Persistent Assistant: Seamless Everyday AI Interactions via Intent Grounding and Multimodal Feedback

Hyunsung Cho
Meta Reality Labs
Redmond, Washington, USA
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
hyunsung@cs.cmu.edu

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

Naveen Sendhilnathan
Meta Reality Labs
Redmond, Washington, USA
naveensendhilnathan@gmail.com

Jonathan Browder
Meta Reality Labs
Redmond, Washington, USA
jonathanbrowder@meta.com

David Lindlbauer
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
davidlindlbauer@cmu.edu

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

Kashyap Todi
Meta Reality Labs
Redmond, Washington, USA
kashyap.todi@gmail.com

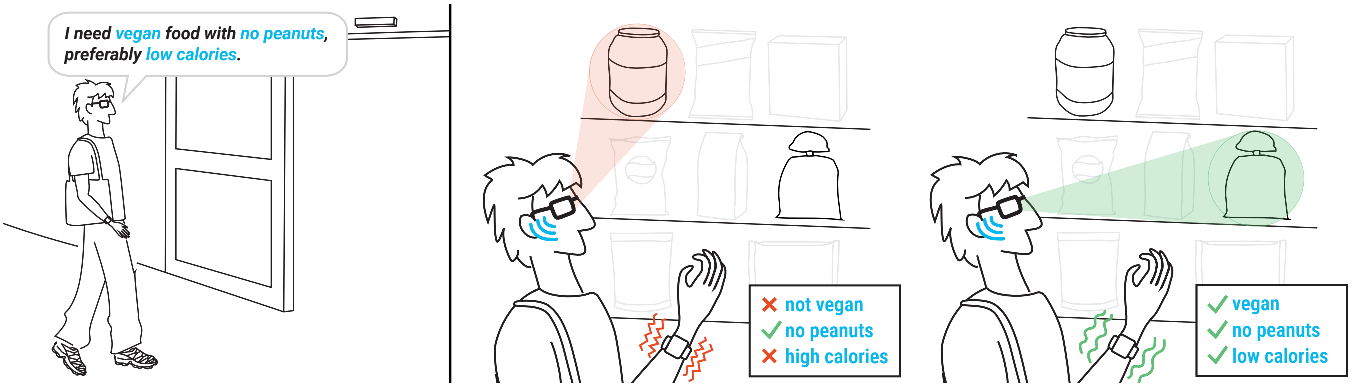

Figure 1: *Persistent Assistant* is a framework to support persistent AI interactions through intent grounding, embodied input, and multimodal feedback. Before entering a supermarket, the user specifies dietary preferences—vegan, no peanuts, and low calories. As they navigate the store and interact with items using gaze and pinch gestures, the system provides immediate feedback via haptic signals and optional speech, indicating whether an item meets their criteria.

Abstract

Current AI assistants predominantly use natural language interactions, which can be time-consuming and cognitively demanding, especially for frequent, repetitive tasks in daily life. We propose *Persistent Assistant*, a framework for seamless and unobtrusive interactions with AI assistants. The framework has three key functionalities: (1) efficient intent specification through *grounded interactions*, (2) seamless target referencing through *embodied input*, and (3) intuitive response comprehension through *multimodal perceptible feedback*. We developed a proof-of-concept system for everyday decision-making tasks, where users can easily repeat queries over multiple objects using eye gaze and pinch gesture, as well as receiving multimodal haptic and speech feedback. Our study shows that multimodal feedback enhances user experience and preference by reducing physical demand, increasing perceived speed, and enabling intuitive and instinctive human-AI assistant interaction. We discuss how our framework can be applied to build seamless and unobtrusive AI assistants for everyday persistent tasks.

CHI '25, Yokohama, Japan
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3714317

CCS Concepts

• **Human-centered computing → HCI theory, concepts and models**; **Interaction techniques**; *Mixed / augmented reality*; *Natural language interfaces.*

Keywords

Wearable AI assistants, grounding, multimodal interaction, gaze and gesture input, haptic and speech feedback

ACM Reference Format:
Hyunsung Cho, Jacqui Fashimpaur, Naveen Sendhilnathan, Jonathan Browder, David Lindlbauer, Tanya R. Jonker, and Kashyap Todi. 2025. Persistent Assistant: Seamless Everyday AI Interactions via Intent Grounding and Multimodal Feedback. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3706598.3714317

1 Introduction

Current personal AI assistants (e.g., Amazon Alexa, Siri) are predominantly conversational, mimicking human-to-human communication by using natural language as the primary mode of communication. This means that when a user provides natural language input, the assistant responds in natural language. Users speak or type verbal commands or queries, such as *"Is this bag of chips vegetarian?"*, *"What is the rating of this restaurant?"*, or *"Wake me up at 7:00 AM."* Assistants read aloud or display verbal responses, e.g., *"Yes, this bag of chips is vegetarian"*, *"The restaurant has a rating of 4.5 stars on Google Maps"*, or *"I have set an alarm for 7 AM"*.

However, when verbal queries are repeated frequently, formulating them precisely and interpreting assistants' verbal responses can be cumbersome and time-consuming. For example, a user searching for gluten-free items in a supermarket may need to repeatedly ask, *"Is this item gluten-free?"* and listen to responses, *"No, these noodles are not gluten-free."* Furthermore, verbal interactions may interfere with users' primary tasks. Speech-based interactions can be disruptive in social settings; challenging to perform in noisy environments; privacy-invasive in sensitive contexts (e.g., libraries, public transit); and conflicting with ongoing activities (e.g., music, phone call). Visual interactions, such as reading text-based responses, can also be distracting, especially in attention-demanding situations like driving or navigating crowded areas.

To address these limitations, we explore the use of non-verbal input and haptic feedback for discreet and unobtrusive wearable AI assistant interactions in persistent tasks (i.e., tasks that occur repeatedly over time, often in a user's daily routine, requiring frequent interaction and continuous engagement). The biggest challenge in adopting non-verbal input and haptic feedback is expressivity. Prior work on vibrotactile patterns focused on designing identifiable tactons representing system actions [5], menu items [64], numbers [6, 11], gestures [60], and physical objects [8]. However, encoding distinct patterns is not scalable to the wide range of dynamic, context-ware information that is exchanged with AI assistants (e.g., "Is this food vegan?" or "Have I met this person before?").

We propose *Persistent Assistant*, a framework that reimagines wearable AI assistant interactions for everyday tasks. By leveraging grounded embodied input and multimodal feedback, our approach captures users' intent seamlessly and delivers intuitive responses without relying on explicit speech-based exchanges. Our framework introduces *Intent Grounding Continuum* and *Speech-Haptic Feedback Continuum*, which enable aligning the assistant's understanding with users' intent, thus simplifying the output of the assistant. Using this structure, Persistent Assistant provides three key functionalities:

- **Efficient Intent Specification through Grounded Interactions:** Establishing a grounded understanding of the user's goals, preferences, and context early in the interaction to provide relevant and concise feedback.
- **Seamless Target Selection with Embodied Input:** Leveraging embodied behaviors such as gaze and gestures to specify targets seamlessly without interrupting the user's ongoing activities.
- **Intuitive Response Comprehension through Multimodal Perceptible Feedback:** Employing a combination of haptic and speech feedback to convey information effectively while minimizing cognitive load.

By embracing more direct forms of input and output on top of language-based communication, we aim to reduce the cognitive and physical demands on users, allowing for more time-efficient, low-effort interactions with AI assistants in daily life.

We contribute a proof-of-concept implementation of the *Persistent Assistant* framework for everyday decision-making tasks. The implementation uses display-free smart glasses (e.g., Ray-Ban Meta smart glasses[1]) and a wristband setup inspired by emerging companion wearable devices, such as smartwatches or EMG wristbands [41], which expand input and output possibilities for wearable AI assistants via more granular interactions through diverse input channels [13, 19, 35] and private, subtle haptic feedback [44, 45, 53, 54, 66]. For instance, as illustrated in Figure 1, when a user enters a supermarket and specifies dietary constraints—vegan, no peanuts, and low calories—the system retains this intent. As the user navigates the aisles, they can simply look at a product and perform a pinch gesture, prompting the assistant to analyze the item and provide immediate haptic feedback, such as a smooth vibration for a match or a jagged pulse for a mismatch. This eliminates the need for repetitive voice queries and streamlines decision-making. This enables users to receive feedback instantly without the need for constant attention to language comprehension or disruption of ongoing tasks.

We conducted a user study to evaluate the user experience of using Persistent Assistant for frequent interactions with an AI assistant and the effectiveness of different feedback designs. Users were tasked with completing a grocery shopping task while interacting with Persistent Assistant. Our findings indicate that multimodal feedback enhanced user experience and increased users' preference by providing timely and concise information without cognitively overloading users. Participants appreciated the immediacy and unobtrusiveness of haptic feedback, particularly when the haptic patterns were memorable and easily interpretable.

Our contributions include the Persistent Assistant framework including the Intent Grounding and Speech-Haptic Feedback Continuums, a proof-of-concept implementation, findings from a user evaluation, and design implications for intuitive, unobtrusive AI assistants for persistent support.

[1] https://www.meta.com/smart-glasses

2 Related Work

2.1 Conversational AI Assistants for Real-world Information Retrieval

Current methods for retrieving information about real-world objects, such as items in a supermarket, typically require explicit user input. Users must identify objects by name, describe their attributes, or provide images to search engines (e.g., Google Search, Bing) or AI assistants (e.g., Siri, ChatGPT, Google Lens). This process creates a disconnect between physical interactions with objects and digital information retrieval. Instead of obtaining information seamlessly, users must pause, formulate a query, and engage with a separate digital interface. This shift disrupts the natural flow of interaction, making the process slower and less intuitive.

To bridge this gap, several approaches integrate digital and physical interactions. XR-Objects [12] uses Augmented Reality (AR) overlays to spatially anchor information on objects, while XAIR [61] ensures AI explanations are understandable to end-users. GazePointAR [30] combines a context-aware voice assistant with wearable AR, using gaze, pointing gestures, and conversation history to disambiguate speech queries. However, they remain heavily reliant on verbal communication, providing information primarily through natural language input and output.

Building on this foundation, researchers in conversation information seeking (CIS) have explored integrating multimodal inputs and outputs to enhance interaction. For example, Deldjoo et al. [10] introduced the concept of multi-modal CIS, showing how non-verbal multimodal inputs such as facial expression, gestures, emotion, eye gaze, and touch pressure can help incorporate richer context and reduce errors. Additionally, multimodal outputs like visual animations, on-screen text, and audio-visual narration improve learning [39] and accessibility, catering to diverse user needs and preferences. These advancements highlight the potential of moving beyond verbal interactions to create more versatile and user-friendly systems.

Our work takes this concept a step further by proposing a framework for persistent AI assistants that utilizes non-visual, lightweight mechanisms, such as haptic feedback, to create a seamless link between physical objects and digital queries. This approach is designed for routine, low-effort interactions, addressing the inefficiencies of language-heavy methods while fostering a more natural integration of digital assistance into everyday tasks.

2.2 Alternative Interaction Modalities for Augmenting Phyical Objects

Studies have explored enhancing real-world interactions by anchoring digital functions to physical objects. Approaches such as Snaplink [7], Snap-to-it [9], and OmniActions [33] detect objects to trigger context-aware functionalities using computer vision. Other approaches use passive vibroacoustic sensing [28], responsive vibroacoustic signals [16], and electromagnetic signals [59] to recognize objects for similar purposes. These methods envision using real-world objects to trigger actions on connected devices, e.g., laptops, smartphones, or the Internet of Things (IoT).

Other approaches focus on embedding and displaying object-related information. For instance, InfoLED [63] uses LED lights to transmit location and status information to AR clients, while LightAnchors [1] leverages point lights for AR anchoring. ODIF [52] fuses information from multiple sources to enhance ubiquitous detection, and Reality Editor [22] overlays graphical elements on tangible interfaces for flexible interaction with "smarter objects." Methods like BlendMR [20] integrate visual interfaces onto object surfaces, while non-visual approaches such as SonoHaptics [8] and Affordance++ [36] use audio-haptic feedback and muscle stimulation to communicate object identities and affordances.

Prior work introduces innovative ways to use object recognition as an anchor to link physical objects with digital functions. We extend these approaches beyond predefined object-action mappings toward dynamic, context-aware interactions. Instead of relying on fixed functions, our system dynamically interprets user intent and provides seamless AI assistant by leveraging direct input modalities (e.g., gaze and gestures) and multimodal feedback (e.g., haptic and audio).

2.3 Multimodal Interaction for Voice Assistants

Research has explored multimodal inputs (e.g., voice, gesture, and gaze) to enhance human-computer interactions, particularly in AR and other immersive environments. Early work like Bolt's "Put-that-there" system [3] demonstrated natural integration of voice with pointing gestures, setting the stage for multimodal interaction. Subsequent studies, such as MAGIC pointing [67], combined gaze and manual inputs to enhance precision and ergonomics. Recent AR research has investigated techniques like head-eye coordination for target selection [27] and gaze-hand alignment for menu interaction [37]

Williams et al. [58] examined user-defined multimodal interactions for object manipulation in AR, while Roider et al. [49] and Neßelrath et al. [42] focused on reducing cognitive load by using gaze to activate speech inputs in cars. Other works, such as Piumsomboon et al. [47] and Irawati et al. [24], evaluated gesture-speech combinations in AR, while Hertel et al. [21] provided a taxonomy for AR interaction techniques, identifying strengths and limitations across modalities. Mayer et al. [40] and Lee et al. [29] further expanded on this research by incorporating gaze to refine voice assistant interactions and resolve ambiguities.

Together, these studies demonstrate the potential of multimodal inputs for naturalistic, efficient interactions. Our framework adopts this concept by integrating gaze, gestures, and haptic feedback to enhance object-centric voice assistant capabilities while simplifying feedback to reduce cognitive load.

2.4 Haptic Feedback and Multimodal Information Encoding

Haptic feedback has been widely explored as a means to enhance non-visual interactions and convey a variety of information, from simple system states to complex emotional expressions. Techniques such as vibrotactile patterns have been used for various purposes: ActiVibe [6] developed patterns to represent numerical progress, while HapticLock [11] used Morse Code vibrations for eyes-free PIN entry. Similarly, Komatsu et al. [25] investigated vibration patterns to reflect system confidence, while VibEmoji [2] leveraged patterns from VibViz [51] to create emoticons that express emotions through

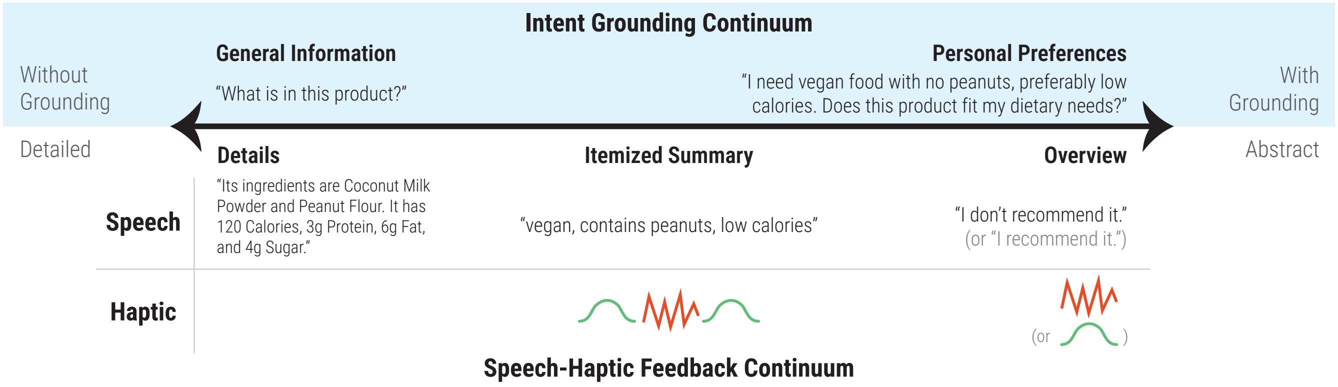
Intent Grounding Continuum
Without Grounding
General Information
"What is in this product?"
Personal Preferences
"I need vegan food with no peanuts, preferably low calories. Does this product fit my dietary needs?"
With Grounding
Detailed
Details
Itemized Summary
Overview
Abstract
Speech
"Its ingredients are Coconut Milk Powder and Peanut Flour. It has 120 Calories, 3g Protein, 6g Fat, and 4g Sugar."
"vegan, contains peanuts, low calories"
"I don't recommend it."
(or "I recommend it.")
Haptic
(or)
Speech-Haptic Feedback Continuum

Figure 2: The Intent Grounding Continuum (top) shows how user queries evolve from general information (e.g., *"What is in this product?"*) to personalized preferences (e.g., *"I need vegan food with no peanuts"*). As grounding increases, speech feedback becomes simpler, transitioning from detailed explanations to concise summaries and abstract recommendations (e.g., *"I don't recommend it"*), as shown in the Speech-Haptic Feedback Continuum (bottom). This simplification opens opportunities for non-verbal modalities like haptics. Smooth green waves represent positive matches, while sharp red waves indicate negative matches, enabling quick and intuitive interpretation of the assistant's responses.

haptic cues. Other work such as VibHand [68] employed on-hand cues for non-visual exploration of digital graphics, demonstrating haptic feedback's versatility in enhancing user experience.

In addition to communication of state or emotion, haptic feedback has proven effective for navigation and interaction. Youn et al. [64, 65] used tactons [5] mapped to wrist rotations for non-visual menu navigation, while Graham-Knight et al. [17] applied smartwatch vibrations to convey frequent texts and categories. Similarly, PocketMenu [46] combined vibrotactile and speech feedback for in-pocket phone control, enhancing efficiency by allowing users to rely on memorized cues. Haptic feedback has also been employed to increase immersion in virtual environments, such as using heartbeat feedback in gaming [43] and maintaining real-world awareness through headset vibrations in VR [55]. Moreover, it guides user interactions. Rhythmic gestures [14] and microgestures [15] provide cues for input, while Roudaut et al. [50] used force feedback to confirm commands. Xu et al. [60] developed semantically relevant vibrotactile patterns for hand gestures.

Despite these advances, existing research has not fully explored the potential of haptic feedback for interactions with AI assistants which involve dynamic exchange of a vast range of information for contextual use. Our work integrates haptic feedback with intent grounding and embodied inputs, such as gaze and gestures, to dynamically encode contextual information into memorable patterns to provide immediate, context-aware responses, facilitating frequent and seamless decision-making without the need for verbal engagement.

3 Persistent Assistant

Persistent Assistant is a conceptual framework for quick and intuitive interactions with wearable AI assistants in everyday tasks.

3.1 Target Interactions and Design Goals

Decision making in complex environments requires users to quickly and sometimes discreetly filter multiple options based on personal preferences, constraints, or contextual queries. For example, a user may need to check dietary restrictions on products in a supermarket (Figure 1), compare menu options in a café, or recall past interactions with someone they meet at a conference. In these scenarios, traditional interactions on smart glasses require users to explicitly verbalize their queries for every entity (e.g., asking "Are these chips vegan?" for dozens of items on a supermarket shelf). However, this approach can be cognitively demanding and socially awkward, especially when repeated over multiple query-response cycles for different products, menus, or people.

Persistent Assistant aims to streamline these interactions by reducing overt, repetitive input while maximizing the responsiveness of the assistant. We target *continual* interactions, where users repeat a same set of queries across multiple entities within a session (e.g., a shopping trip or conference). By leveraging context retention and multimodal input and output, Persistent Assistant ensures users efficiently navigate AI-assisted tasks without unnecessary friction.

3.2 Design Principles and Components

Persistent Assistant redesigns three core processes of wearable AI assistant interaction—intent specification, target selection, and response comprehension—through grounded interactions, embodied input, and multimodal perceptible feedback.

3.2.1 Efficient Intent Specification: Grounded Interactions. In traditional conversational interactions, users must verbally articulate their intent, which can be particularly difficult to be repeated in public, social, noisy, or privacy-sensitive environments. Persistent Assistant mitigates this by grounding user intent early in the interaction, allowing it to be reused throughout a session of repeated interactions. "Grounding" here refers to the process of aligning the assistant's understanding with the user's goals, preferences, and context to ensure effective communication [4].

Intent Grounding Continuum. We introduce the **Intent Grounding Continuum** (Figure 2, top) as a framework for interpreting user intent with increasing specificity—from broad, general questions to specific, personalized queries. This structured approach allows the assistant to provide feedback that matches the level of detail required by the user's intent. By establishing this understanding early, the assistant can deliver concise, focused responses, avoiding unnecessary elaboration or repetitive confirmation.

3.2.2 Seamless Target Referencing: Embodied Input. Once intent is grounded, it can be reused across multiple targets through less elaborate, non-verbal inputs. Persistent Assistant leverages embodied input such as gaze and gestures to facilitate seamless and unobtrusive target referencing. Even in public, social, or privacy-sensitive settings, users can perform a quick glance and a subtle gesture to activate the grounded intent. This allows for unobtrusive interactions without disrupting ongoing conversations or requiring sustained visual attention on a target when intent has already been specified earlier in the session. The seamless target selection can be further extended to support natural behaviors such as pointing or picking up objects in future works, reducing the need for explicit verbal identification of targets. These methods allow the assistant to infer the user's focus of attention with less interruption to the user's ongoing activity.

3.2.3 Intuitive Response Comprehension: Multimodal Perceptible Feedback. Interpreting verbal responses from an assistant can be as challenging as formulating verbal queries, especially when a user's visual and auditory channels are occupied with social interactions or primary tasks. Persistent Assistant uses multimodal haptic and speech feedback to communicate responses in an intuitive and unobtrusive manner.

Speech-Haptic Feedback Continuum. Our **Speech-Haptic Feedback Continuum** (Figure 2, bottom) transitions feedback from detailed verbal descriptions to concise, abstract responses as the assistant has a more specific understanding of user intent. For general queries (e.g., *"What is in this product?"*), detailed speech feedback is provided (e.g., *"Its ingredients are Coconut Milk Powder, Peanut Flour ..."*). As intents become more specific (e.g., *"I need vegan food with no peanuts, preferably low calories. Does this product fit my dietary needs?"*), the feedback transitions to a concise, itemized summary (e.g., *"vegan, contains peanuts, low calories"*). At the highest level of grounding, abstract feedback such as a brief recommendation (e.g., *"I don't recommend it."*) reduces cognitive load and enhances interaction efficiency.

This progression toward *concise* and *abstract* feedback opens opportunities for utilizing modalities beyond speech. For example, an itemized summary can be encoded as a sequence of smooth, sharp jagged, and smooth vibrations to represent positive, negative, and positive matches, respectively. Similarly, a brief recommendation can be conveyed through a single smooth (i.e., positive) or jagged vibration (i.e., negative).

Why haptics? Visual and auditory channels are often engaged in primary tasks, such as conversations, calls, navigation, music listening, or video watching. In these situations, visual or auditory feedback will overlap and interfere with the signals involved in the primary tasks, increasing cognitive load and reducing efficiency [56, 57]. To mitigate this, Persistent Assistant prioritizes haptic feedback, which avoids sensory conflicts and remains perceptible without interfering with existing sensory demands. Haptics aligns with the design principles of minimizing cognitive load and providing a low-effort, private feedback mechanism. Delivered through wearable devices like wristbands, belts, jackets, or shoes, haptic feedback enables persistent, unobtrusive interactions, making it a more seamless alternative to visual or auditory feedback.

Multimodal feedback. While abstract haptic feedback is effective when grounded in context, users may lose track of the grounded intent over time. To address this, Persistent Assistant incorporates speech feedback as an optional complement to haptics (Figure 3, left) to maintain clarity when needed.

To probe different facets on the Speech-Haptic Feedback Continuum (Figure 2), Persistent Assistant employs two distinct multimodal feedback designs, offering varying levels of granularity:

(1) **Haptic Overview + Speech Explanation:** This approach is situated on the abstract end of the spectrum. When the user performs a pinch gesture, a single vibration communicates the degree to which the user's criteria are met: smooth for all criteria satisfied; sharp and jagged for none; and toned-down jagged vibration for partial matches. The patterns were inspired by Xu et al.'s metaphor-based haptic patterns [60]. If the user continues to hold the pinch gesture, the feedback transitions towards the detailed end of the spectrum and provides a speech explanation about which criteria is not met, e.g., "It contains peanuts." Since the haptic feedback is highly abstract, only communicating the valence of the assistant response, the speech feedback gets a complementary role (cf. [38]), providing a richer explanation than haptic feedback.
(2) **Haptic Itemized Summary + Speech Itemized Summary:** This approach probes the middle of the continuum, using *positional encoding* within the vibration sequence to represent query criteria. Each vibration corresponds to a criterion: smooth if satisfied; sharp and jagged if not. By dynamically mapping feedback to positions, the system conveys binary information specific to the user's criteria. This design is inspired by prior work with vibrotactile Morse code, such as HapticLock [11] and ActiVibe [6], which demonstrated the effectiveness of encoding numeric information through discrete tactile patterns. As smaller mappings improve accuracy [64], the number of items is limited to three. Speech feedback *reinforces* these mappings, aiding recall and interpretation if needed.

By balancing these feedback methods, Persistent Assistant aims to find the sweet spot between brief haptic cues and comprehensive speech, accommodating different user preferences and scenarios.

3.2.4 Components Integration: Contextual Mediator. The mediator serves as the cognitive core of the assistant, interpreting embodied inputs, grounding user intents, and generating appropriate multimodal feedback. The mediator processes inputs from various naturalistic modalities and maps them to specific targets within the assistant's knowledge base, enabling swift and accurate identification without the need for detailed verbal commands. The mediator dynamically adjusts its interpretative models to align the user's

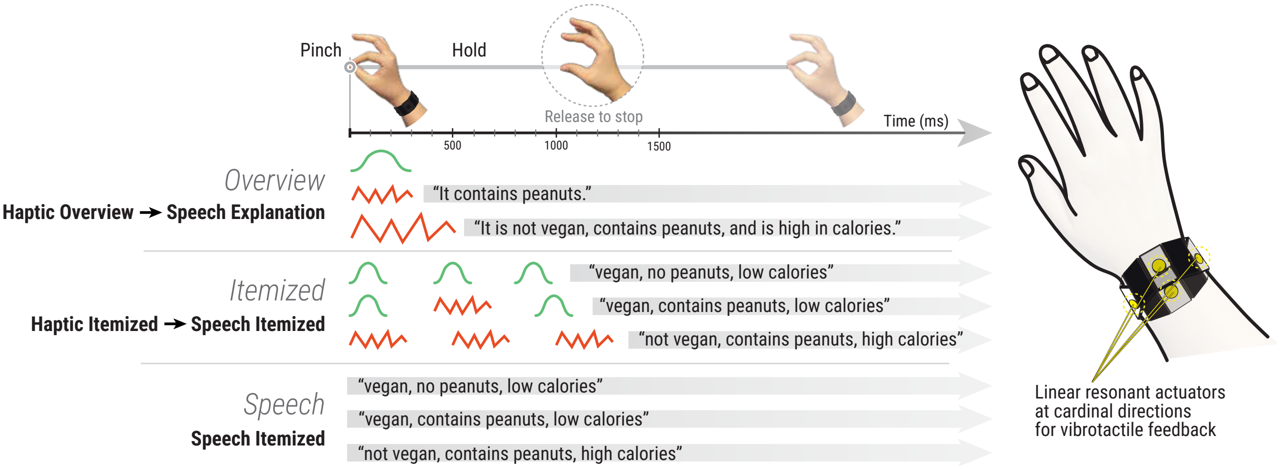
Pinch
Hold
Release to stop
Time (ms)
500
1000
1500
Overview
Haptic Overview → Speech Explanation
"It contains peanuts."
"It is not vegan, contains peanuts, and is high in calories."
Itemized
Haptic Itemized → Speech Itemized
"vegan, no peanuts, low calories"
"vegan, contains peanuts, low calories"
"not vegan, contains peanuts, high calories"
Speech
Speech Itemized
"vegan, no peanuts, low calories"
"vegan, contains peanuts, low calories"
"not vegan, contains peanuts, high calories"
Linear resonant actuators at cardinal directions for vibrotactile feedback

Figure 3: Left: Multimodal speech-haptic feedback designs of Persistent Assistant (*Overview* and *Itemized*) and baseline speech-only feedback in the user study. When the user looks at an object and makes a pinch gesture, *Overview* and *Itemized* begin with concise haptic feedback and allow transition into more detailed speech feedback on demand. The user can release the pinch to stop at any point. In the *Overview* design, smooth green waves correspond to gentle, continuous vibrations, reinforcing a positive match. Sharp red jagged waves are translated into strong, rapid pulses, signaling a full mismatch. Toned-down jagged waves indicate weaker pulses, representing a partial mismatch, with speech providing additional detail on demand. The *Itemized* design uses a sequence of haptic signals to encode query criteria, where each signal corresponds to a criterion (e.g., "vegan" or "low calories") and is reinforced with speech. The *Speech* design serves as a speech-only baseline, delivering detailed verbal summaries without haptic feedback. Right: Illustration of the haptic wristband with four linear resonant actuators at cardinal directions around the wrist.

goals and preferences, and tailors the feedback to the user's current target and the intent grounding.

By integrating these capabilities, the mediator enables Persistent Assistant to function as a cohesive and persistent assistant, maintaining a simple, low-effort, naturalistic interaction style.

3.3 System Pipeline

The system pipeline (Figure 4) illustrates the Persistent Assistant design principles into a cohesive workflow:

1. **Intent Grounding:** The user expresses their preferences when shopping (e.g., "vegan food with no peanuts, preferably low calories"). The system uses this input to establish an early grounding, aligning its understanding with the user's intent, as outlined in Section 3.2.1. This can be specified at the start of a task session, e.g., entering a supermarket (Figure 4a).
2. **Target Referencing with Embodied Input:** Gaze and egocentric camera on smart glasses are used to reference an in-world target, and pinch gesture is used to trigger a query (Figure 4c). When a pinch is detected, Persistent Assistant grabs the camera frame and gaze point, crops the camera image around the gaze point, and passes the crop to the Contextual Mediator.
3. **Contextual Mediator:** A vision language model converts user intent to a query list of binary questions (Figure 4b, top). It infers the target using the user's gaze and surrounding image (Figure 4b, middle). It answers the query list about the object and formats the responses into a structured format for multimodal feedback generation (Figure 4b, bottom).
4. **Multimodal Feedback Delivery:** The system provides feedback based on user intent using a combination of haptic and speech cues (Figure 4d) as described in Section 3.2.3.

By integrating these components, Persistent Assistant demonstrates how a persistent, unobtrusive assistant can facilitate decision-making in real-world contexts.

3.4 Implementation

We implemented a Unity prototype for Meta Quest Pro to simulate smart glasses with eye tracking. The device has a RGB passthrough that simulates display-free smart glasses and eye tracking functionality. It takes a user's voice input to specify the query, using Meta Voice SDK v65.0 to detect and convert the user's speech to text. Then, it updates and stores the current criteria or active 'filters' in the system based on the query. The criteria only need to be set once and remain active for all following queries until updated.

When a user wants to query whether an object meets their criteria, they can look at the object and perform a single pinch gesture. The query always relates to the current criteria. The gesture is detected using Meta Interaction SDK v65.0. When the gesture is detected, the system combines the RGB image of the user's egocentric view and eye tracking to crop the RGB image around the object that the user is currently looking at. The cropped image is used as input to a vision language model (VLM), GPT-4o. Persistent Assistant uses the prompt provided in the Supplementary Materials to perform object recognition, object information retrieval, criteria matching, and result formatting. The last part of the prompt is adapted to different feedback designs, the *Overview* and *Itemized*.

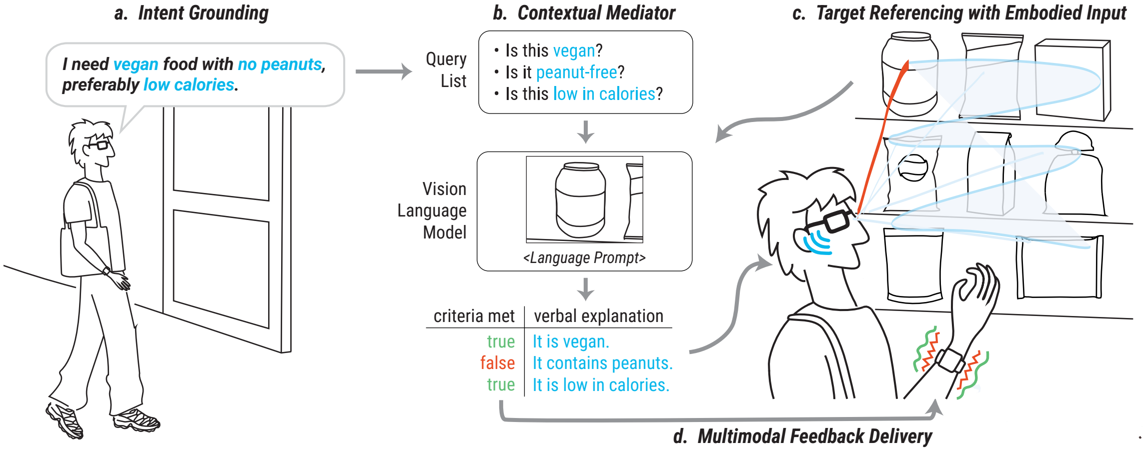

Figure 4: *Persistent Assistant* pipeline. A user enters a supermarket and communicates their dietary constraints to Persistent Assistant, which processes the user intent into a query list, waiting for the user's input. As the user moves through the aisles, they want to check if various products on the shelves meet their constraints. Instead of repeating verbal queries, the user simply looks at the desired item and performs a pinch gesture. Persistent Assistant captures the object referenced by the user's gaze and utilizes the vision language model to evaluate whether the product meets the predefined constraints. Based on the model's response, Persistent Assistant provides multimodal feedback, combining speech and haptic cues, to convey the satisfaction of constraints in a quick and intuitive manner.

The system processes the formatted text from the VLM and presents the speech-haptic feedback accordingly. Feedback continues to play while the user is holding the pinch gesture (Figure 3, left) and stops when the pinch is released. The speech feedback is played through the built-in headset speakers, and the vibrotactile haptic feedback through a custom wristband with four linear resonant actuators placed at four cardinal directions around the wrist (Figure 3, right). We implemented haptic feedback on the wrist instead of the glasses for comfort and distinguishability from passive environmental vibrations while walking.

In *Overview*, the 'all satisfied' vibration is a 300 ms smooth sine wave vibration at 170 Hz. The 'none satisfied' vibration is a 500 ms sharp, jagged vibration, a combination of 170 Hz and 150 Hz sine waves, similar to the 'Cancel' haptic pattern in Xu et al. [60]. The 'partially satisfied' vibration is a 300 ms toned-down sharp, jagged vibration at lower frequencies of 100 Hz and 75 Hz of a similar pattern. In *Itemized*, the 'satisfied' vibration is a 200 ms sine wave at 170 Hz. The 'not satisfied' vibration is a 300 ms version of the 'none satisfied' vibration. The three vibrations are concatenated with a 200-ms delay between each other, determined during pretests for the best distinguishability.

3.4.1 Prompt Evaluation. The main source of ambiguity or uncertainty lies in the prompt processing of the VLM, whether it produces accurate results in a valid format. We tested our two prompts (for Intent Specification and for Contextual Mediator) on the (1) accuracy of generated responses and (2) formatting validity in diverse environments and queries. Details of the tested environments (egocentric images), intent prompts, and cropped images of gaze targets are provided in Appendix A.

For Intent Specification, the prompt could parse sentences with multiple queries (e.g., *"Looking at couches in the furniture store, I'm searching for one that is comfortable, doesn't require assembly, and isn't made of plastic."*) into a query list (Figure 4) with 100% accuracy and zero formatting error, achieving a median latency of 1.35 s (*SD* = 0.27 s).

For Contextual Mediator, which incorporates cropped images, object states, and criteria-based results to generate multimodal feedback, the evaluation for the *Itemized* prompt showed 95.7% accuracy (155 out of 162 criteria items) for both preference alignment and correct information, with zero formatting errors. The median latency was 2.67 s (*SD* = 0.88 s). Errors (4.3%) stemmed from missing attributes in certain cases. The *Overview* prompts attained 98.1% accuracy (53 out of 54 criteria sets) for preference alignment and 92.6% accuracy (50 out of 54 criteria sets) for generating correct feedback. Errors in this case arose from preference misalignment or insufficiently explicit object states. The median latency for this prompt was 1.63 s (*SD* = 0.78 s). These results validate the robustness of the formatting process while identifying areas for improvement in handling nuanced object states and preferences. We discuss limitations and potential challenges in Section 6.

4 User Study: Speech-Haptic Feedback Design

We evaluated the Persistent Assistant interactions of gaze- and gesture-based referencing and speech-haptic feedback through a scenario-based user study. Participants completed a grocery shopping task while interacting with Persistent Assistant.

4.1 Participants

We recruited 12 participants (4 female, 7 male, 1 undisclosed) between 18 and 58 years of age (*M* = 43.9, *SD* = 11.7). All participants had normal or corrected-to-normal vision, hearing, and motor abilities based on self-reports. Participants had occupations of engineer,

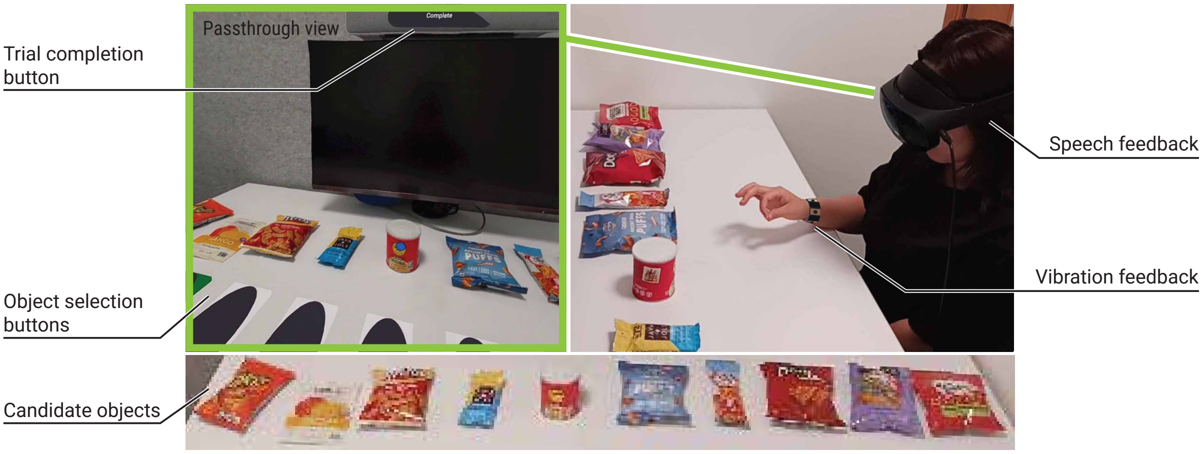

Figure 5: User study setup. A participant *(right)* wears a Meta Quest Pro with passthrough and performs a grocery shopping task with candidate objects on the table *(bottom)*. To query whether an object meets the criteria, the participant looks at the object and performs a pinch gesture, triggering vibration feedback from a wristband (Figure 3) and speech feedback through the AR headset. The participant can then select an object to "purchase" by choosing the corresponding object selection button in AR *(left, bottom)*. Once they finalize their selections for the task, they complete the trial by selecting the AR trial completion button *(left, top)*.

caregiver, teacher, consultant, student, program manager, business intelligence analyst, and rideshare driver. The study was approved by the institution's IRB, and participation was voluntary under informed consent with monetary compensation for their time.

4.2 Study Design

We used a within-subject design with the independent variable *speech-haptic feedback designs*. We compared three designs, illustrated in Figure 3, to assess how well they support rapid decision-making, reduce cognitive load, and enhance user experience: Haptic Overview + Speech Explanation (*Overview*), Haptic Itemized Summary + Speech Itemized Summary (*Itemized*), and Speech Itemized Summary (*Speech*). In addition to the two multimodal designs described in Section 3.2.3, we added *Speech* as a baseline representing current voice assistants. To ensure fairness to the baseline in the highly repetitive nature of the experimental task, we used the same simplified input and structured the response to leverage its strengths by using concise, keyword-based responses rather than full sentences. Prior research by Haas et al. [18] has shown that shorter responses are perceived as similarly useful and likable as full-sentence responses, while being more efficient for command-based tasks and sometimes easier to comprehend. This decision was made to provide the baseline with an advantage that aligns with how voice assistants are optimized for speed and simplicity in real-world usage, ensuring a robust comparison against the more complex multimodal designs.

4.2.1 Task. We designed a task that simulates everyday decision-making scenarios, where users compare options based on preferences and constraints to make optimal choices: shopping three snacks in a supermarket that best meet the given criteria to bring to a friend's party. Ten snacks or objects were given as options (Figure 5), and each object has three attributes: (1) gluten-free / contains gluten, (2) dairy-free / contains dairy, and (3) low calories / high calories. In each trial, the task was to find three snacks that are gluten-free, dairy-free, and low calories among the ten snacks. To simulate a real-world scenario where it is important not only to confirm that an object meets the criteria but also to obtain details about it, we clarified that the gluten and dairy criteria are mandatory dietary restrictions, while the final attribute is an optional preference. The attributes of the ten objects were randomized every trial under a constraint that there should be at least five snacks that meet the required criteria (1) and (2). This design encouraged participants to pay attention to the details of the feedback instead of only waiting for an all-positive signal. The randomly assigned attributes were irrelevant to the actual products' information, and participants were informed about it. This assumes a scenario where participants do not have any prior knowledge about the products. Participants made the selections using virtual buttons associated with each object (object selection buttons in Figure 5) and proceeded to the next trial using the trial completion button with a new set of object attributes. Participants performed five trials per condition, for a total of 20 trials across all four conditions.

4.2.2 Measures. We evaluated the system using both quantitative metrics and qualitative feedback:

Object search performance measures. We calculated the **top-k accuracy (k = 3), precision, recall, and F_1 score** in finding the optimal set of objects that satisfy the given criteria. We also recorded **task completion time** as the duration between the first pinch trigger to check an object's feedback and the last selection or deselection of an object.

Subjective ratings. Participants also provided subjective ratings using the simplified **NASA Task Load Index (TLX)** on a 7-point Likert scale; ratings of **perceived speed, accuracy, and confidence** on a 7-point Likert scale; and a modified version [12] of the

Human-AI Language-based Interaction Evaluation (HALIE) framework [31].

Qualitative feedback. After all conditions, we conducted a **semi-structure interview**, where participants were asked to give **preference rankings** among the three feedback designs, elaborate on their **experiences** with each design, and compare the three designs. In addition, we asked participants to reflect on the initial activity that simulates the conversational interaction with the AI assistant (described in Section 4.4) and compare it with the three designs. Lastly, they were asked to describe their ideal interaction with a wearable AI assistant throughout their daily lives. The interview data was recorded, transcribed, and analyzed by the primary author through a thematic analysis.

4.3 Apparatus

The study was conducted in a private room with a desk in front of the participant, with a row of 10 snacks on the desk, as shown in Figure 5. We used the prototype described in Section 3.4 for the user study, using the Meta Quest Pro headset in passthrough mode. To focus the evaluation on the interactive aspects of the input and output systems, we used precomputed mappings of speech-haptic feedbacks for each object, instead of an online VLM agent, to isolate the effect of latency in the speech-haptic feedback perception.

4.4 Procedure

The study started with a scenario-based introduction for contextual priming about the shopping scenario with an AI assistant. The researcher introduced the multimodal capabilities of current smart glasses in the market, specifically Meta Ray-Ban glasses as an example image. The researcher then introduced the study task scenario of shopping three snacks in a supermarket to bring to a friend's party, which best meet the given criteria. To familiarize participants with AI assistant interactions, the researcher asked participants to act out how they would freely interact with a voice assistant to accomplish the task. The researcher simulated the assistant answering their questions. This familiarization task also provided data on what information they would ask to a voice assistant on the glasses and how they would frame a question or a series of questions to complete the task.

After contextual priming, participants were presented with three task sessions using multimodal feedback designs. The order of the conditions was counterbalanced with a Latin square. After each condition, participants filled out the questionnaires and provided a brief description of their experience. On completion of all three conditions, participants completed semi-structured interview.

5 Results

We analyzed user study data in terms of of optimal object search performance, perceived performance, task load, and interaction quality, as well as qualitative experiences of the speech-haptic feedback designs.

5.1 Quantitative Measures

The quantitative results highlight that all three versions of Persistent Assistant were helpful, enjoyable, satisfactory, responsive, and easy to interact to help the participants complete the task successfully. Among the three feedback designs, the introduction of haptic feedback in *Itemized* and *Overview* showed faster perceived speed than *Speech*. Presenting a summary of all information (*Itemized* and *Speech*) improved perceived accuracy compared to *Overview*.

5.1.1 Optimal Object Search Performance.

The results of top-k accuracy ($k = 3$), precision, recall, F_1 score, and task completion time are summarized in Table 1 in the Appendix B. A repeated measures ANOVA test shows that **there is no statistically relevant difference in object search accuracy and completion time using the three feedback designs**, with the top-3 accuracy, $F(2, 33) = 0.45$, $p = 0.64$ and selection time, $F(2, 33) = 0.07$, $p = 0.93$.

5.1.2 Task Load.

The NASA TLX ratings, summarized in Figure 6 and Table 2, highlight differences in feedback designs across several subscales. We conducted a series Friedman tests to evaluate the effects of feedback design on NASA TLX subscales.

A significant main effect of feedback design was observed for physical demand ($\chi^2(2) = 6.08$, $p = 0.04$, Kendall's $W = 0.25$). Post-hoc Wilcoxon signed-rank tests with Bonferroni correction revealed that *Itemized* ($M = 1.92$, $SD = 0.99$) imposed significantly **lower physical demand** than the baseline *Speech* ($M = 2.42$, $SD = 1.38$; Wilcoxon $W = 0.00$, $p = 0.03$). No significant differences were found between other pairs.

Itemized demonstrated consistently lower scores across several TLX subscales, including mental demand ($M = 2.17$, $SD = 1.11$), performance ($M = 1.33$, $SD = 0.49$), effort ($M = 2.92$, $SD = 1.24$), and frustration ($M = 1.75$, $SD = 1.06$), although these differences were not statistically significant. Descriptive data are available in Table 2. The trend aligns with the qualitative feedback (Figure 5.2), where participants highlighted the concise and balanced information provided by the *Itemized* approach. These findings suggest potential benefits that warrant future investigation.

5.1.3 Perceived Performance.

In short, we found an **increased perceived speed for Persistent Assistant designs, *Itemized* and *Overview***; and an **increased perceived accuracy for itemized summaries, *Itemized* and *Speech***, although actual speed and accuracy did not yield statistically significant differences (Section 5.1.1). The averages and standard deviations of the ratings are summarized in Table 3, Table 2, and Table 4 in the Appendix B.

Friedman tests show significant main effects of the feedback design on perceived speed, $\chi^2 = 12.0$, $p = 0.002$, Kendall's $W = 0.50$, and perceived accuracy, $\chi^2 = 10.69$, $p = 0.004$, Kendall's $W = 0.44$. Post-hoc Wilcoxon signed-rank tests with Bonferroni correction show significant pairwise differences in perceived speed between *Itemized* ($M = 5.67$, $SD = 0.99$) and *Speech* ($M = 3.83$, $SD = 1.47$), $W = 5.00$, $p = 0.012$, and between *Overview* ($M = 5.00$, $SD = 1.05$) and *Speech* ($M = 3.83$, $SD = 1.47$), $W = 3.00$, $p = 0.019$. The difference in perceived speed between *Itemized* and *Overview* was not statistically significant, $W = 5.00$, $p = 0.11$. For perceived accuracy, post-hoc tests indicate significant differences between *Itemized* ($M = 6.25$, $SD = 0.75$) and *Overview* ($M = 5.42$, $SD = 0.90$), $W = 0.00$, $p = 0.01$, and between *Overview* and *Speech* ($M = 6.33$, $SD = 0.89$), $W = 0.00$, $p = 0.009$.

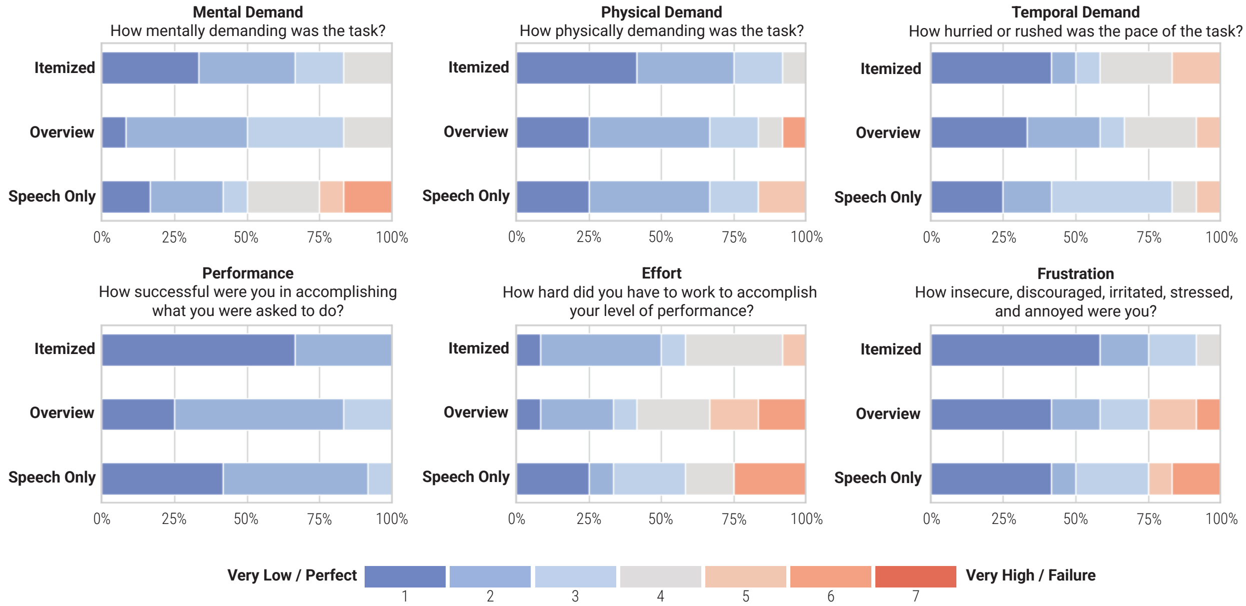

Figure 6: Distribution of NASA Task Load Index responses, with each bar showing the proportion of participants (0-100%) who selected each rating.

5.1.4 Human-AI Language-based Interaction (HALIE). All responses are summarized in Figure 7 with mean and standard deviation values reported in Table 4. All feedback designs were rated positively by most participants in all five criteria to be helpful, enjoyable, satisfactory, responsive, and easy to interact with for task completion. Friedman tests show nonsignificant differences on all criteria among the feedback designs.

5.1.5 Preferences. After all trials, participants ranked the different feedback designs in order of their preference for interacting with AI assistants (Figure 7, bottom right). While subjective preferences varied across participants, all but one participant preferred Persistent Assistant's multimodal summary feedback, either *Itemized* or *Overview*, over *Speech* feedback. The one exception ranked both *Overview* and *Speech* as tied for second place, exhibiting a preference for natural, human-like interactions over time-efficient, robotic interactions.

As part of the study's contextual priming (Section 4.4), participants also experienced a simulation of traditional voice-based conversational interactions with AI assistants for the same task. During the interview, all participants expressed preference for Persistent Assistant's quick and easily repeatable interactions over traditional voice-based conversational interactions for repetitive tasks requiring multiple queries across different items.

5.2 Findings: Dimensions of Feedback in AI-Assisted Interaction

In examining the effectiveness of different feedback modalities, two primary dimensions emerged from our study: the interaction between input specificity and output expressivity, and the differences between speech and haptic modalities in shaping user experience and decision-making.

5.2.1 Specificity of Input vs. Expressivity of Output. Our findings highlight a dynamic interplay between the level of detail in the specificity of user input and the expressiveness of the AI assistant's output.

When the assistant omits some details and provides only the necessary information as in *Overview*, the interaction can be quicker, but it places a greater demand on the user's trust in the system and increases cognitive load. Users may experience uncertainty and feel the need to second-guess their decisions, as P11 explained about the haptic overview condition: *"There is uncertainty. You're not sure what the system actually means with only one buzz."* P11 also described this as a constant decision making: *"Oh, should I listen to it more? Or should I check other items first to do a preview?"*

Along a similar line, when *Overview* provides an ambiguous signal (i.e., 'partially satisfied'), P3 suggested it would be "more natural" if the system says the reason right away without any "buzz." Even after listening to the speech explanation of what is missing in this object, P4 still felt "that reserve of doubt" because they did not have the other criteria. P4 elaborated: *"It's the human tendency to second guess ourselves that kicked in, and trust in her [the assistant's] and the feedback she's giving, or not giving."* This requires the user to actively engage in decision-making, which can be either empowering or induce anxiety depending on their confidence in the system.

On the other hand, when the assistant provides comprehensive information (e.g., a complete list of all item details in *Speech*), users can make decisions with minimal uncertainty, benefiting from the clarity and completeness of the data presented. However, when delivered via speech, it can become repetitive and monotonous.

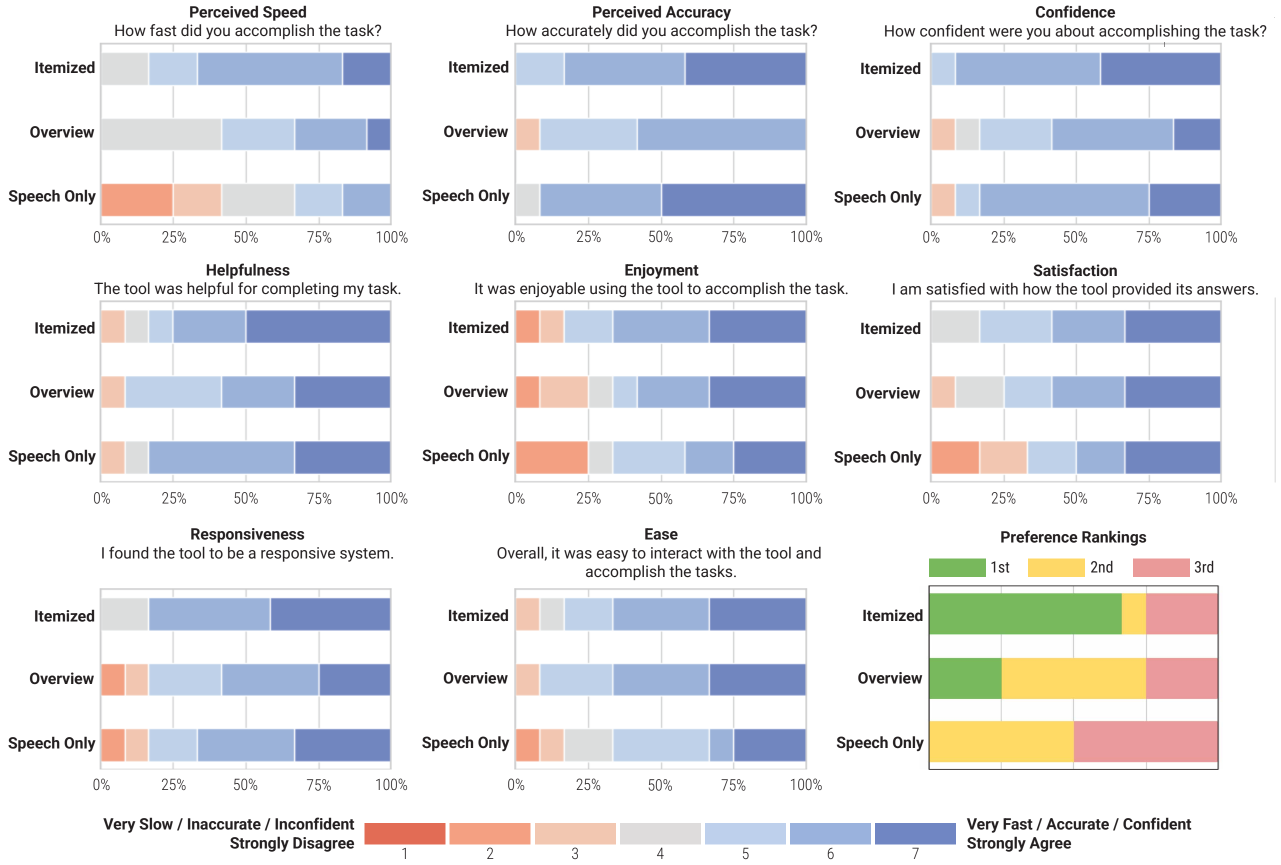

Figure 7: Summary of perceived performance (top row), modified HALIE survey (second row and part of last row), and preference rankings (bottom right).

Participants noted that they perceived the constant stream of information "just too mentally taxing" (P7).

The haptic summary in *Itemized* is a hybrid that balances these dimensions. It provides all the necessary information in a condensed format that is "easier for my brain to process and remember," as described by P7, allowing users to make quick decisions without the verbosity of speech feedback. Participants, like P6, reported that once they "figured out the pattern" or "got the hang of it," they could make decisions much faster, often without needing to listen to the accompanying audio feedback. This approach retained the richness of information while expediting the decision-making process, bridging the gap between too much detail and too little. This is enabled by Persistent Assistant's concept to leverage the capacity of the language model to remember user input and process it on the fly, combined with our haptic feedback design.

Key takeaways.

- *Overview* led to uncertainty in users about what information they contain, and depend on users' trust and experience with the approach.
- *Speech* was appreciated for the additional detail, but was perceived as too invasive and mentally taxing for frequent feedback.
- The haptic *Itemized* balances the amount of information and is well suited for frequent or continuous feedback.

5.2.2 Differences Stemming from Feedback Modalities. Our results revealed that the choice of modality plays a crucial role in shaping user experience. Speech, while universally accessible and easy to understand, was perceived as cumbersome and slow, especially with more details. For example, P6 noted it is "time consuming because I have to listen and be really patient." P4 also added that, when the assistant starts speaking, their "brain took a moment to say, 'Oh, it's her response time.' " This made the participant think "the response from the verbal [feedback]" and their brain's process of "discerning what is being said" are not happening at the same time. The users had to invest more cognitive effort to interpret verbal feedback compared to haptic feedback.

In contrast, haptic feedback was described as being more "instinctive" and "intuitive," offering a direct connection to the user's senses. P8 mentioned that haptic feedback provided a better, quicker understanding with "less energy" and felt "more connected to the senses" than monotonous verbal feedback. The vibrations allowed for immediate recognition and reduced the mental effort required for decision-making. This modality was particularly effective for repetitive tasks where users could quickly learn and internalize the meaning of different haptic patterns, making the interaction feel more "intuitive" and "instantaneous," as P4 describes: *"I became familiar with the vibration. Then, my brain—it already chose."*

However, the effectiveness of the haptics also depended on users' familiarity and comfort with the modality. While many participants described it as a positive and efficient alternative to speech, as reflected in preference rankings (Figure 7), some, like P3, compared the sensation to a "dog buzzer," highlighting a potential negative valence associated with unfamiliar or alert-type vibrations. This feedback underscores the importance of designing haptic cues that are not only functional but also pleasant and engaging, equivalent of natural language feedback like *'This one is great, but it's high calorie.'* as suggested by P3.

Overall, the study suggests that while speech feedback might be better suited for initial learning phases and complex, nuanced conversations, haptic feedback excels in scenarios requiring quick, frequent decisions.

The study also points to the potential of aligning the input and output modalities to strengthen the sensation of instant connection. P11 suggested, for example, if pointing to a target with a finger, *"the haptics even makes more sense because you're mapping the input and output to the same part of your body, very close to each other. I think that's a really nice experience if you can do one. When you experience something in two modalities, it's always annoying. If it occupies two channels, it increases your cognitive load."* P4 shared their account along the similar line: *"The vibration helped the other task [item selection using gaze and gesture] to sink in for me personally."* In summary, the findings indicate that optimizing feedback for Persistent Assistant involves balancing the amount of information provided with the modality used to deliver it.

Key takeaways.

- Speech output was perceived as accessible but cumbersome.
- Haptic feedback was perceived as effective, particularly when the haptic-information mapping can be memorable and easily configured by the user.
- Mapping modalities (e.g., speech query to speech output; pointing to haptic feedback) improved the connection of input and output and improved the general user experience.

6 Discussion

The evaluation highlights how multimodal feedback combined with early intent grounding and embodied inputs can streamline frequent and repetitive AI assistant interactions, presenting a promising alternative to traditional voice-based interactions. Conducting the evaluation in a controlled environment allowed us to isolate and explore specific aspects of the framework's functionality, such as multimodal feedback and embodied input, under consistent conditions. While effective in controlled settings, its deployment in real-world environments presents both challenges and opportunities.

6.1 Challenges in Real-World Deployment

Real-world deployment introduces challenges such as ambiguous target referencing, uncertainty in information access, and latency.

6.1.1 Sloppy and Ambiguous Target Referencing. Accurate target referencing is critical for seamless interactions, but gaze-based methods present several challenges due to inherent tracking errors and noise [8]. These issues become more pronounced in quick, dynamic interactions, where users interact in motion, as brief or unintentional gaze shifts can lead to unintended or incorrect object selection. Advancements in segmentation models such as Segment Anything [48] and SAMURAI [62] offer promising directions. Future work should focus on robust integration into real-time systems.

Similarly, how gaze and segmentation data is input to the LLM influences system accuracy. During our prompt development, we tested various image input formats to convey user's target object. Switching from an egocentric view with a gaze highlight (a white dot and a fixed-size white box around the dot) to a cropped bounding box around the target (Section 3.4.1) improved accuracy by 20.4%. The system sometimes selected the wrong object when the image contained multiple objects, even with highlights. Future work should investigate methods to balance context preservation with precise target communication, potentially through adaptive crop sizes and formats [26].

Interaction techniques can also help users select the correct target through feedback mechanisms. While verbal confirmation (e.g., "The *BBQ flavored potato chip* is not gluten-free, not dairy-free, and high in calories.") provides clarity, it introduces latency. Cursory non-verbal audio-haptic feedback [8] could be a promising alternative by providing quick confirmation while maintaining efficiency.

6.1.2 Object Recognition Reliability and Information Access. In our implementation, we assumed that objects are always recognizable, and sufficient object-related information is available to determine if they meet user criteria. Although feasible with many LLM and online search models, such as Perplexity and ChatGPT, necessary information might not be available online. Future work should consider ways to handle edge cases effectively, for instance, by communicating through a distinct vibration pattern. Xu et al. [60] suggests that users can reliably distinguish up to nine different patterns if well designed, offering opportunities for more nuanced communication.

6.1.3 Latency in Feedback Delivery. The system currently experiences a 4-5 s delay between user input (pinch gesture) and feedback delivery, primarily due to LLM processing times (Section 3.4.1). This latency falls short of the instantaneous interactions expected in real-world applications. Advances in LLM efficiency may reduce processing times, but adjustments to the implementation could further address this issue. For instance, proactively retrieving user intent-related information in the background about all objects in the environment and caching answers could enable near-instantaneous feedback upon user activation.

6.2 Generalizability

Although the technical aspects of our prototype were tested in various environments and tasks (Section 3.4.1), our user study focused on a controlled grocery shopping scenario. We hope to explore the generalizability of Persistent Assistant to real-world interactions with AI assistants in the future, with a particular focus on exploring the diversity of tasks, environments, and evolving user needs. A longitudinal deployment study will provide valuable insights into how users integrate Persistent Assistant into their daily lives over time.

A key question to answer is how well users will be able to recall previously grounded intents, and how often they will need to reground intents across different situations. In dynamic environments, users may expect AI assistants to adapt to their past choices, reducing the need for repeated intent specification. However, contextual factors—such as location, task type, or time of day—may influence whether prior preferences remain relevant. Understanding these factors will help determine whether AI assistants should proactively suggest preferences based on prior interactions or whether users prefer to explicitly reaffirm their needs in each scenario.

Additionally, real-world evaluation will offer insights into the roles of haptic and speech feedback modalities under varying situational constraints. Factors such as environmental noise, user mobility, and task urgency may influence the effectiveness of each modality. Understanding these influences would help balance tradeoffs between privacy, efficiency, and mental load. Since user preferences and accessibility needs vary, these insights could inform the development of personalized or adaptive feedback mechanisms that dynamically adjust to context.

6.3 Expanding the Framework: Opportunities for Future Work

The Persistent Assistant framework offers a flexible foundation for designing context-aware, multimodal AI assistants. In the following, we discuss potential research questions and improvements for each component of the framework.

6.3.1 *Efficient Intent Specification: Grounded Interactions.*

Although our current work focuses on explicit intent grounding—which performed with zero errors in our LLM evaluation—future research could investigate implicit intent grounding to further reduce user effort. By analyzing signals such as users' habitual behaviors and social contexts, an assistant could proactively anticipate user needs. For example, after detecting a potluck event in a calendar invite, the assistant could automatically retrieve dietary preferences of attendees based on memory [69], streamlining tasks like grocery shopping without explicit input. This type of implicit grounding aligns particularly well with the *Overview* feedback design, where users do not need to memorize positional encoding in vibration patterns to interpret criteria. In contrast, the *Itemized* design benefits from explicit verbal articulation of intent, helping users reinforce their awareness of the registered preferences.

Looking ahead, assistants that adapt to both individual and social contexts open up opportunities for deeper investigations into the interplay between explicit and implicit grounding. Questions remain about which scenarios benefit most from implicit assistance versus explicit user input, and about the implications for trust, acceptance, and the system's ability to anticipate needs while maintaining transparency and control.

6.3.2 *Seamless Target Referencing: Embodied Input.*

Eye gaze and pinch gestures represent one form of embodied input, but future work could expand this concept to include more naturalistic behaviors, such as touching or picking up objects. Techniques like semantic segmentation [48], vibroacoustic sensing [16, 28], or RFID [32] could enable robust recognition of these interactions, further reducing the cognitive effort required for target referencing. By integrating embodied input to broader interaction scenarios, we may enhance the assistant's utility in scenarios beyond shopping, such as collaborative tasks or navigation.

6.3.3 *Intuitive Response Comprehension: Multimodal Perceptible Feedback.*

The feedback mechanisms demonstrate the potential of multimodal designs in fostering *intuitive* response comprehension. By leveraging multiple channels—particularly haptic feedback—our approach improved perceived performance and reduced the physical load of interactions, paired with intent grounding. Participants frequently described haptics cues as "instinctive and intuitive" (Section 5.2), highlighting how aligning the input-output modality (for instance, wrist haptics in hand-based interactions) can increase clarity and reduce cognitive effort.

These findings suggest that multimodal feedback provides a valuable additional option, complementing traditional speech-only assistant interactions. By diversifying the channels through which the system and user communicate, it becomes easier to accommodate varying preferences and situational constraints [34]. This approach can streamline information delivery, minimize friction in verbal interactions, and ultimately enhance the overall user experience.

6.3.4 *Components Integration: Contextual Mediator.*

The current implementation processes a single context at a time, with new intent registrations overriding previous settings. Expanding the Contextual Mediator to support persistent context awareness and multi-tasking could improve adaptability. For example, the system could retain recurring intents, allowing users to seamlessly transition between tasks or maintain parallel sessions without reconfiguration. Advanced context awareness could also enable the assistant to prioritize tasks dynamically, further enhancing its flexibility in complex scenarios.

6.4 Reducing Ambiguity in Signals

Ambiguous signals can hinder effective communication between the assistant and the user. While haptic summaries were effective with *Itemized* and the "all satisfied" cue of *Overview*, the "partially satisfied" cue in *Overview* introduced uncertainty about the specific information conveyed. Participants had to hold the pinch gesture longer to wait for the subsequent speech explanation, which disrupted the seamlessness of the interaction.

6.4.1 *Adaptive Feedback Modality Selection.*

To mitigate this uncertainty, one participant suggested skipping ambiguous haptic cues and providing immediate speech explanation. This echoes a suggestion from prior research [23, 38] that when using haptic feedback as the primary information channel, it may require *automatic* translation across modalities. This hints at potential values in future directions to study adaptive multimodal feedback systems that adapt the output modality flexibly to minimize ambiguity.

6.4.2 *Staged Filtering and Flexible Intent Reconfiguration.*

While an automatic modality switching sounds promising, our framework and the study findings also highlight the importance of transparent communication between the user and the assistant for strong grounding. One participant made another suggestion along this line on *staged filtering* through flexible configuration, where users can quickly switch or prioritize the criteria they wanted to apply in real-time. This flexible configuration would enable the assistant to transition along the speech-haptic feedback continuum more swiftly and fluidly, while maintaining the grounding.

6.4.3 Adding Gestural Navigational Interfaces. An extension of the flexible intent configuration would be adding a seamless navigational interface to make the intent configuration even quicker and more structured. From what the findings suggest about the importance of input-output alignment, a gestural navigational interface like Youn et al.'s wrist rotation-based input systems [64, 65] would be a good pair for the haptic-oriented multimodal feedback. Navigational gesture input is a form of the user communicating their quick intent, which establishes the grounding in a flexible manner. Such interfaces would further facilitate application of the Persistent Assistant framework for other types of interactions with AI assistants, such as trigger shortcuts (e.g., controlling smart home devices) or asking questions with freeform responses (e.g., translation support during travel).

6.4.4 Alignment of Input and Output Modalities and Spaces. Participants highlighted the benefits of aligning input and output modalities to improve the intuitiveness of interactions. For example, when a user approaches to grab an item that violates their constraints or preferences, the user receives the haptic feedback on the wrist that is close to the hand grabbing the item. When a user enters a store that is about to close, they get a haptic feedback from the shoes. This modality alignment may help users establish a clear mental model of the interaction, reducing confusion and cognitive load. Investigating and quantifying the effects of input-output alignment on multimodal feedback perception would be an interesting direction to explore.

We envision our Persistent Assistant framework, the intent grounding, and speech-haptic feedback continuum will be the building blocks for future improvements of these interaction techniques.

7 Conclusion

In this paper, we introduced the Persistent Assistant framework and its proof-of-concept implementation, to explore the design of persistent, unobtrusive assistants. We focused on the domain of interacting with AI assistants on a daily basis through frequent, repetitive interactions in various environments. To design low-effort, unobtrusive interactions, Persistent Assistant leverages naturalistic input, early intent grounding, multimodal feedback, and a contextual mediator. Through our proof-of-concept implementation, we demonstrated that combining natural behaviors for target selection, early intent grounding, and balanced multimodal feedback can significantly enhance user experience by providing timely and concise information while minimizing cognitive load. Our user study revealed that the integration of haptic and speech modalities improves the perceived performance, physical load, and preferences in interactions. The Persistent Assistant framework and prototype pave the way for more natural and effortless human-AI interactions for continual, persistent tasks.

Acknowledgments

We thank Romain Nith and Chase Tymms for their valuable insights on haptic interaction; Xingyu Bruce Liu for his thoughtful input on grounding in human-AI interaction; David Moncada, Jom (Pornthep) Preechayasomboon, and Sebastian Freitag for their technical support; and our colleagues at Meta Reality Labs for their feedback, discussions, and support.

References

[1] Karan Ahuja, Sujeath Pareddy, Robert Xiao, Mayank Goel, and Chris Harrison. 2019. Lightanchors: Appropriating point lights for spatially-anchored augmented reality interfaces. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*. 189–196.

[2] Pengcheng An, Ziqi Zhou, Qing Liu, Yifei Yin, Linghao Du, Da-Yuan Huang, and Jian Zhao. 2022. VibEmoji: Exploring user-authoring multi-modal emoticons in social communication. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–17.

[3] 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*. 262–270.

[4] Susan E Brennan. 2014. The grounding problem in conversations with and through computers. In *Social and cognitive approaches to interpersonal communication*. Psychology Press, 201–225.

[5] Stephen Brewster and Lorna M. Brown. 2004. Tactons: structured tactile messages for non-visual information display. In *Proceedings of the Fifth Conference on Australasian User Interface - Volume 28* (Dunedin, New Zealand) *(AUIC '04)*. Australian Computer Society, Inc., AUS, 15–23.

[6] Jessica R. Cauchard, Janette L. Cheng, Thomas Pietrzak, and James A. Landay. 2016. ActiVibe: Design and Evaluation of Vibrations for Progress Monitoring. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 3261–3271. https://doi.org/10.1145/2858036.2858046

[7] Kaifei Chen, Jonathan Fürst, John Kolb, Hyung-Sin Kim, Xin Jin, David E Culler, and Randy H Katz. 2018. Snaplink: Fast and accurate vision-based appliance control in large commercial buildings. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (2018), 1–27.

[8] Hyunsung Cho, Naveen Sendhilnathan, Michael Nebeling, Tianyi Wang, Purnima Padmanabhan, Jonathan Browder, David Lindlbauer, Tanya R. Jonker, and Kashyap Todi. 2024. SonoHaptics: An Audio-Haptic Cursor for Gaze-Based Object Selection in XR *(UIST '24)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3654777.3676384

[9] Adrian A De Freitas, Michael Nebeling, Xiang'Anthony' Chen, Junrui Yang, Akshaye Shreenithi Kirupa Karthikeyan Ranithangam, and Anind K Dey. 2016. Snap-to-it: A user-inspired platform for opportunistic device interactions. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 5909–5920.

[10] Yashar Deldjoo, Johanne R Trippas, and Hamed Zamani. 2021. Towards multi-modal conversational information seeking. In *Proceedings of the 44th International ACM SIGIR conference on research and development in Information Retrieval*. 1577–1587.

[11] Gloria Dhandapani, Jamie Ferguson, and Euan Freeman. 2021. HapticLock: Eyes-Free Authentication for 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, 195–202. https://doi.org/10.1145/3462244.3481001

[12] Mustafa Doga Dogan, Eric J Gonzalez, Andrea Colaco, Karan Ahuja, Ruofei Du, Johnny Lee, Mar Gonzalez-Franco, and David Kim. 2024. Augmented Object Intelligence: Making the Analog World Interactable with XR-Objects. *arXiv preprint arXiv:2404.13274* (2024).

[13] Jacqui Fashimpaur, Amy Karlson, Tanya R Jonker, Hrvoje Benko, and Aakar Gupta. 2023. Investigating Wrist Deflection Scrolling Techniques for Extended Reality. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–16.

[14] Euan Freeman, Stephen Brewster, and Vuokko Lantz. 2016. Do that, there: an interaction technique for addressing in-air gesture systems. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 2319–2331.

[15] Euan Freeman, Gareth Griffiths, and Stephen A Brewster. 2017. Rhythmic micro-gestures: discreet interaction on-the-go. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. 115–119.

[16] Taesik Gong, Hyunsung Cho, Bowon Lee, and Sung-Ju Lee. 2019. Knocker: Vibroacoustic-based object recognition with smartphones. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 3, 3 (2019), 1–21.

[17] John Brandon Graham-Knight, Jon Michael Robert Corbett, Patricia Lasserre, Hai-Ning Liang, and Khalad Hasan. 2021. Exploring Haptic Feedback for Common Message Notification Between Intimate Couples with Smartwatches. In *Proceedings of the 32nd Australian Conference on Human-Computer Interaction* (Sydney, NSW, Australia) *(OzCHI '20)*. Association for Computing Machinery, New York, NY, USA, 245–252. https://doi.org/10.1145/3441000.3441012

[18] Gabriel Haas, Michael Rietzler, Matt Jones, and Enrico Rukzio. 2022. Keep it short: A comparison of voice assistants' response behavior. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–12.

[19] Jooyeun Ham, Jonggi Hong, Youngkyoon Jang, Seung Hwan Ko, and Woontack Woo. 2014. Smart wristband: Touch-and-motion–tracking wearable 3D input device for smart glasses. In *Distributed, Ambient, and Pervasive Interactions: Second International Conference, DAPI 2014, Held as Part of HCI Interational 2014, Heraklion, Crete, Greece, June 22-27, 2014. Proceedings 2*. Springer, 109–118.

[20] Violet Yinuo Han, Hyunsung Cho, Kiyosu Maeda, Alexandra Ion, and David Lindlbauer. 2023. BlendMR: A Computational Method to Create Ambient Mixed Reality Interfaces. *Proceedings of the ACM on Human-Computer Interaction* 7, ISS (2023), 217–241.

[21] 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)*. IEEE, 431–440.

[22] Valentin Heun, James Hobin, and Pattie Maes. 2013. Reality editor: programming smarter objects. In *Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication*. 307–310.

[23] Eve Hoggan and Stephen Brewster. 2007. Designing audio and tactile crossmodal icons for mobile devices. In *Proceedings of the 9th international conference on Multimodal interfaces*. 162–169.

[24] 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: 16th International Conference on Artificial Reality and Telexistence, ICAT 2006, Hangzhou, China, November 29-December 1, 2006. Proceedings*. Springer, 272–283.

[25] Takanori Komatsu, Kazuki Kobayashi, Seiji Yamada, Kotaro Funakoshi, and Mikio Nakano. 2018. Vibrational artificial subtle expressions: Conveying system's confidence level to users by means of smartphone vibration. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–9.

[26] Robert Konrad, Nitish Padmanaban, J Gabriel Buckmaster, Kevin C Boyle, and Gordon Wetzstein. 2024. Gazegpt: Augmenting human capabilities using gaze-contingent contextual ai for smart eyewear. *arXiv preprint arXiv:2401.17217* (2024).

[27] 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*. 1–14.

[28] Gierad Laput, Robert Xiao, and Chris Harrison. 2016. Viband: High-fidelity bio-acoustic sensing using commodity smartwatch accelerometers. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. 321–333.

[29] 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*. 512–520.

[30] 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*. 1–20.

[31] Mina Lee, Megha Srivastava, Amelia Hardy, John Thickstun, Esin Durmus, Ashwin Paranjape, Ines Gerard-Ursin, Xiang Lisa Li, Faisal Ladhak, Frieda Rong, et al. 2022. Evaluating human-language model interaction. *arXiv preprint arXiv:2212.09746* (2022).

[32] Hanchuan Li, Eric Whitmire, Alex Mariakakis, Victor Chan, Alanson P Sample, and Shwetak N Patel. 2019. IDCam: Precise item identification for AR enhanced object interactions. In *2019 IEEE International Conference on RFID (RFID)*. IEEE, 1–7.

[33] Jiahao Nick Li, Yan Xu, Tovi Grossman, Stephanie Santosa, and Michelle Li. 2024. OmniActions: Predicting Digital Actions in Response to Real-World Multimodal Sensory Inputs with LLMs. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–22.

[34] Xingyu Bruce Liu, Jiahao Nick Li, David Kim, Xiang'Anthony' Chen, and Ruofei Du. 2024. Human I/O: Towards a Unified Approach to Detecting Situational Impairments. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–18.

[35] Yang Liu, Chengdong Lin, and Zhenjiang Li. 2021. WR-Hand: Wearable armband can track user's hand. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3 (2021), 1–27.

[36] Pedro Lopes, Patrik Jonell, and Patrick Baudisch. 2015. Affordance++ allowing objects to communicate dynamic use. In *Proceedings of the 33rd annual acm conference on human factors in computing systems*. 2515–2524.

[37] Mathias N Lystbæk, Ken Pfeuffer, Jens Emil Sloth Grønbæk, and Hans Gellersen. 2022. Exploring gaze for assisting freehand selection-based text entry in ar. *Proceedings of the ACM on Human-Computer Interaction* 6, ETRA (2022), 1–16.

[38] Karon E MacLean, Oliver S Schneider, and Hasti Seifi. 2017. Multisensory haptic interactions: understanding the sense and designing for it. In *The Handbook of Multimodal-Multisensor Interfaces: Foundations, User Modeling, and Common Modality Combinations-Volume 1*. 97–142.

[39] Richard E Mayer and Roxana Moreno. 2003. Nine ways to reduce cognitive load in multimedia learning. *Educational psychologist* 38, 1 (2003), 43–52.

[40] 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*. 1–10.

[41] Meta. 2024. https://www.meta.com/blog/quest/surface-emg-wristband-electromyography-human-computer-interaction-hci/?srsltid=AfmBOopWgP7BHvzY22cES0E15K1IROuOVLxk1zA6ztCjII2hdv4SdiyI

[42] 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)*. IEEE, 190–193.

[43] Erik Pescara, Alexander Wolpert, Matthias Budde, Andrea Schankin, and Michael Beigl. 2017. Lifetact: utilizing smartwatches as tactile heartbeat displays in video games. In *Proceedings of the 16th International Conference on Mobile and Ubiquitous Multimedia*. 97–101.

[44] Evan Pezent, Aakar Gupta, Hank Duhaime, Marcia O'Malley, Ali Israr, Majed Samad, Shea Robinson, Priyanshu Agarwal, Hrvoje Benko, and Nick Colonnese. 2022. Explorations of wrist haptic feedback for AR/VR interactions with Tasbi. In *Adjunct Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–5.

[45] Evan Pezent, Ali Israr, Majed Samad, Shea Robinson, Priyanshu Agarwal, Hrvoje Benko, and Nick Colonnese. 2019. Tasbi: Multisensory squeeze and vibrotactile wrist haptics for augmented and virtual reality. In *2019 IEEE World Haptics Conference (WHC)*. IEEE, 1–6.

[46] Martin Pielot, Anastasia Kazakova, Tobias Hesselmann, Wilko Heuten, and Susanne Boll. 2012. PocketMenu: non-visual menus for touch screen devices. In *Proceedings of the 14th international conference on Human-computer interaction with mobile devices and services*. 327–330.

[47] 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)*. IEEE, 73–82.

[48] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. 2024. SAM 2: Segment Anything in Images and Videos. *arXiv preprint arXiv:2408.00714* (2024). https://arxiv.org/abs/2408.00714

[49] 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*. 210–214.

[50] Anne Roudaut, Andreas Rau, Christoph Sterz, Max Plauth, Pedro Lopes, and Patrick Baudisch. 2013. Gesture output: eyes-free output using a force feedback touch surface. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2547–2556.

[51] Hasti Seifi, Kailun Zhang, and Karon E MacLean. 2015. VibViz: Organizing, visualizing and navigating vibration libraries. In *2015 IEEE World Haptics Conference (WHC)*. IEEE, 254–259.

[52] Jannis Strecker, Khakim Akhunov, Federico Carbone, Kimberly García, Kenan Bektaş, Andres Gomez, Simon Mayer, and Kasim Sinan Yildirim. 2023. MR Object Identification and Interaction: Fusing Object Situation Information from Heterogeneous Sources. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 3 (2023), 1–26.

[53] Akifumi Takahashi, Yudai Tanaka, Archit Tamhane, Alan Shen, Shan-Yuan Teng, and Pedro Lopes. 2024. Can a Smartwatch Move Your Fingers? Compact and Practical Electrical Muscle Stimulation in a Smartwatch. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–15.

[54] Yudai Tanaka, Neil Weiss, Robert Cole Bolger-Cruz, Jess Hartcher-O'Brien, Brendan Flynn, Roger Boldu, and Nicholas Colonnese. 2024. ReaWristic: Remote Touch Sensation to Fingers from a Wristband via Visually Augmented Electro-Tactile Feedback. In *2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 951–960.

[55] Dimitar Valkov and Lars Linsen. 2019. Vibro-tactile feedback for real-world awareness in immersive virtual environments. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 340–349.

[56] Alexander Wang, Yi Fei Cheng, and David Lindlbauer. 2024. MARingBA: Music-Adaptive Ringtones for Blended Audio Notification Delivery. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–15.

[57] Alexander Wang, David Lindlbauer, and Chris Donahue. 2024. Towards Music-Aware Virtual Assistants. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–14.

[58] 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.

[59] Robert Xiao, Gierad Laput, Yang Zhang, and Chris Harrison. 2017. Deus EM Machina: on-touch contextual functionality for smart IoT appliances. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 4000–4008.

[60] Shan Xu, Sarah Sykes, Parastoo Abtahi, Tovi Grossman, Daylon Walden, Michael Glueck, and Carine Rognon. 2024. Designing Haptic Feedback for Sequential Gestural Inputs. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–17.

[61] Xuhai Xu, Anna Yu, Tanya R Jonker, Kashyap Todi, Feiyu Lu, Xun Qian, João Marcelo Evangelista Belo, Tianyi Wang, Michelle Li, Aran Mun, et al. 2023. Xair: A framework of explainable ai in augmented reality. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–30.

[62] Cheng-Yen Yang, Hsiang-Wei Huang, Wenhao Chai, Zhongyu Jiang, and Jenq-Neng Hwang. 2024. SAMURAI: Adapting Segment Anything Model for Zero-Shot Visual Tracking with Motion-Aware Memory. arXiv:2411.11922 [cs.CV] https://arxiv.org/abs/2411.11922

[63] Jackie Yang and James A Landay. 2019. Infoled: Augmenting led indicator lights for device positioning and communication. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*. 175–187.

[64] Eunhye Youn, Taejun Kim, and Geehyuk Lee. 2024. WristMenu with Tactons: An Eyes- and Ears-Free Menu with Tactons Describing Menu Items in the Wrist Rotation Space. *International Journal of Human–Computer Interaction* 40, 9 (2024), 2314–2325. https://doi.org/10.1080/10447318.2022.2159780 arXiv:https://doi.org/10.1080/10447318.2022.2159780

[65] Eunhye Youn, Sangyoon Lee, Sunbum Kim, Youngbo Aram Shim, Liwei Chan, and Geehyuk Lee. 2021. WristDial: An Eyes-Free Integer-Value Input Method by Quantizing the Wrist Rotation. *International Journal of Human–Computer Interaction* 37, 17 (2021), 1607–1624. https://doi.org/10.1080/10447318.2021.1898848 arXiv:https://doi.org/10.1080/10447318.2021.1898848

[66] Eric M Young, Amirhossein H Memar, Priyanshu Agarwal, and Nick Colonnese. 2019. Bellowband: A pneumatic wristband for delivering local pressure and vibration. In *2019 IEEE World Haptics Conference (WHC)*. IEEE, 55–60.

[67] 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*. 246–253.

[68] Kaixing Zhao, Marcos Serrano, Bernard Oriola, and Christophe Jouffrais. 2020. VibHand: On-Hand Vibrotactile Interface Enhancing Non-Visual Exploration of Digital Graphics. *Proc. ACM Hum.-Comput. Interact.* 4, ISS, Article 207 (nov 2020), 19 pages. https://doi.org/10.1145/3427335

[69] Wazeer Deen Zulfikar, Samantha Chan, and Pattie Maes. 2024. Memoro: Using Large Language Models to Realize a Concise Interface for Real-Time Memory Augmentation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–18.

A Test Cases for Prompt Evaluation

A.1 Couches

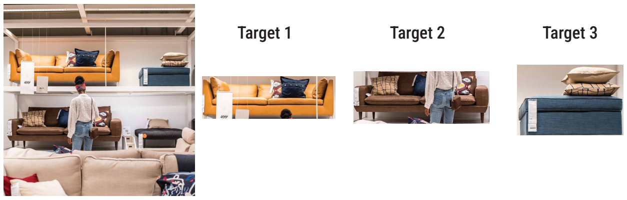

Figure 8: Target images used in prompt evaluation for a couch shopping scenario.

Object States.

- Mustard Leather Couch: leather material, does not recline, fits standard door, assembly required, mid-century modern style.
- Brown Suede Couch: suede material, does not recline, fits standard door, no assembly required, firm cushions.
- Blue Ottoman Couch: fabric material, does not recline, fits standard door, assembly required, includes hidden storage.
- Beige Plush Couch: synthetic fabric, does not recline, fits standard door, no assembly required, stain-resistant.

Intent Prompts.

(1) "I'm at the furniture store. I'm looking for a couch that reclines, is made of leather, and fits through a standard door."
(2) "Looking at couches in the furniture store, I'm searching for one that is comfortable, doesn't require assembly, and isn't made of plastic."
(3) "I want a couch that is stain-resistant, doesn't use synthetic fabric, and has a pull-out bed."

A.2 Cars

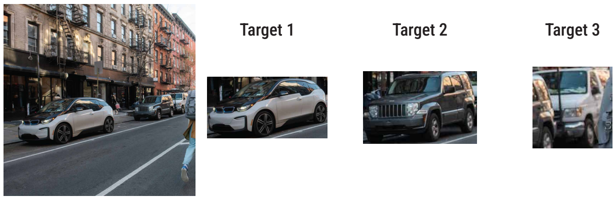

Figure 9: Target images used in prompt evaluation for a car shopping scenario.

Object States.

- White BMW i3: electric vehicle, fabric seats, compact size, price under $40,000, supports Apple CarPlay, not suitable for off-roading, small trunk space.
- Black Jeep Cherokee: fuel-efficient for an SUV, leather seats, price above $40,000, good for off-roading, heated seats, supports Apple CarPlay, hybrid option available, over 15 feet long, moderate trunk space.
- White Van: gas-powered, basic interior, no heated seats, no Apple CarPlay, not good for off-roading, under $40,000, over 15 feet long, large trunk space.

Intent Prompts.

(1) "I'm looking for a vehicle that is fuel-efficient, doesn't have leather seats, and is priced under $40,000."
(2) "I'm comparing cars. I want one that has heated seats, supports Apple CarPlay, and good for off-roading."
(3) "I want to compare cars that are hybrid, not over 15 feet long, and have a large trunk space."

A.3 Plants

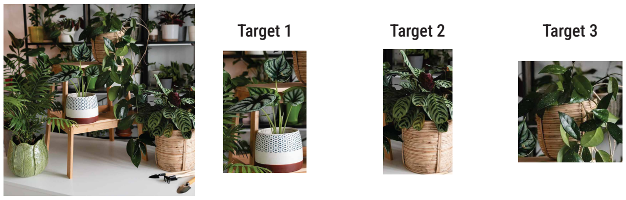

Figure 10: Target images used in prompt evaluation for a plant shopping scenario.

Object States.

- Parlor Palm (Chamaedorea elegans): low-maintenance, pet-friendly, thrives in low light, needs moderate watering, non-toxic to pets.
- Calathea (Calathea makoyana): non-toxic to pets, thrives in low light, medium maintenance, needs high humidity, native to tropical regions.
- Hoya (Hoya carnosa): low-maintenance, toxic to pets, prefers bright indirect light but tolerates low light, blooms periodically.
- Alocasia (Alocasia Polly): medium maintenance, toxic to pets, prefers bright indirect light, needs consistent watering, striking foliage.

Intent Prompts.

(1) "I need plants that are low-maintenance and pet-friendly for a dimly lit living room."
(2) "I'm looking for plants that are native to the region, not toxic to pets, and thrive in low sunlight."
(3) "I need flowers that bloom year-round, don't attract bees, and require minimal watering."

A.4 Restaurant Menu

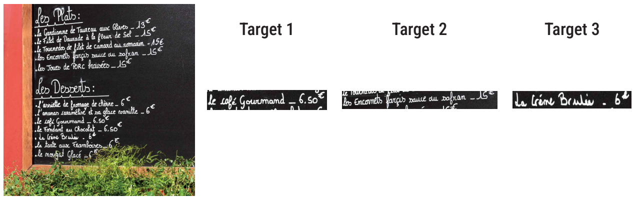

Figure 11: Target images used in prompt evaluation for a restaurant menu, where the targets are menu items.

Object States.

- La Gardianne de Taureau aux Olives: gluten-free, contains meat, no mayonnaise, not served with a side salad.
- Le Filet de Daurade à la Fleur de Sel: gluten-free, dairy-free, contains fish, no mayonnaise, served with a side salad.
- Le Tournedos de Filet de Canard au Romarin: gluten-free, contains dairy (butter), no mayonnaise, not served with a side salad.
- Les Encornets Farcis Sauce au Safran: contains gluten (stuffing), dairy, no mayonnaise, served with a side salad.
- Les Joues de Porc Braisées: gluten-free, contains meat, no mayonnaise, served with a side salad.
- L'assiette de Fromage de Chèvre: contains dairy, gluten-free, no nuts, no added sugar, no fresh fruit.
- L'ananas Caramélisé et sa Glace Vanille: dairy (ice cream), no gluten, no nuts, no added sugar, fresh fruit included.
- Le Café Gourmand: contains gluten (small pastries), dairy, no nuts, added sugar, no fresh fruit.
- Le Fondant au Chocolat: contains dairy, gluten, no nuts, added sugar, no fresh fruit.
- La Crème Brûlée: contains dairy, gluten-free, no nuts, added sugar, no fresh fruit.
- La Tarte aux Framboises: contains gluten, dairy, no nuts, added sugar, fresh fruit included.
- Le Nougat Glacé: contains dairy, nuts, no gluten, added sugar, no fresh fruit.

Intent Prompts.

(1) "Looking at this café menu, I want a dish that is gluten-free, doesn't include mayonnaise, and is served with a side salad."
(2) "I'm deciding on a dish that is dairy-free, doesn't contain nuts, and includes chocolate."
(3) "I want food that are vegan, don't have added sugar, and come with fresh fruit."

A.5 Restaurants

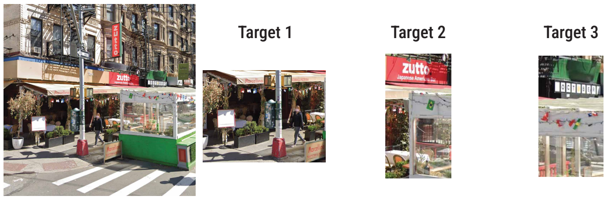

Figure 12: Target images used in prompt evaluation for restaurant search.

Object States.

- Marcellino Trattoria & Pizzeria: serves vegetarian dishes, serves alcohol, offers outdoor seating, plays soft background music, specializes in Italian cuisine, has family-sized portions, offers a kids' menu, does not accept mobile payments.
- Zutto Japanese American Pub: serves vegetarian options, does not specialize in ramen, serves alcohol, has indoor seating only, plays moderate music, specializes in Japanese-American fusion, accepts mobile payments, portions are individual-sized.
- Restaurant OCTOBER (Green Sign): limited vegetarian options, serves alcohol, offers outdoor seating, plays loud music, specializes in a casual bar menu, does not accept mobile payments, portions are individual-sized.

Intent Prompts.

(1) "Walking down this street, I'm looking for a restaurant that serves vegetarian dishes, doesn't play loud music, and has outdoor seating."
(2) "On this street, I want to find a place that specializes in ramen, isn't crowded right now, and accepts mobile payments."
(3) "Looking at these restaurants, I need one that offers family-sized portions, doesn't serve alcohol, and has a kids' menu."

A.6 Snacks

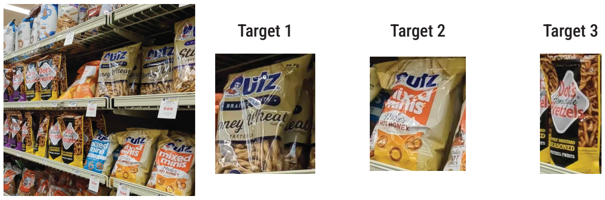

Figure 13: Target images used in prompt evaluation for snack shopping in a supermarket

Object States.

- Dot's Homestyle Pretzels (Seasoned): contains gluten, no artificial flavors, 200 calories per serving, not vegan, contains palm oil, recyclable packaging, low protein, does not contain peanuts, moderate sugar content.
- Utz Honey Wheat Pretzels: contains gluten, contains natural honey flavor, 180 calories per serving, not vegan (contains honey), no palm oil, recyclable packaging, moderate protein, does not contain peanuts, moderate sugar content.
- Utz Mixed Minis (Sea Salt): gluten-free, no artificial flavors, 130 calories per serving, vegan, contains palm oil, non-recyclable packaging, low protein, does not contain peanuts, low sugar content.
- Utz Mixed Minis (Hot Honey): contains gluten, contains natural honey flavor, 160 calories per serving, not vegan (contains honey), no palm oil, recyclable packaging, low protein, does not contain peanuts, moderate sugar content.
- Snack Factory Pretzel Crisps (Sea Salt): contains gluten, no artificial flavors, 110 calories per serving, vegan, no palm oil, recyclable packaging, low protein, does not contain peanuts, low sugar content.

Intent Prompts.

(1) "In this snack aisle, I want chips that are gluten-free, don't have artificial flavors, and are under 150 calories per serving."
(2) "Looking at these options, I'm choosing snacks that are vegan, not made with palm oil, and come in recyclable packaging."
(3) "In the snack section, I'm looking for one that is high in protein, doesn't contain peanuts, and has low sugar content."

B User Study Results

Table 1: Optimal object search performance. Averages (M) and standard deviations (SD) of top-3 accuracy, precision, recall, F_1 score, and task completion time.

Feedback	Top-3 Accuracy		Precision		Recall		F_1 Score		Completion Time	
	M	SD	M	SD	M	SD	M	SD	M	SD
Itemized	0.801	0.249	0.709	0.275	0.808	0.232	0.745	0.249	32.4 s	23.3 s
Overview	0.788	0.247	0.749	0.277	0.792	0.242	0.766	0.261	33.5 s	18.1 s
Speech	0.830	0.191	0.715	0.254	0.836	0.191	0.760	0.219	35.8 s	12.4 s

Table 2: Simplified NASA Task Load Index on a 7-point Likert scale, where 1 is 'Very Low' (or 'Perfect' for Performance), and 7 is 'Very High' (or 'Failure' for Performance). The feedback design showed a main effect on Physical Demand ($\chi^2(2) = 6.08, p = 0.0478$, Kendeall's W = 0.225). A post-hoc Friedman test showed a significant difference between *Itemized* and *Speech* (W = 0.00, p = 0.0339). $^*p < 0.05$

Feedback	Mental Demand		Physical Demand*		Temporal Demand		Performance		Effort		Frustration	
	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Itemized	2.167	1.115	1.917	0.996	2.667	1.670	1.333	0.492	2.917	1.240	1.750	1.055
Overview	2.583	0.900	2.417	1.443	2.500	1.446	1.917	0.669	3.667	1.670	2.583	1.832
Speech	3.333	1.775	2.417	1.379	2.583	1.240	1.667	0.651	3.333	1.923	2.750	1.960

Table 3: Perceived performance on a 7-point Likert scale, 1 being 'Very Low' and 7 being 'Very High'. The feedback design showed main effects on perceived speed ($\chi^2(2)$ = 12.00, p = 0.0025, Kendall's W = 0.500) and perceived accuracy ($\chi^2(2)$ = 10.69, p = 0.0048, Kendall's W = 0.445). A post-hoc Friedman test showed significant pairwise differences on perceived speed between *Itemized* and *Speech* (W = 5.00, p = 0.0122) and between *Overview* and *Speech* (W = 3.00, p = 0.0186); and perceived accuracy between *Itemized* and *Overview* (W = 0.00, p = 0.0141) and between *Overview* and *Speech* (W = 0.00, p = 0.0094). $^*p < 0.05$; $^{**}p < 0.01$.

Feedback	Perceived Speed**		Perceived Accuracy**		Confidence	
	M	SD	M	SD	M	SD
Itemized	5.667	0.985	6.250	0.754	6.333	0.651
Overview	5.000	1.045	5.417	0.900	5.500	1.168
Speech	3.833	1.467	6.333	0.888	5.917	1.084

Table 4: Modified HALIE Questionnaire on a 7-point Likert scale, where 1 is 'Strongly Disagree, and 7 is 'Strongly Agree'.

Feedback	Helpfulness		Enjoyment		Satisfaction		Responsiveness		Ease	
	M	SD	M	SD	M	SD	M	SD	M	SD
Itemized	6.000	1.348	5.583	1.621	5.750	1.138	6.083	1.084	5.750	1.288
Overview	5.750	1.215	5.250	1.815	5.583	1.379	5.417	1.564	5.833	1.193
Speech	5.917	1.240	4.833	1.946	5.000	2.000	5.583	1.621	5.000	1.595