics.
(2) Walkie-Talkie, a context-aware VA for XR headsets, leveraging natural gaze for low-effort query disambiguation.
(3) Insights from VR and AR evaluations, including a two-part lab study and real-world deployment, revealing challenges and opportunities for gaze- and AI-driven XR systems.

2 The Design of Walkie-Talkie

Walkie-Talkie is a novel context-aware multimodal voice assistant (VA) for wearable XR that leverages gaze and speech over time. With Walkie-Talkie, our goal is to enable users to gaze naturally at their surroundings and ask ambiguous queries instinctively. For example, a user might scan items on a shelf and ask, "*Which is the healthiest?*", allowing Walkie-Talkie to infer context from their everyday gaze and speech behavior. In its first iteration, *Walkie-Talkie v1*, we encode gaze and speech features into text and integrate them into a hard prompt for LLM processing. We built Walkie-Talkie v1 in *Unity 2021.3.15f1*[1] using the *Meta XR SDK*[2], with the *Meta Quest Pro*'s built-in eye tracker enabling rapid prototyping. Figure 2 provides an overview of the system architecture.

System Activation. Since the Quest Pro lacks voice activation, participants initiated queries by saying "*Hey glass*", after which the researcher activated Walkie-Talkie via a laptop key press, simulating state-of-the-art VAs (*e.g.,* Apple Siri).

Gaze and Speech Capture. The Quest Pro's eye tracker (1.652 ± 0.699 degrees error [38]) continuously logged users' natural gaze behavior, storing the most recent 30 seconds of fixations over 50ms [17] (*e.g.,* "*Gazed at apple for 0.4 seconds*"), as well as saccades or unfixated gazes (*e.g.,* "*Gazed at nothing for 0.6 seconds*"). This structure emphasizes key spatiotemporal features—gaze duration, target, and order [29]—for LLM interpretation. Then, to account for gaze tracking inaccuracies, each entry included up to five nearby objects, ordered by proximity to the tracked gaze position within a five-degree cone (*e.g.,* "*(Nearby objects: melon, kiwi)*"). Speech was transcribed via *Wit.ai*[3], with word utterance timings logged (*e.g.,* "*Said 'this' 0.2 seconds ago*") to help the LLM align speech with gaze.

Data Processing and Query Response. Gaze and speech data were processed using *GPT-4* [27] with a custom-designed hard prompt (Figure 2). Responses were read aloud via text-to-speech.

3 User Study

We conducted a two-part within-subjects lab study with 12 participants to evaluate Walkie-Talkie's ability to process longitudinal natural gaze data for query disambiguation, comparing it to a custom baseline system. The study focused on accuracy (*i.e.,* how well each system predicted speech referents) and naturalness (*i.e.,* how well participants could maintain natural gaze behaviors). While both systems employ a prompt-based approach, the baseline relied on explicit gaze input through a heuristic method inspired by prior

[1] https://unity.com
[2] https://assetstore.unity.com/packages/tools/integration/meta-xr-all-in-one-sdk-269657
[3] https://wit.ai

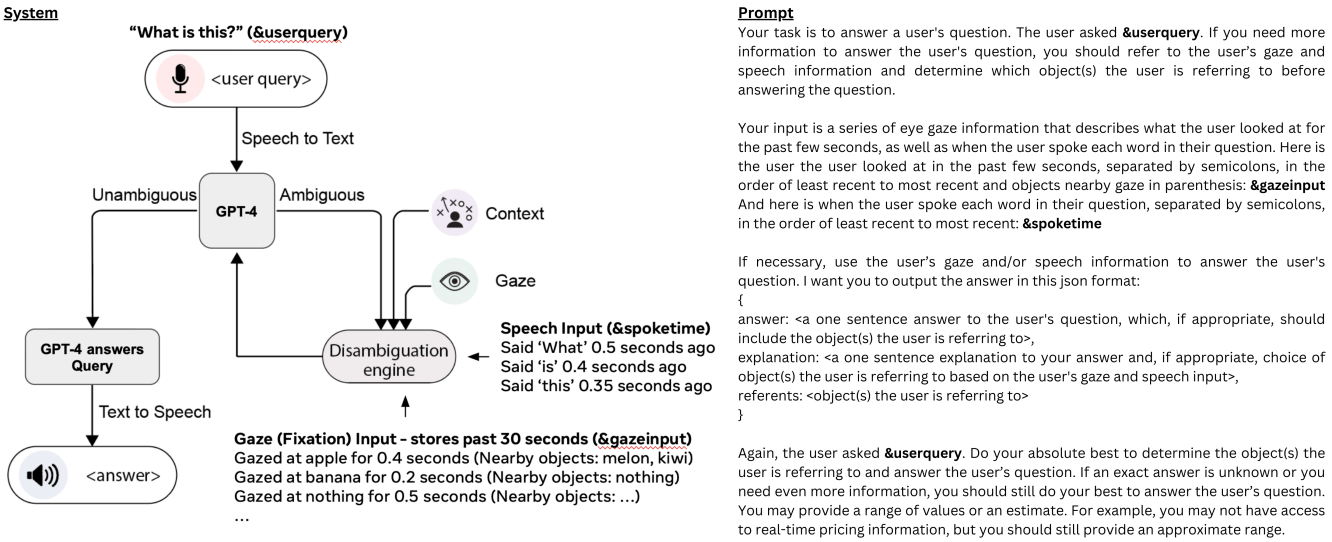
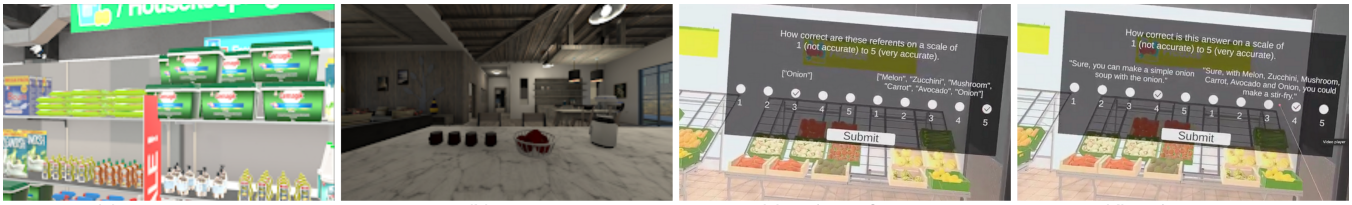


Figure 2: System overview and hard-prompt design of Walkie-Talkie v1. The system captures users' natural gaze over time and compiles a 30-second gaze history, including targets, durations, and nearby objects (*e.g.,* "*gazed at apple for 0.4 seconds (nearby objects: melon, kiwi)*"). This is combined with the timing of spoken words (*e.g.,* "*said 'What' 0.5 seconds ago*") and the query itself, forming a comprehensive hard prompt sent to an LLM (*i.e.,* GPT-4) for processing.



(a) Grocery Store (b) Home (c) Rating Referents (d) Rating Answers

Figure 3: Components of the controlled VR study: (a) VR grocery store environment, (b) VR home environment, (c) referent-rating UI panel, where users rated the accuracy of predicted referents for the baseline and Walkie-Talkie simultaneously on a 5-point Likert scale, and (d) answer-rating UI panel, which served a similar purpose but focused on response accuracy.

work [2, 19, 24, 33], capturing gaze data either when a keyword was spoken (*e.g.,* commonly-spoken pronouns like "*this*" [20]) or at the end of the query. The study was conducted in VR to control conditions and eliminate confounding factors like object recognition errors. Participants (7 female, 5 male, $Mean_{age}$=28.8 years, SD_{age}=5.5) had some prior experience with XR, VA, and AI chat systems, having used each technology a few times but not regularly.

Part 1: Free-Form Queries. Part 1 evaluated Walkie-Talkie and the baseline system's ability to predict speech referents for spontaneous queries, the quality of their responses, and how participants formed ambiguous queries. For 15 minutes, participants explored a VR grocery store (Figure 3a) and posed queries freely, with both systems generating responses simultaneously. The predicted referents were first displayed side by side and rated on a 5-point Likert scale (Figure 3c), followed by the answers (Figure 3d). To reduce bias and encourage natural gaze behavior, participants were told only that their gaze was being tracked, with the baseline labeled as "*System A*" and Walkie-Talkie as "*System B*". Afterward, participants answered open-ended questions about their experience, perceived system accuracy, and approaches to gaze and query formulation.

Part 2: Researcher-Defined Tasks. Part 2 evaluated system usability and the impact of different gaze processing strategies on participants' natural gaze behavior, preferences, and system accuracy. Participants completed 12 tasks using each system separately across a VR grocery store (Figure 4a) and home environment (Figure 4b), with counterbalanced system and environment order. We designed tasks with varying complexity, ranging from those requiring a single referent to those involving multiple or past referents. Participants were encouraged to make multiple attempts until satisfied with the system's answer, allowing them to refine their gaze behavior. Metrics included 5-point Likert scale ratings for referent and response accuracy, raw gaze data, and post-task questionnaire responses using SUS [5] and NASA-TLX [15]. Follow-up interviews further explored participants' perceptions of accuracy, naturalness, and their own gaze behavior across the two systems.

Data Analysis. Quantitative data were analyzed using Wilcoxon signed-rank tests for paired comparisons and Mann-Whitney U tests when the number of queries differed between conditions. Qualitative data were analyzed using reflexive thematic analysis [3, 4].

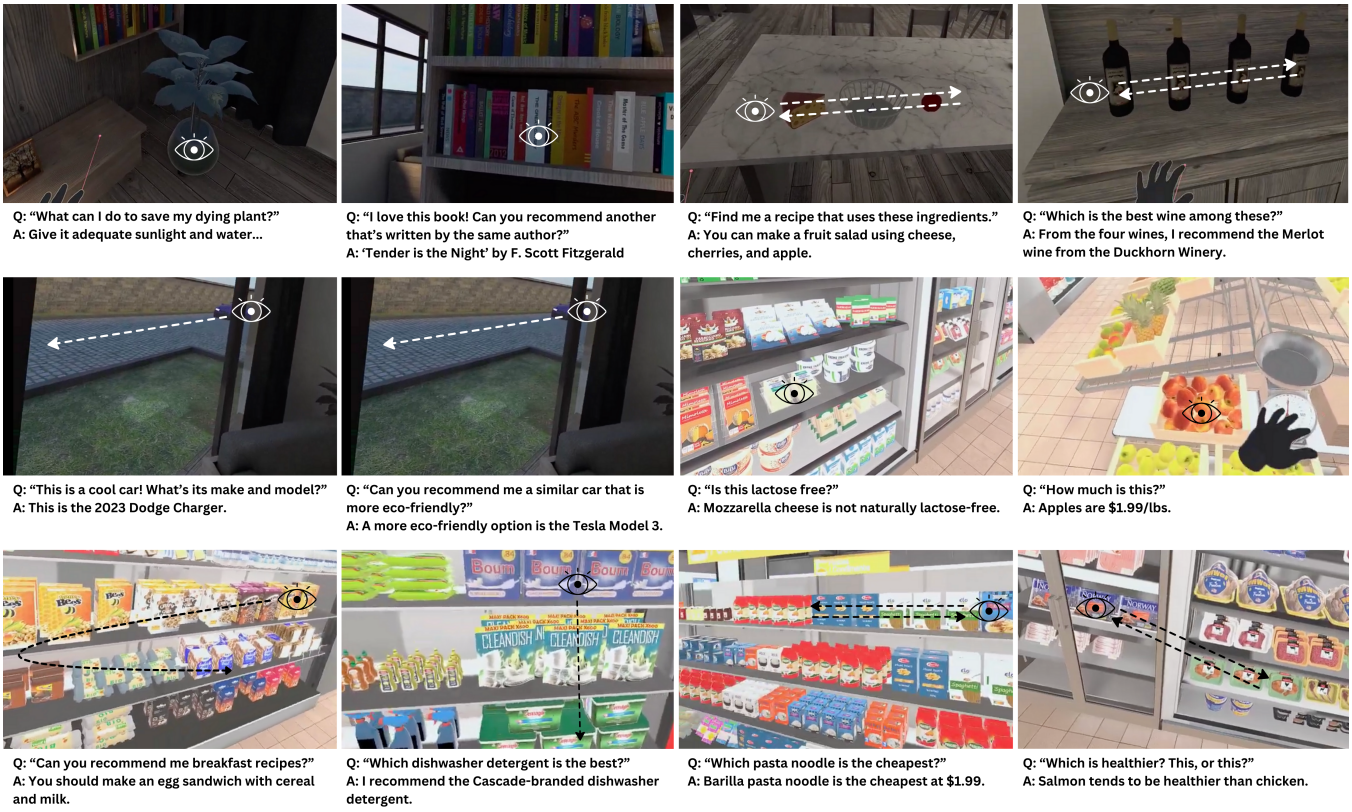


Figure 4: Example queries asked when completing the 12 researcher-defined scenarios in Part 2 of the VR study.

4 Results

We analyzed data from *593 query attempts*—305 spontaneous queries from Part 1 and at least 288 (2 systems x 12 tasks x 12 participants) from Part 2. In Part 2, we used all attempts for referent and answer accuracy, as they reflected system performance across retries, but only the final satisfactory attempt for gaze data, as it captured participants' most refined gaze input. Overall, Walkie-Talkie outperformed the baseline in accuracy, gaze naturalness, and usability.

4.1 Perceived Accuracy

Walkie-Talkie showed significantly higher accuracy than the baseline in Part 1, both for referent prediction ($V = 73$, $p < 0.001$) and answer accuracy ($V = 588$, $p < 0.001$). Similarly, in Part 2, it outperformed the baseline in referent prediction ($W = 14,779$, $p < 0.001$) and answer accuracy ($W = 16,140$, $p < 0.001$). Usability ratings (SUS) were initially not significant ($V = 16.5$, $p = 0.08$), but after removing outliers, they favored Walkie-Talkie ($V = 5.5$, $p < 0.05$). Cognitive load measures (NASA-TLX) showed no significant differences ($V = 51.5$, $p = 0.35$), even after outlier removal ($V = 40.5$, $p = 0.53$). See Figure 5 for boxplots of all collected metrics.

Participant feedback aligned with these results, with 11 of 12 participants preferring Walkie-Talkie's accuracy. The baseline sporadically returned "*unsure*" (8/12), missed key referents or failed to identify any (7/12), and misclassified gaze targets as larger nearby objects like "*shelf*" or "*table*" (3/12). P9 explained, "*My gaze often drifts, so I may unintentionally look between nearby objects and the object I actually care about, like the shelf holding the cereal boxes,*" which made it harder for the baseline to capture precise referents, especially when they were smaller or further away. P11, who preferred the baseline's cautious approach, reasoned, "*The baseline system is always right when it says something, though it sometimes doesn't say anything... I prefer an AI that is absolutely correct, even if it responds less often.*"

For single-referent queries (*e.g.,* "*Is this healthy?*"), half of participants found little difference between systems, while the other half preferred Walkie-Talkie for its accuracy. This is because the baseline sometimes missed intended targets, even when participants tried to keep their gaze steady, causing some to wonder if they had "*moved [their] eyes too early*" (P3). For multi-referent tasks, Walkie-Talkie was widely preferred, with eight participants praising its ability to handle complex comparisons (*e.g.,* "*Which is healthiest among these?*"). However, five participants noted Walkie-Talkie sometimes included unintended referents; for example, P2 observed, "*When I asked, 'Which fruit has the highest vitamin C content?' Walkie-Talkie compared six or seven fruits when I really meant just three.*"

4.2 Gaze Behavior

In Part 2, we collected and processed raw gaze data to compare participants' gaze behavior between Walkie-Talkie and the baseline, following prior work [35]. We first transformed 3D gaze positions

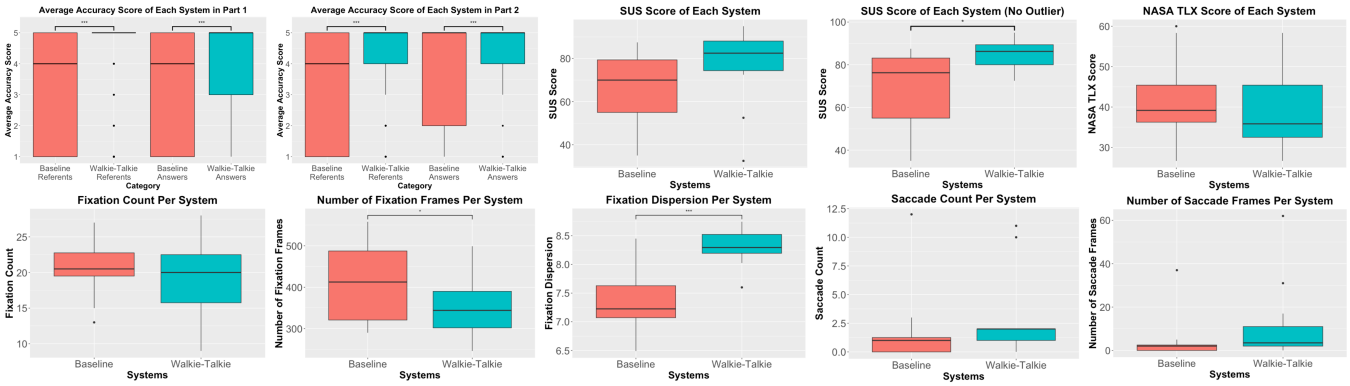


Figure 5: Box plots comparing the Baseline and Walkie-Talkie systems across all collected quantitative metrics, including Part 1 & 2 referent and answer accuracy, SUS (with and without outliers), NASA-TLX, fixation/saccade counts, fixation/saccade frame durations, and fixation dispersion. Statistically significant differences are indicated in relevant plots.

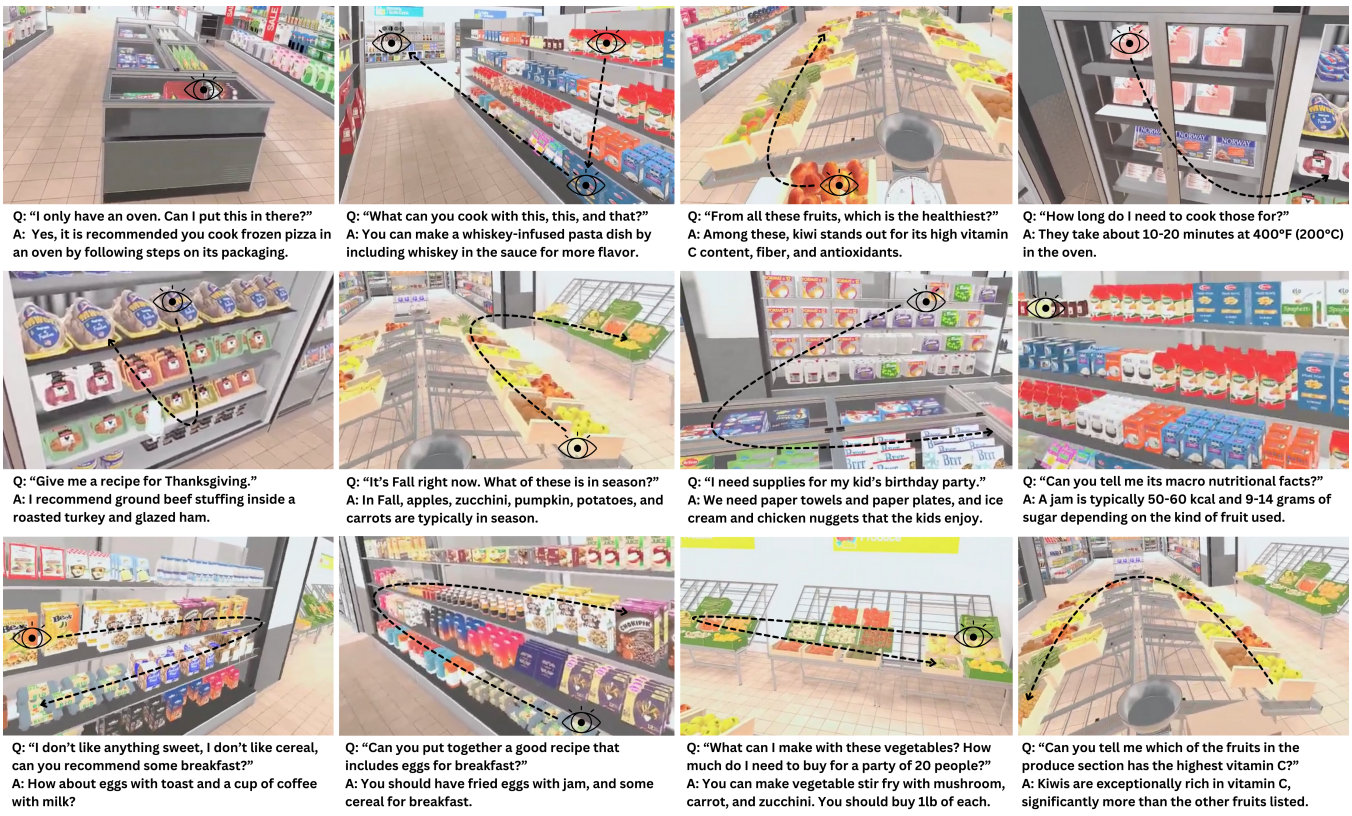


Figure 6: 12 examples of free-form queries participants asked in Part 1 of the VR study.

from the eye-in-head reference frame to gaze-in-world using head orientation data [7]. Gaze velocity was then computed as the angular displacement between consecutive samples divided by the time interval, with velocities exceeding 800°/s filtered as noise [11]. Missing data were linearly interpolated. Saccades were detected using the IN-VT method for samples exceeding 70° and lasting 17–200 ms [34]. Fixations were identified using the I-DT method, where time windows with dispersion under 1° and durations between 50–1500 ms were labeled as fixations [34].

Quantitative gaze analysis revealed that participants exhibited more natural gaze behavior with Walkie-Talkie than the baseline (Figure 5). Fixation durations (*i.e.,* total frames spent fixating) were significantly shorter ($V = 61.5$, $p < 0.05$), and fixation dispersion (*i.e.,* the average distance between fixation points and the fixation

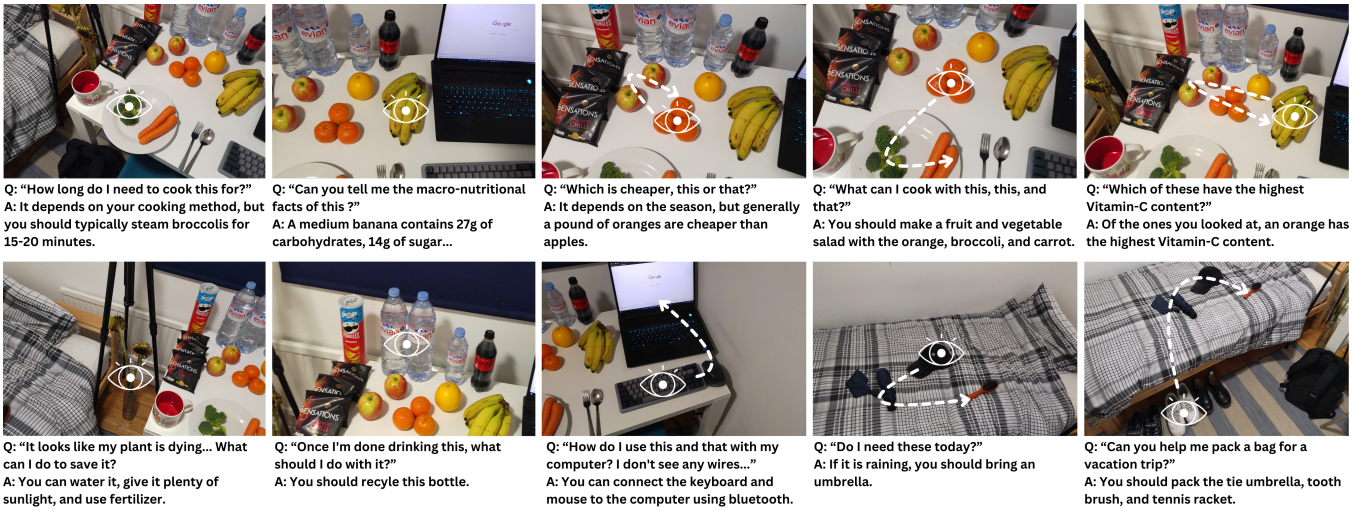


Figure 7: 10 example queries the first author asked to Walkie-Talkie v2 in AR.

centroid [12]) was significantly wider ($W = 11$, $p < 0.001$), suggesting Walkie-Talkie encouraged more dynamic and exploratory gaze patterns. Additionally, saccade durations (*i.e.,* total frames spent saccading) showed a trend toward being longer with Walkie-Talkie but was not statistically significant ($W = 44$, $p = 0.10$). Other metrics, including average gaze velocity ($W = 66$, $p = 0.76$), fixation counts ($V = 49.5$, $p = 0.43$), and saccade counts ($V = 9.5$, $p = 0.25$), did not differ significantly between conditions. See Figure 5.

Participants described their gaze as more dynamic with Walkie-Talkie, using terms like "*more natural*" (P1, P2, P3, P5, P6, P11), "*scanning*" (P3, P7, P12), "*continuously moving*" (P6, P8), "*freer*" (P3, P12), "*lasso-like looping*" (P5, P10), "*sweeping*" (P11), "*gliding*" (P10), "*jumping*" (P2), and "*bouncing*" (P1). P8 and P9 mentioned that they "*gave almost no thought into how to look at things*" with Walkie-Talkie, due to its higher accuracy. However, several participants noted their eye movements were not entirely natural, as they adjusted to assist Walkie-Talkie in understanding their intentions (P3, P4, P12). Some also reiterated concerns about Walkie-Talkie over-predicting referents, particularly in cluttered scenes.

4.3 Query Formation

Of the 305 free-form queries in Part 1 (see Figure 6 for examples), 238 (78.0%) contained pronouns, with "*this*" (136/238), "*that*" (41/238), and "*these*" (24/238) being the most common. This aligns with prior research on the prevalence of pronouns in ambiguous speech [10, 20]. Participants found ambiguous queries, especially those with pronouns, instinctive and natural. P5 observed they spoke more concisely to Walkie-Talkie, which consistently tracked their gaze, compared to people who might miss context: "*I'd just say 'this cereal' instead of explaining further, like 'the cereal down below.'*" However, P12 noted that the system's lack of personal context sometimes required more detailed queries: "*People are already aware of my goals, like eating healthy, so I don't have to say, 'What's healthiest among these?' to my friends. Instead, I'd say something like 'what's good?' if we're at a restaurant trying to order.*"

5 System Evaluation in Wearable AR

In our VR study, we demonstrated that multimodal foundation models, particularly LLMs, show promise in processing longitudinal natural gaze data to resolve ambiguous queries in controlled conditions. However, in real-world scenarios, factors like latency and CV limitations can affect Walkie-Talkie's performance. To assess its behavior under these constraints, we deployed the system on a *Microsoft HoloLens 2*, iterating twice to refine its capabilities. This evaluation offers insights into designing and deploying context-aware, always-available AI agents for wearable XR devices.

5.1 Walkie-Talkie v2: LLMs + YOLO-WORLD

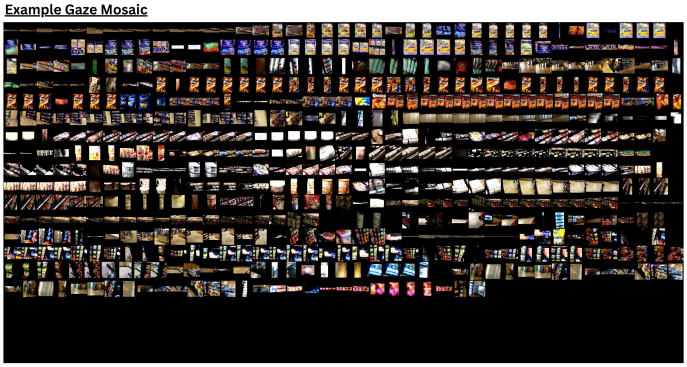
The first AR iteration of Walkie-Talkie on the HoloLens 2 (*Walkie-Talkie v2*) closely mirrors the Quest Pro implementation (Walkie-Talkie v1), with additional features inspired by *GazePointAR* [19].

System Implementation. Walkie-Talkie v2 streams 424x240 @ 30FPS video feed using *hl2ss* [9] to an external PC running *YOLO-World* [8], a real-time open-vocabulary object detector. The HoloLens' built-in gaze tracker (1.5–3° error margin [26]) maps the user's gaze coordinates to detected objects through projection. Then, gaze targets and speech data over time are integrated into a hard prompt for a *GPT-4o* LLM [28], as in the VR system (Figure 2). To activate Walkie-Talkie, users say "*Hey Glass*"[4] and the HoloLens responds, "*Hi, I'm listening.*" [19]. Speech is converted to text and word timings via *Microsoft Azure*'s speech-to-text API [25], while the gaze log continuously updates at 23.8 FPS. LLM responses are vocalized with Azure's text-to-speech API [25].

Preliminary Evaluation. The first author tested Walkie-Talkie v2 in 10 real-world scenarios, such as comparing fruit prices, finding recipes, and interacting with household items (see Figure 7). Walkie-Talkie v2 supported fluid eye movements and resolved several ambiguous queries but struggled with object detection limitations, including misclassifications (*e.g.,* confusing a white table for

[4]https://learn.microsoft.com/en-us/windows/mixed-reality/mrtk-unity/mrtk2/features/input/speech

Example Gaze Mosaic



Prompt

Your task is to answer a user's query clearly and directly. The user's query may be ambiguous, in which case you should answer it based on their gaze history provided as a mosaic image.

The mosaic image represents the user's gaze history over time, up to **{MAX_TIME}** seconds. Each tile corresponding to **{SAMPLE_INTERVAL}** seconds of gaze input, arranged temporally from top-left (oldest gaze) to bottom-right (most recent gaze). Black areas act as borders or indicate unused tiles if fewer than **{MAX_TIME}** seconds of gaze were recorded.

The user's query may refer to object(s) or event(s) at any point in the gaze history. Use the gaze mosaic to identify the object(s) or event(s) the user is likely referring to, then provide an accurate, concise response to their question. Timestamps for the spoken words in the query will help align the user's speech with their gaze. When answering, you must:
1. Use plain, natural, and conversational language that is easy to understand.
2. Provide a direct answer that explicitly identifies the objects or events referenced.
3. Avoid mentioning the gaze mosaic, analysis, or reasoning process. Focus solely on the user's query and the objects/events.

Here is a few example queries and responses:
Query: "What is this?"
Response: "This is a red apple."
Query: "What is this machine used for?"
Response: "This is a coffee machine, and it can make a delicious cup of espresso."
Query: "Which is healthier, this or that?"
Response: "The orange is healthier than the banana due to its higher Vitamin C content."

{QUERY} To reiterate, your response must be one clear, conversational sentence that directly answers the user's query. Do not describe your reasoning or analysis. Just give the answer.

Figure 8: Example gaze mosaic representation of a user's gaze history in a local grocery store, alongside the updated hard prompt design of Walkie-Talkie v3.

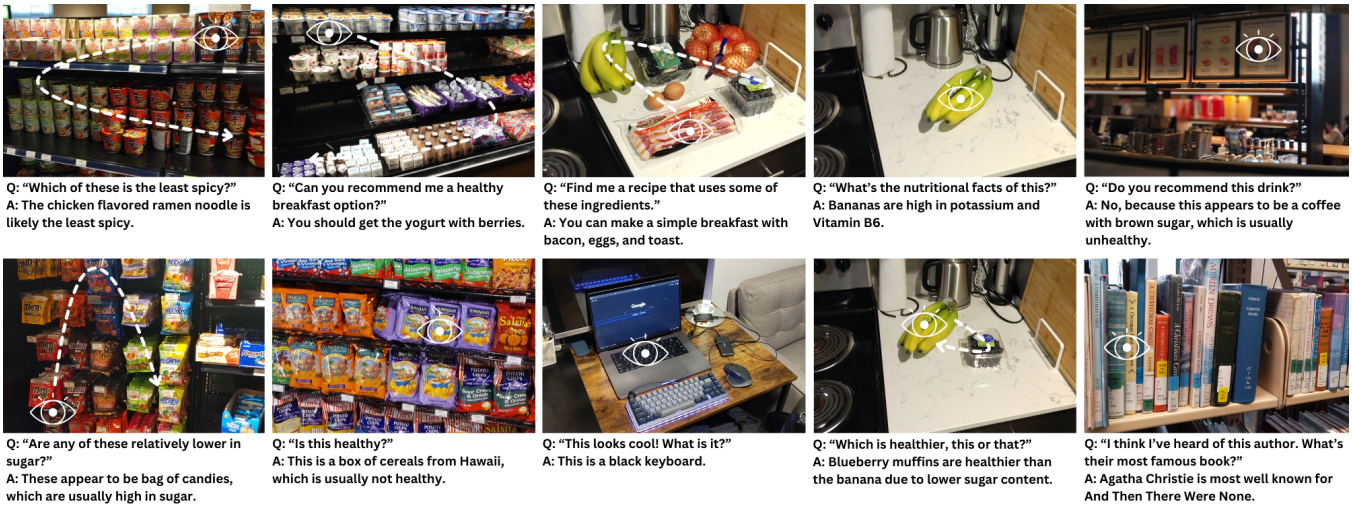


Figure 9: 10 example queries participants asked to Walkie-Talkie v3 in AR.

a refrigerator) and numerous unrecognized objects. We concluded that current efficient open-vocabulary object detectors alone may be insufficient for the complexity of real-world environments.

5.2 Walkie-Talkie v3: VLMs + Gaze Mosaic

While the LLM and open-vocabulary object detector approach showed promise, it struggled to handle the diversity of objects in real-world environments. To address this, we explored using a vision-language model (VLM) to improve Walkie-Talkie's gaze target classification capabilities (*Walkie-Talkie v3*).

System Implementation. We enhanced Walkie-Talkie by incorporating a VLM for better open-vocabulary object detection, while keeping the same user flow as previous versions. To process gaze data and the video feed efficiently without exceeding token limits, we implemented a gaze-driven cropping strategy using *RepViT-SAM* [36], a lightweight variant of the *Segment Anything Model (SAM)* [16]. The user's projected gaze coordinate is passed as input to generate polygonal crops from the video stream, which are then converted to rectangular crops every 0.2 seconds and stored for 15 seconds. When the user speaks, these crops form a chronologically-ordered 4096×8192 mosaic, with the oldest gaze target in the top left and the most recent towards the bottom right. If no gaze data is available—due to lack of fixation (*e.g.,* gaze tracker failure, closed eyes), SAM failures, or having gazed for less than 15 seconds—black rectangles may populate parts of the gaze mosaic. This mosaic is integrated into a hard prompt alongside speech data, including word utterance timings, and processed by *GPT-4o*'s vision feature [28] (Figure 8). Gaze data and the video feed are sent via hl2ss, with text data transferred using a custom Flask-based TCP solution. The response is vocalized by the HoloLens.

User Study and Findings. Five participants (PP1–PP5, 3 female, 2 male, $Mean_{age}$=27.2 years, SD_{age}=4.1) evaluated Walkie-Talkie v3 across three environments: grocery store, library, and home. Participants asked 67 free-form queries in total (Figure 9), and Walkie-Talkie v3 successfully handled many ambiguous queries (41/67), demonstrating a significant improvement over the previous LLM-based approach. The VLM's ability to align gaze-driven crops with speech timing was particularly promising, with nearly

all queries resulting in at least some predicted referents. However, CV limitations remained: misclassifications, such as "*a case of blueberries*" being misidentified as "*a blueberry muffin*" (PP3) and "*a bag of chips*" as "*a box of cereal*" (PP1), as well as small crops leading to inaccuracies, like misidentifying a book's author due to poor crop resolution (PP4). Additionally, rapid head movements caused inaccurate or occasional "*unsure*" responses (PP1, PP2, PP5) due to blurred crops. Despite these issues, participants were impressed with Walkie-Talkie v3, noting that even in misclassification instances, predicted referents were reasonable and aligned with the intended targets (*e.g.,* "*a green box*" predicted as "*a dollar bill*" (PP5)). PP1 was particularly impressed that Walkie-Talkie v3 not only recognized their referent as an instant noodle but also distinguished between different types and brands, correctly answering their query, "*Which of these is the least spicy?*" Lastly, participants felt their gaze was natural; however, like in the VR study, they noted it was not entirely natural due to awareness of being tracked.

Reflections on Walkie-Talkie v3. Overall, using a VLM with gaze-driven crops is a promising approach to resolving ambiguous queries, as it balances accuracy, latency, and input data efficiency for foundation models. Compared to an LLM paired with a separate object detector (Walkie-Talkie v2), this method achieved higher accuracy while maintaining a reasonable response time ($Mean$=10.1s) and reducing the need for video-based inputs. That said, with advancements in *retrieval-augmented generation (RAG)* [22, 23] and real-time multimodal models (*e.g., Google Gemini*'s Multimodal Live API [1]), our approach may serve as a stepping stone rather than a final solution. Given the overhead of these emerging techniques—including data storage, relevance filtering, and large-scale data transmission—our crop-based approach provides a lightweight alternative as foundation models continue to evolve.

6 Discussion and Conclusion

Walkie-Talkie explores the use of LLMs and VLMs to process longitudinal natural gaze data for query disambiguation. Results from our two-part VR study show that multimodal foundation models can enable more accurate, natural gaze-based interactions. AR evaluation demonstrates Walkie-Talkie's real-world feasibility, though challenges like over-predicting referents and misclassifications remain. Below, we discuss future opportunities and areas for improvement.

Future of Always-Available AI Agents. Walkie-Talkie leveraged frozen multimodal foundation models, which performed surprisingly well despite likely not being trained on gaze data. Its performance could be further improved through fine-tuning [41] or prompt-tuning [21]. To achieve this, a dataset capturing natural gaze behavior, along with other human inputs like gestures and speech, and ground truth intent, must be assembled, similar to *VOILA-A* [40]. Additionally, future systems should explore whether LLMs and VLMs can process other common human inputs, such as hand gestures, to enhance interaction flexibility [10, 19]. Moreover, systems should support not only present and recent past referents but also those from the distant past, enabling queries like "*Where did I leave my keys?*". Lastly, techniques like retrieval-augmented generation (RAG) [22, 23] and real-time models (*e.g.,* Gemini's Multimodal Live API [1]) could help overcome these limitations, enabling AI to retrieve relevant data from larger datasets, eventually achieving life-logging [14]. As discussed earlier, the approach used in Walkie-Talkie v3 offers a lightweight alternative by reducing existing overhead associated with data storage, selection, and transmission—though, over time, advancements may make these challenges less of a concern. Multimodal foundation models on wearable XR devices could offer powerful capabilities but still face unresolved technical and privacy challenges.

AI Explainability. As XR systems become increasingly AI-driven, explainable AI (XAI) features are essential [39]. Feedback from participants emphasized the need for transparency in how Walkie-Talkie tracks gaze and generates answers (P7, P8, P11, P12, PP1). Participants suggested Walkie-Talkie should provide visual confirmation of gaze tracking accuracy (P7), vocalize explanations alongside answers to help users understand erroneous responses and adjust their gaze or rephrase queries (P11, PP1), and provide multiple answer options, like a Google search result, for users to choose from (P8, P12). Incorporating XAI features will enhance trust and usability by making the AI's decision-making process more transparent.

Context-Aware Dialogue. Walkie-Talkie supports context-aware dialogue, but token limitations restricted its ability to store and use past data. Future work should look into expanding memory capacity to enhance dialogue capabilities. For example, if the system selects the wrong referent, users should be able to follow up with, "*No, I meant that one*" (P12).

Usability and Gaze Behavior in Long-Term Use. We encourage researchers to conduct in-the-wild studies [32], allowing users to integrate context-aware VAs into their daily lives without researcher intervention. Such studies can provide deeper insights into query formation, gaze behavior shifts, and overall system performance in real-world settings. Additionally, we recommend comparing users' gaze patterns with a wearable VA to their natural gaze behavior without technological intervention. This can further clarify how naturally users engage with gaze-based interactions in these systems. At the same time, while longitudinal gaze data enables more natural interactions, awareness of being tracked may still influence behavior. That said, participants did not find this uncomfortable (unlike with explicit gaze systems), and their gaze adjustments may reflect a natural tendency to assist AI-based systems. Over time, as users grow accustomed to the system, these adjustments may also become effortless, requiring little to no conscious adaptation.

We hope readers continue to explore the space of always-available AI agents on wearable XR, particularly for ambiguous question-answering.

References

[1] Google AI. 2025. Multimodal Live API. https://ai.google.dev/api/multimodal-live
[2] 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
[3] 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
[4] 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
[5] J. B. Brooke. 1996. SUS: A 'Quick and Dirty' Usability Scale.
[6] Donna K. Byron and James F. Allen. 1998. Resolving Demonstrative Anaphora in the TRAINS93 Corpus.
[7] Ricardo Chavarriaga, Pierre W Ferrez, and José del R Millán. 2008. To err is human: Learning from error potentials in brain-computer interfaces. In *Advances in cognitive neurodynamics ICCN 2007*. Springer, 777–782.
[8] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. 2024. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16901–16911.
[9] Juan C Dibene and Enrique Dunn. 2022. HoloLens 2 Sensor Streaming. *arXiv preprint arXiv:2211.02648* (2022).
[10] 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
[11] Stefan Dowiasch, Svenja Marx, Wolfgang Einhäuser, and Frank Bremmer. 2015. Effects of aging on eye movements in the real world. *Frontiers in human neuroscience* 9 (2015), 46.
[12] Andrew T Duchowski and Andrew T Duchowski. 2017. *Eye tracking methodology: Theory and practice*. Springer.
[13] Kerstin Gidlöf, Annika Wallin, Richard Dewhurst, and Kenneth Holmqvist. 2013. Using eye tracking to trace a cognitive process: Gaze behaviour during decision making in a natural environment. *Journal of eye movement research* 6, 1 (2013).
[14] Cathal Gurrin, Alan F. Smeaton, and Aiden R. Doherty. 2014. LifeLogging: Personal Big Data. *Foundations and Trends® in Information Retrieval* 8, 1 (2014), 1–125. https://doi.org/10.1561/1500000033
[15] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*, Peter A. Hancock and Najmedin Meshkati (Eds.). Advances in Psychology, Vol. 52. North-Holland, 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9
[16] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment Anything. arXiv:2304.02643 [cs.CV] https://arxiv.org/abs/2304.02643
[17] Rakshit Kothari, Zhizhuo Yang, Christopher Kanan, Reynold Bailey, Jeff B Pelz, and Gabriel J Diaz. 2020. Gaze-in-wild: A dataset for studying eye and head coordination in everyday activities. *Scientific reports* 10, 1 (2020), 2539.
[18] 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
[19] 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* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 408, 20 pages. https://doi.org/10.1145/3613904.3642230
[20] Geoffrey Leech, Paul Rayson, and Andrew Wilson. 2001. *Word frequencies in written and spoken English: Based on the British National Corpus*. Routledge.
[21] Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691* (2021).
[22] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv:2005.11401 [cs.CL] https://arxiv.org/abs/2005.11401
[23] Jiahao Nick Li, Zhuohao Jerry Zhang, and Jiaju Ma. 2024. OmniQuery: Contextually Augmenting Captured Multimodal Memory to Enable Personal Question Answering. arXiv:2409.08250 [cs.HC] https://arxiv.org/abs/2409.08250
[24] 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
[25] Microsoft. 2024. Azure AI Speech. https://azure.microsoft.com/en-us/products/ai-services/ai-speech.
[26] Microsoft. 2024. Eye tracking on HoloLens 2. https://learn.microsoft.com/en-us/windows/mixed-reality/design/eye-tracking.
[27] OpenAI. 2023. GPT-4. https://openai.com/research/gpt-4
[28] OpenAI. 2024. Hello GPT-4o. https://openai.com/index/hello-gpt-4o/.
[29] Jacob L. Orquin and Simone Mueller Loose. 2013. Attention and choice: A review on eye movements in decision making. *Acta Psychologica* 144, 1 (2013), 190–206. https://doi.org/10.1016/j.actpsy.2013.06.003
[30] Alexander Pastukhov and Jochen Braun. 2010. Rare but precious: microsaccades are highly informative about attentional allocation. *Vision research* 50, 12 (2010), 1173–1184.
[31] Alexander Plopski, Teresa Hirzle, Nahal Norouzi, Long Qian, Gerd Bruder, and Tobias Langlotz. 2022. The Eye in Extended Reality: A Survey on Gaze Interaction and Eye Tracking in Head-worn Extended Reality. *ACM Comput. Surv.* 55, 3, Article 53 (March 2022), 39 pages. https://doi.org/10.1145/3491207
[32] Yvonne Rogers, Paul Marshall, and John M Carroll. 2017. *Research in the Wild*. Vol. 10. Springer.
[33] 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
[34] 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
[35] Naveen Sendhilnathan, Ting Zhang, Ben Lafreniere, Tovi Grossman, and Tanya R Jonker. 2022. Detecting Input Recognition Errors and User Errors using Gaze Dynamics in Virtual Reality. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–19.
[36] Ao Wang, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. 2024. RepViT-SAM: Towards Real-Time Segmenting Anything. arXiv:2312.05760 [cs.CV] https://arxiv.org/abs/2312.05760
[37] Zeyu Wang, Yuanchun Shi, Yuntao Wang, Yuchen Yao, Kun Yan, Yuhan Wang, Lei Ji, Xuhai Xu, and Chun Yu. 2024. G-VOILA: Gaze-Facilitated Information Querying in Daily Scenarios. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 2, Article 78 (May 2024), 33 pages. https://doi.org/10.1145/3659623
[38] Shu Wei, Desmond Bloemers, and Aitor Rovira. 2023. A Preliminary Study of the Eye Tracker in the Meta Quest Pro. In *Proceedings of the 2023 ACM International Conference on Interactive Media Experiences* (Nantes, France) *(IMX '23)*. Association for Computing Machinery, New York, NY, USA, 216–221. https://doi.org/10.1145/3573381.3596467
[39] 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
[40] Kun Yan, Lei Ji, Zeyu Wang, Yuntao Wang, Nan Duan, and Shuai Ma. 2023. Voila-A: Aligning Vision-Language Models with User's Gaze Attention. arXiv:2401.09454 [cs.CV] https://arxiv.org/abs/2401.09454
[41] Liang Zhang, Katherine Jijo, Spurthi Setty, Eden Chung, Fatima Javid, Natan Vidra, and Tommy Clifford. 2024. Enhancing Large Language Model Performance To Answer Questions and Extract Information More Accurately. arXiv:2402.01722 [cs.CL]