AROMA: Mixed-Initiative AI Assistance for Non-Visual Cooking by Grounding Multimodal Information Between Reality and Videos

Zheng Ning
zning@nd.edu
University of Notre Dame
Notre Dame, Indiana, USA

Leyang Li
lli27@nd.edu
University of Notre Dame
Notre Dame, Indiana, USA

Daniel Killough
dkillough@wisc.edu
University of Wisconsin-Madison
Madison, Wisconsin, USA

JooYoung Seo
jseo1005@illinois.edu
University of Illinois
Urbana-Champaign, Illinois, USA

Patrick Carrington
pcarrington@cmu.edu
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

Yapeng Tian
yapeng.tian@utdallas.edu
University of Texas at Dallas
Richardson, Texas, USA

Yuhang Zhao
yuhang.zhao@cs.wisc.edu
University of Wisconsin-Madison
Madison, Wisconsin, USA

Franklin Mingzhe Li
mingzhe2@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

Toby Jia-Jun Li
toby.j.li@nd.edu
University of Notre Dame
Notre Dame, Indiana, USA

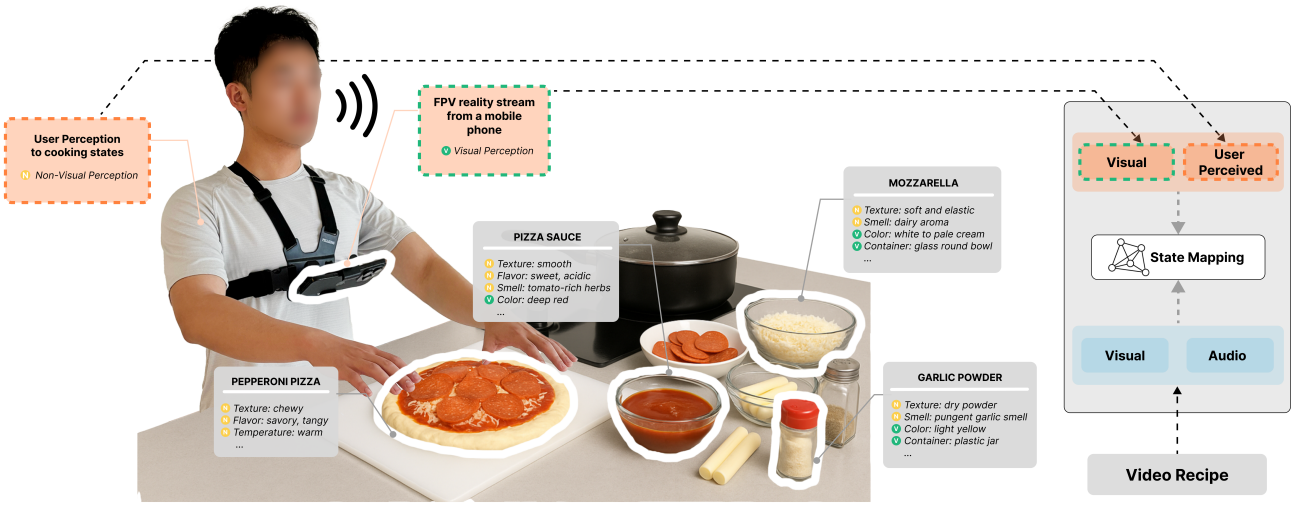

Figure 1: An illustration of how a blind or low-vision (BLV) user uses the AROMA in the kitchen. AROMA helps the user access video recipes by allowing them to communicate their perceived non-visual information (N) about the food, such as texture, smell, taste, etc., along with real-time visual information captured by a wearable camera (V). The system then responds to their questions, referring to the knowledge from the video recipe during cooking. Meanwhile, AROMA proactively monitors the cooking process through the real-time video stream and raises alerts when specific criteria are met.

Abstract

Videos offer rich audiovisual information that can support people in performing activities of daily living (ADLs), but they remain largely inaccessible to blind or low-vision (BLV) individuals. In cooking, BLV people often rely on *non-visual* cues—such as touch, taste, and smell—to navigate their environment, making it difficult to follow the predominantly *audiovisual* instructions found in video recipes. To address this problem, we introduce AROMA, an AI system that provides timely responses to the user based on real-time, context-aware assistance by integrating non-visual cues perceived by the user, a wearable camera feed, and video recipe content. AROMA uses a mixed-initiative approach: it responds to user requests while also proactively monitoring the video stream to offer timely alerts and guidance. This collaborative design leverages the complementary

UIST '25, Busan, Republic of Korea

ACM ISBN 979-8-4007-2037-6/2025/09
https://doi.org/10.1145/3746059.3747650

strengths of the user and AI system to align the physical environment with the video recipe, helping the user interpret their current state and make sense of the steps. We evaluated Aroma through a study with eight BLV participants and offered insights for designing interactive AI systems to support BLV individuals in performing ADLs.

CCS Concepts

• **Human-centered computing** → **Accessibility systems and tools**; **User interface programming**.

Keywords

video recipes, cooking, multimodal perception, accessibility

ACM Reference Format:

Zheng Ning, Leyang Li, Daniel Killough, JooYoung Seo, Patrick Carrington, Yapeng Tian, Yuhang Zhao, Franklin Mingzhe Li, and Toby Jia-Jun Li. 2025. Aroma: Mixed-Initiative AI Assistance for Non-Visual Cooking by Grounding Multimodal Information Between Reality and Videos. In *The 38th Annual ACM Symposium on User Interface Software and Technology (UIST '25), September 28-October 1, 2025, Busan, Republic of Korea.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3746059.3747650

1 Introduction

Videos have become a critical resource for learning how to perform activities of daily living (ADLs) [35, 47]. For instance, platforms like YouTube have more than 500 hours of video content uploaded every minute [75], making it one of the most lively communities to share instructional content [27]. As more instructions are shared exclusively in video format, videos have become an indispensable source for certain types of knowledge [4, 56]. Unlike single-modality formats such as text or static images, videos offer rich, multimodal instruction through visual demonstrations, spoken explanations, and audio cues. Despite their advantages, instructional videos for ADLs remain less accessible to blind or low-vision (BLV) individuals [41], as those videos are often designed with sighted viewers in mind. Specifically, in videos, instructors frequently rely on visual demonstrations without offering sufficient verbal detail, or they use vague references like "this" or "that" without context [41], making it difficult for BLV users to follow along.

Although efforts such as audio descriptions—narrated explanations of key visual elements, actions, and scene transitions [15, 67]—and video question-answering systems [8, 74] aim to improve accessibility, instructional videos still pose challenges for BLV users attempting to perform tasks in the real world. One key issue is the mismatch between what is demonstrated in the video and what the user is experiencing in real-time. Specifically, BLV users primarily rely on non-visual sensory modalities, such as tactile feedback, smell, and sound, to track progress and make decisions during the activity [34, 35]; however, these sensory modalities are frequently underrepresented or acknowledged in video content. As a result, BLV users often: (i) struggle to determine how their current state compares with what is shown in the video, and (ii) find it difficult to decide on the next appropriate steps based on their own sensory experiences.

Among the many activities demonstrated through instructional videos, cooking is particularly important. It is essential to support independence, health and emotional well-being [10, 41, 57]. Prior research shows that people with vision impairments are enthusiastic about video recipes, as they often include rich auditory cues (e.g., sizzling sounds) and are frequently produced by professional chefs [33, 41]. However, visually impaired cooks primarily consume these videos for entertainment or inspiration, rather than for direct task guidance, due to the lack of detailed visual descriptions [41]. This disconnect highlights a fundamental challenge for BLV individuals — bridging the gap between their *non-visual* sensory inputs (e.g., aroma, texture, or sound) and the instructional content primarily conveyed through *visual* and *auditory* modalities in video recipes.

Prior research has offered valuable insights for designing AI systems that support BLV individuals in cooking by referencing video recipes [34, 41, 47]. However, a critical gap remains: enabling collaboration between users and AI agents that leverages their complementary perceptual and cognitive strengths to align real-time, multimodal information from the physical environment with instructional knowledge in the video.

To address this challenge, we introduce Aroma[1], an interactive AI system that enables real-time collaboration between a BLV user and a multimodal agent during cooking. Users can interact with the system to access procedural step information, verify the current state of cooking, and receive targeted guidance for correcting errors, etc., by directly querying the video recipe based on their own non-visual perceptions of the cooking process. Following a mixed-initiative design paradigm [22], the system also continuously monitors the cooking states through a wearable camera feed, proactively analyzing the scene and raising alerts when misalignments are detected.

We evaluated Aroma through a user study with eight BLV participants. Each user study session took place in the participant's own kitchen or another preferred location by the participant. During the user study, each participant used the system to reproduce a dish of their choice from a list of three video recipes. Based on the results, we assess the system's usability and identify key contextual challenges BLV users face when cooking with video-based instructions. We also propose design implications to inform the development of interactive AI systems aimed at improving video accessibility and supporting BLV individuals in performing ADLs.

To sum up, our paper presents the following contributions:

- Aroma, an interactive mixed-initiative multimodal AI system that supports BLV users in following instructional cooking videos by leveraging the complementary perceptual and cognitive strengths of the user and the system to align information from the physical environment with instructional knowledge in the video.
- A user study with eight BLV participants to validate the usability and effectiveness of Aromain realistic cooking tasks.
- Design insights and findings for designing interactive AI systems that help BLV individuals access multimodal instructions and perform daily activities.

[1]Aroma is an acronym for **A**ugmenting **R**ecipe **O**rchestration with **M**ultimodal **A**ssistance

2 Related Work

In this section, we review three key areas of related work that inform our research. First, we explore theoretical frameworks and assistive approaches designed to support individuals with sensory impairments in performing daily activities (Sec. 2.1). Second, we discuss prior empirical findings and system work specifically addressing the challenges faced by BLV individuals in cooking (Sec. 2.2). Lastly, we review research on AI models and interactive systems developed to enhance video content accessibility for BLV users (Sec. 2.3). Throughout this review, we highlight how Aroma builds upon and extends these existing approaches by integrating real-time visual information with users' inherent non-visual perceptual capabilities in real-time mixed-initiative cognitive assistance for activities of daily living.

2.1 Assisting People with Sensory Impairments in Performing Daily Activities

Prior research has explored various approaches to support people with sensory impairments in performing daily activities, which can be categorized into two main areas: (i) facilitating cognitive processes to compensate for sensory input limitations such as visual, auditory, and olfactory [19, 24, 26, 44, 55, 58, 60, 69], and (ii) assisting users in overcoming physical barriers that hinder task execution [32, 35, 38, 39, 45]. Examples of the latter include designing Augmentative and Alternative Communication (AAC) for non-verbal users [70], using eye-tracking devices for hands-free manipulation [21], designing accessible robots [12], etc. Our work is more closely aligned with the first category, focusing on addressing cognitive challenges that arise when one or more perceptual modalities are impaired.

Key theoretical frameworks inform this space. The design principles of mixed-initiative user interfaces [22] highlighted the importance of system-initiated actions in helping users detect and recover from errors. This is particularly crucial for individuals with sensory impairments, who may be unable to perceive such errors independently [48, 64]. Similarly, multimodal disambiguation theory highlights the importance of integrating information across multiple modalities to more effectively interpret user intent [62].

Building on those theories, prior research has applied assistive approaches across a variety of scenarios. To enhance navigation capabilities for BLV users, prior work has combined auditory and haptic cues to support spatial awareness [14, 23, 65]; Patil et al. explored using additional gestures on white canes to control the smart device of a BLV user [63]; Zhao et al. augmented the audiovisual information in a virtual reality (VR) scenario to enhance content understanding for low-vision users [76]. For Deaf and Hard of Hearing (DHH) users, prior work has explored augmenting sound effects in VR [7], adding extra indicators [46] and haptics [59] to deliver a better experience for them; and captioning and visualizing non-speech sounds for a better video consumption experience [3]. Furthermore, the integration of olfactory feedback into VR environments has been explored to enhance user immersion and potentially aid those with olfactory impairments [52, 53].

The design rationales of Aroma are informed by the theories and insights discussed above. Specifically, cooking, as a naturally multimodal activity, requires individuals to integrate information from various sensory channels to understand and manage the process [33, 34]. Aroma enables BLV users to communicate their non-visual sensory observations to the system and pairs this input with real-time visual analysis. This human-AI collaboration follows a mixed-initiative paradigm to bridge the gap between sensory perception and visual instruction, offering timely, context-aware support throughout the cooking process.

2.2 Assisting BLV People in Cooking

Cooking is an important daily activity that supports independence and improves the quality of life for blind or low-vision (BLV) individuals [10, 33, 57]. Prior research has examined difficulties across various stages and design insights for assistive systems that support this task [33].

One major challenge lies in accessing and interpreting recipes. Text-based recipes are traditionally adopted because they are easy to follow through OCR [66] and text-to-speech [61]; However, they often omit critical visual context. For instance, what is *"cook until golden brown"* remains unclear [41]. Video-based recipes, while rich in multi-modal information, introduce new difficulties, such as lacking structured navigation, insufficient verbalization of visual content, and increased cognitive load required to recall [25, 41]. Although video recipes were reported as less accessible, the multimodal information was reported as entertaining and inspiring for people with vision impairments [41]. Recent research has also examined how BLV users engage with cooking instructions across modalities. For instance, Li et al. found that BLV cooks often prefer structured, chunked formats, tactile representations, and hands-free interaction mechanisms [41]. Strategies like reordering steps, simplifying language, and annotating sensory checkpoints (e.g., smells, sounds) can enhance recipe usability.

Another significant hurdle is recognizing and interpreting real-time cooking states, such as ingredient readiness, food completion, or the location of utensils. Li et al [34] conducted a contextual inquiry study, identifying eight classes of contextual information that BLV people actively seek. These include spatial layout, object status, and dynamic properties like temperature or completion, which serve important design rationales for Aroma. It has also pointed out that BLV users would develop intentional, embodied associations with objects (e.g., placing a spoon at a known angle) to ground information retrieval in spatial memory. These workarounds, however, are effective but fragile, especially in situations involving multitasking or shared kitchen environments.

To address those issues, one relevant system is CookAR [29], which augments the affordance of appliances and objects in cooking through a head-mounted AR system for low-vision users. Another similar system is OSCAR [36, 37], which provides context-aware feedback to the user when a task is completed by tracking object statuses. Specifically, OSCAR focuses on Step prediction in cooking, demonstrating that tracking object status changes in video significantly improves the accuracy, and step prediction is one of the important features in assisting non-visual cooking. In contrast, AROMA explores the dynamics of a real-time, human-AI partnership that focuses on assisting the cooking process as a whole, rather than the order of a particular step.

Aroma's voice-based model and mixed-initiative architecture are grounded in this body of research. The system is designed to align with BLV users' established preferences for interaction and feedback. For example, when describing a cooking step, the system provides a concise explanation that includes the step name, estimated duration, and expected outcome—an approach consistent with the preferred instruction style identified in prior work.

2.3 Accessing Video Content for BLV Individuals

Accessing video content remains a significant challenge for BLV individuals [50, 51]. Prior research in this area can be broadly categorized into two areas: i) the support for sequential video consumption, where information is accessed in temporal order; ii) non-sequential information retrieval, where key information from the video is extracted based on user needs without requiring users to watch the entire video.

Audio descriptions (AD) are a primary method for enabling temporal video consumption by inserting narrated descriptions of visual content either inline (without pausing the video) or through extended descriptions (with pauses for additional narration) [1]. Prior research in this space has looked at automating the generation of ADs [72] and new interaction paradigms such as layered audio descriptions that allow users to explore video content non-linearly while preserving the temporal flow [60].

When used in an information retrieval context, systems have been developed to help BLV users retrieve relevant information without watching the entire video. For example, Shortscribe [71] presents hierarchical summaries of short-form videos to quickly convey the gist. Another popular approach is to support question answering (QA) based on videos. For example, the AI community has made progress in Video Question Answering (VideoQA), where models generate natural language answers to queries grounded in video content directly [30, 43, 68]. Recent foundation models, trained on vast audiovisual corpora, have further pushed the boundaries of open-ended video understanding [16, 18].

The design of Aroma builds on these efforts by exploring how video recipes can serve as a rich instructional resource in real-world cooking scenarios, where the physical context often diverges from what is shown in the video. From a human-centered design perspective, Aroma also investigates how to leverage the non-visual perceptual strengths of BLV users, such as touch, smell, and sound, to support multi-modal interaction with video content, enabling users to cook effectively while accessing and aligning relevant information from the recipe video.

3 Aroma System

We designed and implemented Aroma, a mixed-initiative system that couples BLV users' non-visual cues with real-time visual information to support users in cooking. The system offers on-demand, conversational assistance and proactively detects errors during the cooking process by aligning information from the physical environment with instructional knowledge in the video and provide timely and context-aware guidance to the user.

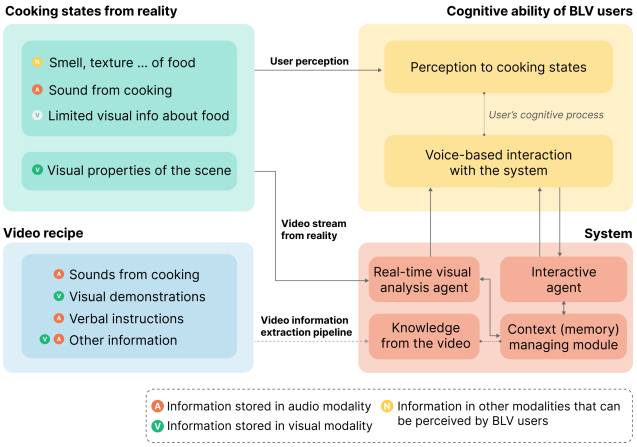

Figure 2: An illustration of the architecture of the system, and the corresponding input/output for each component. The system comprises four modules to extract knowledge from video recipes, analyze visual information from real-world cooking states, interact with the user through sound, and a module to manage history and agent context.

3.1 Design Goals

We identified the following design goals for Aroma, inspired by prior findings on non-visual cooking (detailed in Sec. 2.2).

DG1 **Provide both on-demand and proactive support.** The system should respond promptly to requests initiated by the user to address their immediate needs. This is particularly important given the cognitive demands of cooking and the difficulty of transferring knowledge from video recipes to real-world scenarios [9, 73]. Additionally, the system should proactively monitor the cooking process to detect potential errors and offer corrective suggestions, recognizing that BLV users may not always be aware when a mistake occurs.

DG2 **Bridge the gap between real-world cooking states and video content.** The system should fuse the user's non-visual perception knowledge *("I can feel the dough is sticky.)"* with real-time visual analysis *("The dough is too thin compared to the recipe's demonstration.")* to align the user's actual cooking state with the intended recipe steps. In addition, the system should flexibly align real-world sensory input with video instructions, rather than enforcing rigid mappings. To be effective, it must also interpret and adapt to the user's context by handling ambiguities and offering disambiguation strategies when needed.

DG3 **Support flexible and accessible interaction with video recipes.** Referencing video recipes during cooking can impose a high cognitive load due to their length and multimodal nature [33, 41]. The system should accommodate diverse user needs by offering multiple forms of support, such as concise step-by-step guidance, detailed explanations, and the ability to jump directly to relevant video segments.

3.2 Example Usage Scenario

This section presents an example usage scenario in which Jane, a congenitally blind user, uses Aroma to prepare Spaghetti Bolognese. Each system feature mentioned corresponds to those described in Sec. 3.4, and is labeled as Ft.{ID}.

Preparation Jane starts by watching a video on the recipe and listens to it from beginning to end. While this gives her a general idea of the steps, she cannot remember everything, and many details are unclear. To get more help, she opens Aroma and loads the video. The system first analyzes the pre-recorded instructional video; it then uses Jane's mobile phone as a first-person camera to stream the kitchen environment in real time while capturing her voice commands for processing.

Step 1: Boiling Pasta Following the tutorial, Jane fills a pot with water and places it on the stove to boil. As Jane fills the pot and brings the water to a boil, the real-time monitoring agent aligns and compares her actions with the reference from the corresponding tutorial video. It detects that Jane has forgotten to add salt. In response, the system proactively alerts Jane with: *"It appears you have not added salt to your boiling water. Adding salt enhances the flavor of the pasta."* (Ft.2).

Step 2: Sautéing Onions and Garlic Jane recalls that her next step is to sauté chopped onions and garlic in olive oil. While preparing the ingredients, Jane feels uncertain about the knife technique, so she asks, *"How do I properly chop onions?"*. The system responds with: *"Onion should be chopped into small chunks"*, which was retrieved and synthesized from the video content (Ft.1). Jane finds the initial response too general, so she tries chopping the onion for a few samples, using her sense of touch to judge the thickness. She then follows up with the system, showing a sample and asking, *"I'm chopping my onion this thin, is this correct?"* After a brief pause, the system replies, *"According to the video recipe, the onions are chopped into thicker square slices. For making sauce, there's no need to chop them finely."* (Ft.1). Jane is satisfied with the answer and decides to continue cooking.

Step 3: Cooking the Sauce After chopping the onions, Jane forgets what her next step is, so she asks: *"What's my next step?"*. The system automatically takes the history of her previous actions into consideration, retrieves the content from the video, and responds to her with: *"Next, you should prepare the sauce. Heat oil in a pan, sauté the onions and garlic, and then incorporate tomato sauce and herbs."* (Ft.1). Jane feels uncertain about this response, so she says *"play the video recipe"* to ask the system to replay the relevant segments about *chopping onions* from the original video recipe. Jane can also control the clip by saying *"pause"* or related commands (Ft.3).

Step 4: Combining Pasta and Sauce In the final step, Jane needs to drain the pasta and mix it with the sauce before serving with grated Parmesan cheese. Here, she wants to reflect on the previous step, so she asks, *"Did I drain the pasta already or is it still in the water?"* After a short thinking process, the system replies based on the memory of the process (Ft.1).

3.3 System Overview

Aroma operates through a mixed-initiative architecture. Rather than relying solely on user prompts or system-driven guidance, Aroma continuously integrates: *non-visual perceptions* (see Fig. 2 ■): the BLV user's own sensory cues, such as smell, touch, and taste, conveyed verbally (e.g., describing the food texture or tasting for saltiness); *visual information* (Fig. 2 ■ V): real-time analysis of cooking states and ingredients from a wearable first-person camera (shown in Fig. 1); and *video recipe knowledge* (Fig. 2 ■): information extracted and structured from the given video recipe.

There are four primary components (see Fig. 2 ■) to coordinate reality information input in different modalities and process user interaction. Specifically, knowledge from the corresponding video recipe is stored in an accessible JSON form from a video information extraction pipeline. In parallel, a real-time visual analysis agent (implemented with a visual LLM, details in Sec. 3.5) streams the reality visual information into the system and monitors the cooking states by aligning the visual information with the video content. An interactive agent implemented based on another LLM handles the users' requests based on their perceived non-visual information, visual information from the visual agent, and the context. Video recipe knowledge, the periodically analyzed kitchen visual information, along with all user–agent conversational data, is stored as memory and retrieved by the context management module to provide adaptive context for each system response.

To coordinate user requests with evolving kitchen scenes, we model interaction as a deterministic state machine containing nine events and seven states drawn from prior studies [33, 34, 41]. Fig. 3 illustrates this framework. Each event is identified by the LLM based on the current visual scene and the user's voice request. For agent-initiated events (events 5 and 6), there was no need for user voice information. Every state links to a prompt template that guides the backend LLM. When the system enters a new state, it generates a response based on the current sensory input, video recipe knowledge, and the session context.

3.4 Key Features

Following the mixed-initiative interaction paradigm, Aroma implements the following features to fulfill the design goals outlined in Sec. 3.1.

3.4.1 Ft.1: Contextualized Responses Grounded in Visual and Non-Visual Perception. Since cooking involves multi-modal information (as shown in Fig. 1), where both visual and non-visual inputs are important to decide a cooking state, Aroma integrates visual data — captured through a wearable camera — and non-visual information verbalized by the user to represent the cooking state. This combined perceptual input is then grounded in the instructional content derived from the video recipe, as illustrated in Fig. 1.

Upon receiving a user query, the system immediately leverages the query that integrates visual and non-visual information to determine the current *event*, defined in the state machine framework (see Fig. 3), and the system transitions into one of the following four response categories:

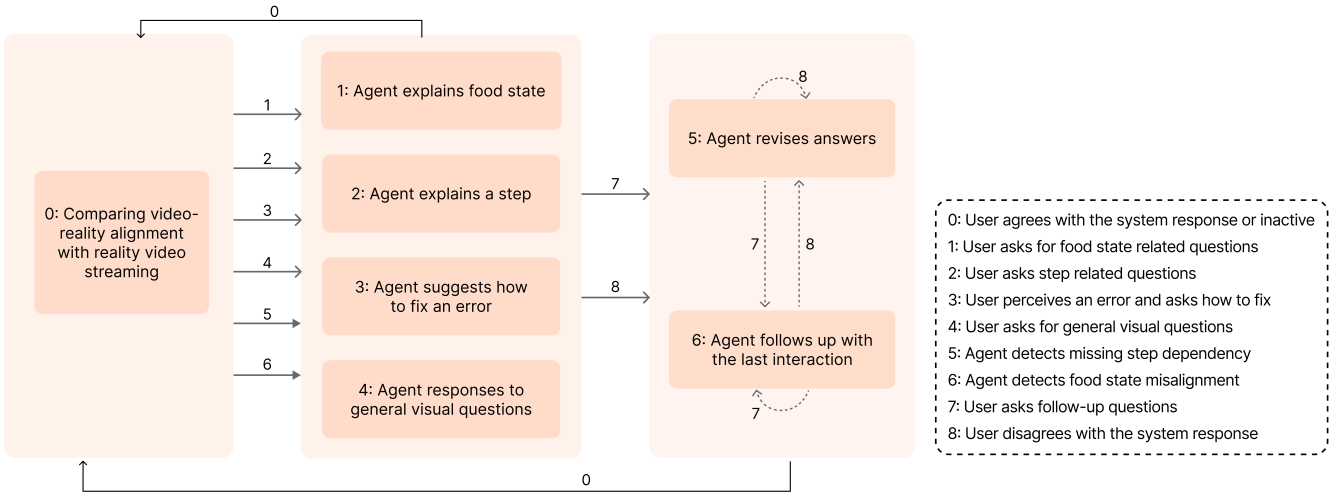

Figure 3: The state-machine-like framework in AROMA to handle various user queries and reality visual scene changes. The system starts at the initial state (state 0), and transitions between different states are determined by various events. Each event is decided by the current visual scene and the user's query (if available) by an LLM. When transitioning to a new state, an LLM generates a response using a predefined prompt tailored to that state. The response is subsequently transformed into audio output through a text-to-speech (TTS) service and delivered to the user.

- **Food State Responses** (state 1 in Fig. 3): inquiries about the current condition of the food. Triggered by event 1 from state 0.
- **Step-Related Responses** (state 2 in Fig. 3) — Clarifications regarding specific steps in the cooking process. Triggered by event 2 from state 0.
- **Problem-Solving Responses** (state 3 in Fig. 3) — Assistance when encountering issues or uncertainties. Triggered by events 3, 5, and 6 from state 0.
- **General Visual Questions Responses** (state 4 in Fig. 3) — Broader inquiries related to visual guidance on cooking procedure or recipe. Triggered by event 4 from state 0.

Responses are designed to deliver concise and critical information relevant to the user's immediate needs. For example, if a user inquires, *"Is the chicken cooked through?"* The system provides a targeted response as: *"The chicken is lightly browned externally but requires additional cooking, as the internal temperature has not reached 165°F."* — deliberately omitting extraneous details to minimize cognitive load.

If the query is either a follow up for more details (event 7 in Fig. 3) or the previous error wrong (event 8 in Fig. 3) — particularly because a user is unsatisfied with a default response due to vagueness, ambiguity, or limitations in model interpretation — the state will transit to the corresponding one (state 6 or state 5 in Fig. 3, respectively) where the backend LLM of the agent is prompted to pay special attention to the additional information provided by the user.

The system adapts dynamically to the user's non-visual perceptual input, such as information obtained through touch, sound, or smell, and incorporates that context into more grounded responses. This feature leverages the language model's capacity to synthesize multi-modal information and maintain context over multi-turn interactions.

3.4.2 Ft.2: Proactive Monitoring of the Cooking Process. Previous research [34] highlights the increased cognitive load BLV users face when accessing recipe instructions during cooking, which often prevents them from noticing errors during the cooking process. To address this challenge, AROMA adopts a mixed-initiative interaction model [22], blending user-initiated requests with system-initiated assistance to reduce friction and offload cognitive effort.

Specifically, AROMA leverages a video analysis agent that continuously monitors the cooking environment through a first-person camera (see Fig. 1). It performs two key operations every two seconds: (i) it generates objective observations of the current scene, and (ii) it compares those observations with reference knowledge extracted from the video recipe to make judgments.

Specifically, the agent is instructed to observe and describe:

- The specific cooking action being performed
- The corresponding recipe step
- The visible food items, ingredients, and kitchenware
- Any identifiable cooking-related sounds

Based on these observations, the agent determines:

- Whether the observed activity is relevant to the recipe
- If relevant, whether the step is being executed correctly
- Whether any required steps have been missed
- Whether the user has advanced to a new step

When the agent detects a deviation, such as a missed or incorrectly performed step, it alerts the user and provides corrective instructions. These situations map to event 5 and event 6 in Fig. 3 in the state machine, prompting a transition to a new interaction state (state 3 in Fig. 3), which in turn triggers the appropriate system response.

3.4.3 Ft.3: Accessing Video Segments and Memory Stored. To support DG3, AROMA enables users to access not only information from

the current scene but also segments from the original instructional video, and memory of previous user-agent interactions and visual information stored by the real-time visual analysis agent every given time interval (2 seconds in our setting).

To access video segments, users simply make a voice request e.g. *"Replay the part that tells me what ingredients I should prepare"*, and the backend LLM will interpret the semantic meaning of the request and play the segmented parts automatically to the user. Noticeably, the system automatically retrieves the relevant video segments from the video for the response it generates from Ft.1 (Sec. 3.4.1) or Ft.2 (Sec. 3.4.2). This allows the users to quickly refer back to the video recipe to find *evidence* if needed. To do this, users can issue commands like "play" or "pause" after receiving a response. This feature is designed to complement the conversational guidance and reduce cognitive effort. The design rationale for it is grounded on prior research, which highlights the effectiveness of AI-provided concrete examples in strengthening user trust [6]. In the context of cooking, it is also a common practice for users to navigate to the appropriate parts from the original recipe [34].

To further support DG1 and DG2 and to address the practical challenges of managing multi-step cooking tasks, the system also allows users to retrieve information from earlier stages of the session. Aroma maintains a comprehensive record of both the conversational history and the automatic visual analysis results every 2 seconds. Users can access this information by making a retrieval-related request, such as *"Did I already add the garlic?"*, and the system will retrieve the information from the context (memory) managing module (Fig. 2 ■) and play it to the user automatically, enabling users to reflect on or resume prior steps without guesswork.

3.5 Implementation Details

The system is implemented using Next.js[2]. We use OpenAI Realtime API[3] to continuously transcribe users' voice commands. The real-time visual analysis agent is achieved by periodically making the same request to the Multimodal Live API from Gemini[4]. In addition, we used GPT-4o-mini[5] to generate user-initiated responses (see Sec. 3.4.1) and extract relevant clips (Sec. 3.4.3) from the original video at the sentence level. We use OpenAI Text-to-Speech API[6] to convert text-based responses into audio.

To extract audio and visual information from video recipes, we use a set of AI models to obtain text-based descriptions from each modality separately. We first separate the vocal part from the video — typically containing spoken instructions — and store the transcribed text. We then parse the transcript at the sentence level and use PySceneDetect[7] package to extract key frames representing different visual scenes corresponding to each sentence's time interval. Next, we use GPT-4o to generate visual descriptions for each sentence by passing in all associated key frames. The model is prompted to describe the cooking steps, as well as the appearance, relative position, and relationships between ingredients and kitchenware. We use GAMA [17] to generate text-based descriptions for environmental sounds such as sizzling or simmering. The extracted video information is compiled into a structured JSON file.

Before the system starts running, a user can configure the speed of text-to-speech according to their preference. During runtime, it automatically returns to the initial state (state 0) either when the user indicates satisfaction or when the system remains idle for a set period (5 seconds in our implementation), as shown in the state machine illustration in Fig. 3.

4 User Study

To evaluate the usability and effectiveness of Aroma and to understand how BLV users use Aroma in cooking tasks, we conducted a study with eight BLV users[8]. The research questions are:

- RQ1: How effective is Aroma for assisting BLV users in cooking.
- RQ2: How do BLV users perceive the agency, control, and helpfulness when collaborating with Aroma in cooking?
- RQ3: What is the mental process of the user when coordinating with complementary perceptions from the system?

4.1 Participants

We recruited eight BLV participants for the study. Participants were screened using a demographic questionnaire to confirm that their visual acuity was worse than 20/200 and that they had no physical or medical conditions affecting mobility or the ability to handle kitchen tools. We did not put strict screening criteria on the prior cooking experience of potential participants. The average age of the participants was 37.1 (σ = 13.7). Five participants were female, and three were male. Detailed demographic information is shown in Table 1.

4.2 Study Setting

We conducted the user study in participants' own kitchens (P3, P5, and P8) or at alternative locations of their choice. We prepared three video recipes (details in Table 2), and the participant can freely choose one of them. Noticeably, we do not expect to draw any conclusions about the differences across video recipes, especially given the small sample size. These videos were preprocessed by Aroma using the knowledge extraction pipeline described in Sec. 3. All ingredients were prepared in advance by the experimenter and placed on a table. The participants were not aware of their exact positions; therefore, they used Aroma to explore and make confirmations. No heat-generating appliances or sharp knives were used during the sessions, per an Institutional Review Board (IRB) request to minimize the risk of the study protocol. In the three video recipes we selected, none of the steps require the use of sharp knives. However, some steps involve cutting tasks, such as dividing dough, that can be done using a safe knife. For any steps involving heat, participants were instructed to skip them. Each study session lasted about one hour, and participants were compensated with a 50 USD gift card for their time.

[2]https://nextjs.org/
[3]https://platform.openai.com/docs/guides/realtime
[4]https://ai.google.dev/gemini-api/docs/live
[5]https://platform.openai.com/docs/models/gpt-4o-mini
[6]https://platform.openai.com/docs/guides/text-to-speech
[7]https://www.scenedetect.com/

[8]This study was approved by the IRB at our institution.
[9]These videos can be accessed at https://www.youtube.com/watch?v=[VideoID]

ID	Age	Gender	Onset	Level of Visual Impairment	Occupation	Cooking Frequency
P1	34	M	Congenital	Blindness with some light/color perception	Professor	About once a week
P2	33	F	Congenital	Total Blindness	Self-employed	About once a week
P3	33	M	Congenital	Blindness with some light/color perception	Massage therapist	Almost every day
P4	25	F	Congenital	Total Blindness	Student	Less than once a week
P5	62	M	Acquired	Total Blindness	Chef	Almost every day
P6	28	F	Congenital	Total blindness	Student	About once a week
P7	24	F	Acquired	Blindness with some light/color perception	Student	Less than once a week
P8	58	F	Congenital	Total Blindness	Massage therapist	2–4 times a week

Table 1: Participant demographics for our user study

4.3 Study Process

After the consent process and a brief overview of the study procedure, the researcher informed the participant about the recipe categories listed in Table 2, and the participant selected one they liked from the list. The researcher then chose a different video recipe to use as an example for demonstrating the system's features. Specifically, each feature described in Sec. 3.4 was demonstrated. During the process, the researcher also addressed any questions the participant had.

To begin the main cycle of the study, participants listened to the full video recipe to develop a high-level understanding of the dish, relying primarily on auditory cues. They were instructed to retain as much detail as possible to support their performance during the cooking session.

Before the cooking process began, the researcher helped each participant wear a chest strap outfitted with a mobile phone that streamed a first-person video feed to the system. This chest-mounted setup was chosen based on prior findings showing that it offers greater comfort and more stable footage compared to head-mounted alternatives [34]. The participant then attempts to cook the selected dish with the help of Aroma.

After completing the cooking task, participants filled out a post-study questionnaire, rating Aroma's usability, usefulness, and features on a 7-point Likert scale. Then, we conducted a semi-structured interview to understand participants' experiences, perceptions, and expectations when using Aroma. The interviews lasted approximately 10–20 minutes and were audio-recorded for transcription and analysis. Specifically, we asked open-ended questions about participants' overall impressions of the system, perceived usefulness of individual features, and how the interaction compared to their prior cooking experiences. We also followed up on specific moments we observed during the session, such as when participants made follow-up requests or deviated from the video instructions, to elicit their underlying reasoning and challenges. Additionally, for features rated highly or poorly in the questionnaire, we probed further to understand the rationale behind their ratings.

We conducted a thematic analysis [5] of the interview transcripts, recordings, and observational notes from the researcher. One researcher reviewed all the data and open-coded segments [11] that reflected user behaviors, perceptions, and preferences. These initial codes were iteratively grouped into broader themes through axial coding, with regular discussion among the research team to refine interpretations and resolve discrepancies.

Additionally, we annotated the recording of each study session, labeling user requests, the events, and the system's responses. We then evaluated Aroma on two aspects: (i) the accuracy of mapping each request to the corresponding user-initiated event shown in Fig. 3. The accuracy is calculated by dividing the correctly mapped user queries by the total number of user queries, excluding user follow-ups (events 7 and 8). To do this, two researchers manually annotated the category of events for each request as the ground truth. Non-agreement cases were discussed to reconcile differences. (ii) the initial factual accuracy of the resulting response, without any user follow-ups (states 1-4 as shown in Fig. 3). We compute it as the proportion of answers labeled correct out of all user queries. Similarly, the factual boolean labels are manually labeled by the researchers. Noticeably, we do not focus on the response quality in this analysis—this is reflected separately in the qualitative results from the user study. A response is considered correct if it meets both requirements: a.) it directly addresses the user's query; b.) the response contains no factual errors.

Index	Video ID[9]	Recipe	Length
1	lgg6luYfQ1w	Pepperoni Pizza	6:18
2	mixdagZ-fwI	Beef Taco	5:45
3	lH7pgsnyGrI	Miso Soup	3:41

Table 2: Video recipes for user study.

5 Results and Findings

In this section, we begin by presenting quantitative results on the system's accuracy in mapping user requests to the correct events, as well as the accuracy of its immediate responses, which are generated in states 1–4, as shown in Fig. 3. Then we report the overall usability ratings from participants, user feedback on each feature, and the reasons behind their perceptions. We then closely examine the specific requests made by BLV users. Building on this, we analyze

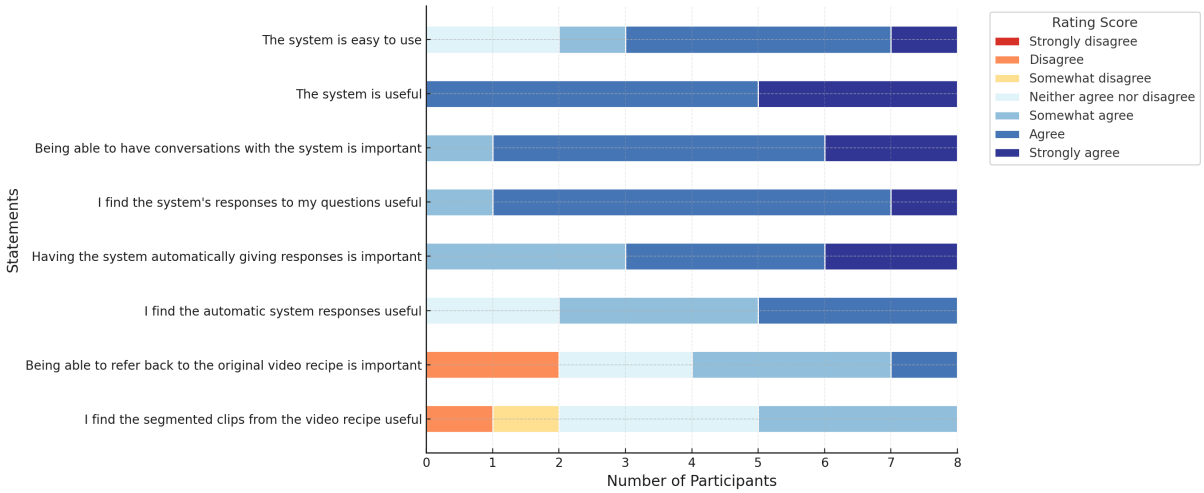

Figure 4: Participants' ratings of the usability and usefulness statements of Aroma

their mental models when utilizing the additional perception and cognitive capabilities of Aroma.

5.1 Quantitative Results on System Performance

Participants frequently issued requests to Aroma. Table 3 reports the total request count, the system's accuracy in mapping each request to the correct event in Fig. 3, and the factual accuracy of its immediate responses. Across all participants, event-mapping accuracy averaged 0.82, while factual accuracy averaged 0.67.

Most mapping errors stemmed from the diverse ways participants phrased the same type of request. For example, when users sought clarification on a specific step (event 2), one might explicitly ask, *"Could you explain this step for me?"* while another might phrase it more colloquially, e.g., *"What's going on now?"*, which occasionally caused the model to misclassify the query as event 4, in which the backend LLM is prompted to give a generic response rather than step-specific guidance. Although such replies could be factually accurate, they did not satisfy users' informational needs and undermined their experience. One of the participants reflected on this as: *"When I ask about a step, I expect details like how long it takes and how to perform it, but I remember once the system just simply told me what this step was."* As Aroma relies on prompts for event mapping, we anticipate that incorporating user-specific few-shot examples in the future will reduce these errors and enhance user experience.

Response errors most often arise from hallucinations in the underlying language model. As explained in Sec. 3, Aroma answers each user-initiated query by conditioning on the runtime session history. As this context window lengthens, the accuracy degrades accordingly, which is a critical limitation of most LLM-based systems [49, 54]. To mitigate this problem, especially for future deployments that support BLV users tackling more complex recipes, the underlying algorithm of the context (memory) managing module (Fig. 2 ■) should be enhanced to reduce the redundancy of the memory, and enable more efficient searching algorithm for retrieving the appropriate context, rather than purely relying on the LLM capability.

5.2 System Usability and User Perception

Overall, participants reported high satisfaction with Aroma's usability and its effectiveness in supporting cooking tasks. Post-study questionnaire results indicate that conversational features, including receiving immediate responses to questions and asking follow-ups (Sec. 3.4.1), were particularly well-received. These features were rated as both highly important ($\mu = 6.13$, $\sigma = 0.60$) and useful ($\mu = 6.00$, $\sigma = 0.50$). System-initiated assistance (Sec. 3.4.2) also received strong importance ($\mu = 5.88$, $\sigma = 0.78$) and ($\mu = 5.13$, $\sigma = 0.78$) useful ratings, suggesting that participants appreciated the system's ability to monitor progress and offer timely support without prompting. Lastly, the ability to replay segments of the recipe video (Sec. 3.4.3) was not seen as important as the previous ones ($\mu = 4.13$, $\sigma = 1.36$), reflecting its role as a supplementary aid rather than a critical component of interaction. Fig. 4 summarizes the distribution of participant ratings across these features.

5.2.1 Conversational interaction supports flexible access to video recipes, enhances agency, and reduces cognitive load for the user. Participants emphasized that conversational interaction allowed them to engage with content more flexibly and on their own terms, rather than being constrained by the linear structure of a video. For

Part.	Total Queries	Correct Mappings	Mapping Accuracy	Correct Responses	Response Accuracy
P1	10	9	0.90	8	0.80
P2	6	5	0.83	5	0.83
P3	13	11	0.85	9	0.69
P4	7	5	0.71	4	0.57
P5	16	12	0.75	10	0.63
P6	11	9	0.82	8	0.73
P7	8	7	0.88	5	0.63
P8	12	10	0.83	7	0.58

Table 3: System performance across participants

instance, P3 appreciated being able to re-ask questions and request clarifications: *"I could ask what ingredients I need for the next step without hearing back from the video again."* This ability to access information non-linearly helped reduce stress, especially when compared to traditional video consumption. As P1 noted: *"With a video, once you miss something, it's hard to go back and catch it."*

Going beyond, when system outputs were ambiguous or incomplete, users often followed up with contextual cues based on their own non-visual perception to refine or confirm the answer. For example, P5 first asked the agent for an overview of the kitchen setup, then used that as a scaffold to ask specifically about the location of the pizza dough: *"This helps me build a hierarchical understanding of what's on my counter."* Similarly, P4 recounted using touch to distinguish between two plates when the system ambiguously referred to "a round plate": *"I could feel there were two plates. So I picked one (from the texture) and asked, 'Is this the pepperoni?' and got confirmation. That helps me be sure."*

5.2.2 Proactive feedback was helpful when precise and well-timed. Several participants found that proactive feedback helped them stay on track and feel more confident moving forward. For example, P6 recalled, *"The system automatically detected a wrong step where I was about to put the miso paste before adding the tofu into the soup."* Similarly, P2 and P3 appreciated the reminders to roll the pizza crust edge to cover the string cheese—something they had missed, but the system intervened with a useful reminder.. These interventions reduced their cognitive load and stress, as P3 noted: *"I know I have an assistant by my side to detect some, if not all, problems."*

The challenge of this feature, however, was the timing and relevance of the interventions. Participants were frustrated when the system intervened during intentional deviations. For instance, P5 skipped a cutting step due to difficulty, yet the system still prompted her to complete it. Additionally, due to inevitable time delays, responses sometimes failed to align precisely with user actions. Despite these issues, precise proactive messages were generally well-received, as they could be easily ignored. As P2 described, the system felt like a "background assistant" rather than a distraction.

5.2.3 Segmented video replays were useful when original video clips were accessible. In general, participants found the feature useful when the original audio information—instructions from the speaker, for instance—was accessible and informative. In these cases, replays helped users quickly *"verify step sequences"* (P6) or confirm specific actions without explicitly asking the system to answer it, which takes extra time (P2). However, the usefulness of this feature is limited for inaccessible video clips. For example, P3 noticed that: *"I tried this feature several times, but then I realized I couldn't get enough information I wanted, compared with other features."*

5.3 What Requests are BLV Users Making?

To better understand how BLV participants interacted with Aroma over cooking, we analyzed the types of requests they made throughout the sessions. Drawing from both observation and interview data, we identified key categories of requests, their underlying motivations, and implications for future system design.

5.3.1 Maintain awareness of procedural flow. In our study, participants frequently relied on the system for procedural guidance. Nearly all participants asked questions such as *"What should I do next?"* (P1) and *"I have finished preparing the meat, what's the next step?"* (P6). These requests were especially common when participants had already made several other requests in the previous step, or when they wanted to skip a specific step—for example, a participant asked *"What should I do after baking the pizza?"* (P2). Addressing this, future systems could consider automatically responding when detecting the completion of a step. However, determining the "end of a procedure" is inherently ambiguous and may place greater demands on the model's performance and accuracy.

5.3.2 Confirm the types and locations of foods and ingredients. In addition to procedural support, participants often used the system to confirm the identity or location of ingredients, especially when non-visual sensory cues alone were insufficient. Specifically, participants often turned to the AI system when multiple items had similar textures, smells, or spatial positions. For example, P4 asked questions like *"Is this the mozzarella?"* when she felt a moist texture, but was not sure whether it was mozzarella or pepperoni. In other cases, users directly sought help from the visual agent to locate items. For instance, P5 asked, *"Do you see anything that looks like pizza dough?"* These interactions illustrate a collaborative dynamic between user and system: participants brought non-visual sensory impressions to the conversation, while the AI contributed visual grounding. Together, they formed a hybrid perception loop, helping users confidently identify and locate items in their environment.

5.3.3 Clarification and response customization through follow-up requests. Participants frequently followed up on initial agent responses to clarify ambiguous instructions or to tailor answers to their individual preferences. For instance, P3 asked the system to *"order it from top left to bottom right"* after asking *"tell me what you see..."* as first, explicitly requesting a spatially structured response aligned to his habits. Similarly, P1 asked to *"Describe what you see in a clockwise direction."* and P4 asked to *"Describe the items to my left one by one."*

In other cases, users rephrased or elaborated their questions when the initial answer lacked sufficient detail or clarity. For example, P5 attempted to form a stuffed pizza crust and asked repeatedly about the 90-degree folding step, iterating: *"Is this one the video said to rotate 90 degrees? Or am I doing it wrong?"*. These interactions emphasize the importance of supporting iterative, dialogic clarification—allowing users to gradually refine their understanding and guide the system toward more useful, personalized responses.

5.4 How Non-Visual and Visual Perception and Cognition Complement Each Other?

Our study revealed how BLV users integrate visual assistance from the AI system into their existing non-visual cooking strategies. Rather than replacing their tactile, auditory, olfactory, and gustatory skills, Aroma acted as a complementary scaffold that enhanced these well-practiced abilities. We observed several key patterns in how users combined visual and non-visual information during the cooking process:

5.4.1 Non-visual perception as the foundation for assessing food state. For participants who had experience in cooking, they began with a robust, non-visual schema for gauging food readiness and

identifying ingredients. These strategies included interpreting texture, shape, container types, and spatial orientation through touch. Although participants who had less cooking experience (P4, P8) also expressed that the additional visual information is *"additional help"* (P8). Specifically, P5, who is an experienced blind chief, described how the use of various bowl shapes helps them identify ingredients without needing to touch the ingredients directly: *"I use different shapes and sizes and textures. So if the computer couldn't answer me, I could."* Similarly, P3 explained how he could *"feel the edge of the pan so I know how far I need to stretch the dough."* In another instance, a participant confirmed a sauce's identity by taste rather than asking the agent.

These examples illustrate that many users used the AI primarily to validate their own perceptions. As P5 reflected: *"You do know by touch that the crust is uniform without asking the computer"*.

5.4.2 Resolving ambiguities in object and step identification. Participants frequently used Aroma to verify what had inferred through touch or sound, especially when encountering ambiguous objects. As P5 explained while handling pizza dough: *"I think this is probably triangular... I can tell by touch, but I'm just confirming."*. This interplay of tactile inference and visual confirmation was especially important when participants encountered objects with similar textures or forms. For instance, P2 described struggling to distinguish between string cheese and mozzarella by touch or taste alone, noting that *"they all have similar textures and flavors, and the video doesn't say what they look like either."* In another moment, when the participant (P1) did not understand how to shape the crust with chopped cheese, they explicitly asked: *"Did the video say anything about rotating 90 degrees?"* and later double-checked by asking: *"Is this what the video said about folding it?"*, showing their reliance on the visual properties of the food at this moment.

5.4.3 Improving spatial awareness of ingredients and tools. Participants used the AI's egocentric vision not only to confirm objects' identities but also to orient themselves within the physical workspace. For example, during the preparation of each dish, nearly every participant asked the system to help locate ingredients and cookware, and confirm if the ingredient is in the correct position. For instance, P2 asked questions such as: *"Am I spreading the pizza sauce on the center of the dough?"*. We also observed that participants combined their demonstrations with the system's visual feedback to ask spatial questions more directly, such as: *"If cheese sticks were evenly placed here?"* (P3)—while simultaneously pointing to the pizza crust with their finger.

5.4.4 User expectations: extending visual perception beyond food and ingredients to include gestures from users. In Aroma , the real-time visual analysis agent was originally designed to visually track and interpret the ingredients and tools in the process. However, from the creative uses of the participants during the study, we observed opportunities to improve their experience. Specifically, P6 once requested to use schemas like "clockwise direction" for the system to describe objects. Meanwhile, as noticed by P3, he prefers responses based on hand-relative coordinates (e.g., "left of your right hand"), while others like P4 don't show a particular preference for this.

These observations suggest two promising directions for future versions of the system. First, responses could be personalized based on users' preferred spatial language or cognitive mapping strategies. Second, incorporating gesture recognition could provide additional context, improving the accuracy and relevance of visual feedback in real-time interactions.

6 Discussion

Our study of Aroma highlights that designing assistive technologies for BLV users requires more than providing additional sensory information — it demands systems that collaborate with users' embodied skills, adaptive strategies, and personal expertise. We discuss how future systems can move toward human-AI co-reasoning, support multimodal and sensor-integrated interactions, and balance proactive assistance with user privacy and agency. Finally, we reflect on how Aroma's design principles may generalize to other activities of daily living (ADLs), informing more flexible, inclusive, and context-aware assistive technologies.

6.1 From Providing Additional Sensory Information to Co-Reasoning

Our study shows that BLV users engage in cooking not just by the additional visual information, but through a rich, embodied non-visual cognitive ability developed over time, which is grounded in touch, smell, sound perception, and their spatial memory etc. Therefore, we suggest that assistive technology should go beyond simply asking, *"how to compensate what is missing?"*, future systems should consider *"how to work with the user's existing and ongoing cognitive process?"* Just as P5 explained: *"I use different shapes and sizes and textures. So if the computer couldn't answer me, I could."*

In Aroma, we leverage the text-based reasoning capability of LLMs to contextualize its response based on non-visual clues given by the user through prompting. However, it is inevitable that the responses will occasionally be incorrect, ambiguous, or not actionable (for instance, providing instructions that have visual descriptions) — depending on the performance of the model itself. In Aroma we try to reduce such effects by carefully designing the prompts that we use, looking ahead, future systems could build on this foundation by incorporating specialized models that offer greater robustness and reliability in interpreting and responding in non-visual contexts.

6.2 Multimodal User Interaction Beyond Verbalization

Aroma's current interaction paradigm primarily leverages spoken commands and perceptual descriptions provided by the user. However, in the user study, we also noticed that participants naturally employed a variety of non-verbal cues for expression throughout their cooking processes. For instance, we observed frequent gestures such as pointing, tapping, or sweeping hand movements to reference objects or to indicate spatial queries. This observation is consistent with prior research that emphasizes the potential of hand- or wrist-based gestures for supporting non-visual interaction, such as navigating to the next step in a recipe [33]. A potential improvement could be integrating a robust gesture-recognition component to interpret these signals [33], which would augment

the visual understanding capability of the system beyond a simple end-to-end LLM call and further reduce the cognitive load of verbalizing for BLV users.

Going beyond gestures that have explicit semantic meaning, tracking user expressions and micro-gestures could also serve as effective cognitive indicators [20]. Specifically, indicators such as hesitation pauses, head movement trajectory could all serve as indicators of cognitive load, confusion, or uncertainty. Integrating those non-verbal cues in the memory management component of Aroma (see Fig. 2 ■) and adapting the context dispatch strategy accordingly, presents an opportunity for Aroma to deliver results that are better contextualized to the user's cognitive state.

6.3 Toward Multimodal, Sensor-Integrated Interaction Paradigms

In the current design of Aroma, users verbalize their sensory experiences, offering cues beyond visual information to help the system determine the cooking state and provide appropriate guidance for next steps. However, several participants noted that this process can be cognitively demanding.

Sensor augmentation offers a path forward: thermal sensors, smart utensils, visual object/action recognition [37], or even audio classifiers could reduce this burden by allowing the system to proactively support users based on real-time sensor signals. These approaches could support a more seamless mixed-initiative flow, where AI agents mediate the timing and content of intervention based on contextual uncertainty and user need.

Yet, our study reaffirms that embodied, subjective knowledge, such as taste or smell, remains essential to cooking. Even in sensor-rich environments, systems should preserve and support this form of expertise. This also aligns with Dourish's view of embodied interaction [13], where technology is most effective when it is grounded in the body and lived practice. Aroma demonstrates that subjective perception is not only valid input, but often the most reliable and personally meaningful cue in practice.

6.4 Designing for Procedural Fluidity vs. Procedural Fidelity

While Aroma was designed to support users in following the procedural flow of video recipes, our findings reveal that BLV users often engage in cooking practices that intentionally diverge from strict procedural fidelity. Participants might skip, reorder, or modify recipe steps to accommodate personal preferences, available ingredients, or contextual constraints. These observations highlight an important future direction for assistive systems: designing for procedural fluidity rather than enforcing rigid adherence to predefined steps. Future systems could model users' intent to distinguish between errors and deliberate adaptations, enabling more collaborative interaction [33]. For instance, when users skip or combine steps for efficiency or convenience, the system could proactively adjust its guidance to reflect the new procedural context. This design approach respects the autonomy and expertise of BLV users while maintaining safety and alignment with instructional content.

6.5 Balancing Proactivity with Privacy and Agency in Sensor-Rich Environments

Aroma's mixed-initiative design, particularly its proactive monitoring features, was generally well-received; however, participants' experiences also underscore the importance of balancing system proactivity with user agency and privacy considerations. While timely interventions helped prevent errors and supported task progression, participants occasionally expressed frustration when the system intervened during intentional deviations or when proactive prompts conflicted with user preferences. This highlights an emerging design challenge for future assistive technologies operating in personal or sensor-rich environments [2]. Systems should enable configurable levels of proactivity, allowing users to specify preferences for monitoring sensitivity, intervention types, or contexts in which proactive assistance is appropriate. Designing consent-driven and context-aware proactivity mechanisms would preserve user control while offering tailored support aligned with individual privacy expectations and situational needs.

6.6 Extending Aroma's Framework to Other Activities of Daily Living

While our work focuses on cooking, the design principles underlying Aroma—grounding real-world sensory perception with instructional video content through mixed-initiative interaction—hold promise for supporting other activities of daily living (ADLs). Tasks such as makeup application [40], home repair [31], arts and crafts [42], or gardening [28] similarly involve multimodal perception, procedural knowledge, and embodied expertise. Each domain presents unique sensory demands and interaction challenges, necessitating domain-specific adaptations. Nevertheless, the core approach of leveraging user-provided non-visual cues, integrating real-time visual analysis, and aligning with video-based instruction could be generalized across ADLs. Future research should explore the transferability of Aroma's framework, contributing to the development of universal assistive technologies that support BLV users across diverse everyday activities.

7 Limitations

While our system demonstrates the potential of AI-enhanced, non-visual cooking support through multimodal grounding with video recipes, the current implementation reveals several limitations that suggest important directions for future work.

Firstly, our user study involved eight BLV participants, which is a limited sample size. In addition, heat and sharp knives are not used due to safety concerns. Therefore, the study may not fully represent the range and steps of cooking activities that BLV users routinely perform. In contrast, Aroma is a proof of concept that shows how a human–AI partnership can combine visual and non-visual cues to perceive the kitchen state and translate video-based knowledge into actions in reality.

Another limitation is the unstable latency and response timing of the backend multi-modal LLMs that affect the user experience. Despite leveraging the Gemini multimodal API and OpenAI's real-time audio transcription to minimize delay, response latency remains a significant limitation in both system-initiated and user-initiated interactions. For system-initiated prompts, such as real-time visual

analysis of the user's current cooking state, visual processing occasionally lagged behind action. This led to situations where *"The agent confirmed a step I already did 5 seconds ago."* as P7 noted. This delay may also compromise safety in time-sensitive steps. User-initiated queries also experienced varying response times depending on request complexity and length. For example, when a user asked, *"Can you tell me what you see right now? And order it from top-left to bottom-right?"* (P4), the response required full-scene analysis, leading to a multi-second pause that broke task flow. Participants expressed frustration with these lags in the interview: *"Sometimes the response is just not perfect, or I should say it's too slow"* (P3). Improving model optimization and pipelining inference tasks in parallel is crucial to achieving more fluid and dependable interaction.

In addition, the current version of Aroma requires preprocessing the video offline, which may limit recipe access for BLV people. The instructional video must first be parsed to extract stepwise procedural knowledge, ingredient states, and visual targets — a process we have not implemented on-the-fly. In the interview, we were frequently asked questions like *"What if I want to try something from another YouTube channel? Can it still help me?"* This motivates future work on scalable pipelines that can process arbitrary video recipes near real-time to support "plug-and-play" recipe access.

8 Conclusion

In this work, we present Aroma, a mixed-initiative, multimodal AI system that supports BLV users in cooking by grounding their real-time non-visual sensory input and first-person video stream from a wearable camera with the audiovisual information from a video recipe. Through a study with eight BLV participants, we demonstrate the effectiveness of Aroma and highlight the insights for designing AI systems that not only provide additional visual information but also recognize and respond to the unique perceptual states of BLV users in performing real-world tasks.

Acknowledgments

This work was supported in part by a Notre Dame-IBM Technology Ethics Lab Award, an NVIDIA Academic Hardware Grant, a Google Research Scholar Award, a Gift from Adobe Inc., the National Eye Institute of the National Institutes of Health R01EY037100, and NSF CMMI-2326378. Any opinions, findings, or recommendations expressed here are those of the authors and do not necessarily reflect the views of the sponsors.

References

[1] Nayyer Aafaq, Ajmal Mian, Wei Liu, Syed Zulqarnain Gilani, and Mubarak Shah. 2019. Video description: A survey of methods, datasets, and evaluation metrics. *ACM Computing Surveys (CSUR)* 52, 6 (2019), 1–37.

[2] Tousif Ahmed, Roberto Hoyle, Kay Connelly, David Crandall, and Apu Kapadia. 2015. Privacy concerns and behaviors of people with visual impairments. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 3523–3532.

[3] Oliver Alonzo, Hijung Valentina Shin, and Dingzeyu Li. 2022. Beyond Subtitles: Captioning and Visualizing Non-speech Sounds to Improve Accessibility of User-Generated Videos. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '22)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3517428.3544808

[4] Ava Bartolome and Shuo Niu. 2023. A Literature Review of Video-Sharing Platform Research in HCI. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–20. doi:10.1145/3544548.3581107

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

[6] Carrie J. Cai, Jonas Jongejan, and Jess Holbrook. 2019. The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19)*. Association for Computing Machinery, New York, NY, USA, 258–262. doi:10.1145/3301275.3302289

[7] Xinyun Cao and Dhruv Jain. 2024. Supporting Sound Accessibility by Exploring Sound Augmentations in Virtual Reality. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '24)*. Association for Computing Machinery, New York, NY, USA, 1–5. doi:10.1145/3663548.3688525

[8] Santiago Castro, Mahmoud Azab, Jonathan Stroud, Cristina Noujaim, Ruoyao Wang, Jia Deng, and Rada Mihalcea. 2020. LifeQA: A Real-life Dataset for Video Question Answering. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association, Marseille, France, 4352–4358. https://aclanthology.org/2020.lrec-1.536/

[9] Paul Chandler, , and John Sweller. 1991. Cognitive Load Theory and the Format of Instruction. *Cognition and Instruction* 8, 4 (Dec. 1991), 293–332. doi:10.1207/s1532690xci0804_2 Publisher: Routledge _eprint: https://doi.org/10.1207/s1532690xci0804_2.

[10] Minsuk Chang, Leonore V. Guillain, Hyeungshik Jung, Vivian M. Hare, Juho Kim, and Maneesh Agrawala. 2018. RecipeScape: An Interactive Tool for Analyzing Cooking Instructions at Scale. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3173574.3174025

[11] Juliet M. Corbin and Anselm Strauss. 1990. Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative Sociology* 13, 1 (March 1990), 3–21. doi:10.1007/BF00988593

[12] Yinpei Dai, Run Peng, Sikai Li, and Joyce Chai. 2024. Think, Act, and Ask: Open-World Interactive Personalized Robot Navigation. http://arxiv.org/abs/2310.07968 arXiv:2310.07968.

[13] Paul Dourish. 2001. *Where the Action Is: The Foundations of Embodied Interaction*. The MIT Press. doi:10.7551/mitpress/7221.001.0001

[14] Junchi Feng, Giles Hamilton-Fletcher, Todd E. Hudson, Mahya Beheshti, Maurizio Porfiri, and John-Ross Rizzo. 2025. Haptics-based, higher-order sensory substitution designed for object negotiation in blindness and low vision: Virtual Whiskers. *Disability and Rehabilitation. Assistive Technology* (Feb. 2025), 1–20. doi:10.1080/17483107.2025.2458112

[15] Louise Fryer. 2016. *An Introduction to Audio Description: A practical guide*. Routledge, London. doi:10.4324/9781315707228

[16] Lishuai Gao, Yujie Zhong, Yingsen Zeng, Haoxian Tan, Dengjie Li, and Zheng Zhao. 2024. LinVT: Empower Your Image-level Large Language Model to Understand Videos. doi:10.48550/arXiv.2412.05185 arXiv:2412.05185 [cs] version: 2.

[17] Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. 2024. GAMA: A Large Audio-Language Model with Advanced Audio Understanding and Complex Reasoning Abilities. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 6288–6313. https://aclanthology.org/2024.emnlp-main.361

[18] Google Cloud. 2024. Vertex AI Video Understanding - Generative AI on Google Cloud. https://cloud.google.com/vertex-ai/generative-ai/docs/multimodal/video-understanding Accessed: 2025-04-08.

[19] Anhong Guo, Junhan Kong, Michael Rivera, Frank F. Xu, and Jeffrey P. Bigham. 2019. StateLens: A Reverse Engineering Solution for Making Existing Dynamic Touchscreens Accessible. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology (UIST '19)*. Association for Computing Machinery, New York, NY, USA, 371–385. doi:10.1145/3332165.3347873

[20] Ernest A. Haggard and Kenneth S. Isaacs. 1966. Micromomentary facial expressions as indicators of ego mechanisms in psychotherapy. In *Methods of Research in Psychotherapy*, Louis A. Gottschalk and Arthur H. Auerbach (Eds.). Springer US, Boston, MA, 154–165. doi:10.1007/978-1-4684-6045-2_14

[21] Keita Higuch, Ryo Yonetani, and Yoichi Sato. 2016. Can Eye Help You? Effects of Visualizing Eye Fixations on Remote Collaboration Scenarios for Physical Tasks. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. Association for Computing Machinery, New York, NY, USA, 5180–5190. doi:10.1145/2858036.2858438

[22] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems the CHI is the limit - CHI '99*. ACM Press, Pittsburgh, Pennsylvania, United States, 159–166. doi:10.1145/302979.303030

[23] Felix Huppert, Gerold Hoelzl, and Matthias Kranz. 2021. GuideCopter - A Precise Drone-Based Haptic Guidance Interface for Blind or Visually Impaired People. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–14. doi:10.1145/3411764.3445676

[24] Gaurav Jain, Basel Hindi, Connor Courtien, Xin Yi Therese Xu, Conrad Wyrick, Michael Malcolm, and Brian A. Smith. 2023. Front Row: Automatically Generating Immersive Audio Representations of Tennis Broadcasts for Blind Viewers. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*. Association for Computing Machinery, New York, NY, USA, 1–17. doi:10.1145/3586183.3606830

[25] Lucy Jiang, Crescentia Jung, Mahika Phutane, Abigale Stangl, and Shiri Azenkot. 2024. "It's Kind of Context Dependent": Understanding Blind and Low Vision People's Video Accessibility Preferences Across Viewing Scenarios. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–20. doi:10.1145/3613904.3642238

[26] Mohammad Kianpisheh, Franklin Mingzhe Li, and Khai N Truong. 2019. Face recognition assistant for people with visual impairments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 1–24.

[27] Benjamin Lafreniere, Tovi Grossman, and George Fitzmaurice. 2013. Community enhanced tutorials: improving tutorials with multiple demonstrations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. Association for Computing Machinery, New York, NY, USA, 1779–1788. doi:10.1145/2470654.2466235

[28] Loretto Lambe. 1995. Gardening: A multisensory experience. In *Making leisure provision for people with profound learning and multiple disabilities*. Springer, 113–130.

[29] Jaewook Lee, Andrew D. Tjahjadi, Jiho Kim, Junpu Yu, Minji Park, Jiawen Zhang, Jon E. Froehlich, Yapeng Tian, and Yuhang Zhao. 2024. CookAR: Affordance Augmentations in Wearable AR to Support Kitchen Tool Interactions for People with Low Vision. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology (UIST '24)*. Association for Computing Machinery, New York, NY, USA, 1–16. doi:10.1145/3654777.3676449

[30] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L. Berg, Mohit Bansal, and Jingjing Liu. 2021. Less Is More: ClipBERT for Video-and-Language Learning via Sparse Sampling. 7331–7341. https://openaccess.thecvf.com/content/CVPR2021/html/Lei_Less_Is_More_ClipBERT_for_Video-and-Language_Learning_via_Sparse_Sampling_CVPR_2021_paper.html

[31] Barbara Leporini, Michele Rosellini, and Nicola Forgione. 2020. Designing assistive technology for getting more independence for blind people when performing everyday tasks: an auditory-based tool as a case study. *Journal of Ambient Intelligence and Humanized Computing* 11 (2020), 6107–6123.

[32] Franklin Mingzhe Li, Di Laura Chen, Mingming Fan, and Khai N Truong. 2021. "i choose assistive devices that save my face" a study on perceptions of accessibility and assistive technology use conducted in china. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.

[33] Franklin Mingzhe Li, Jamie Dorst, Peter Cederberg, and Patrick Carrington. 2021. Non-Visual Cooking: Exploring Practices and Challenges of Meal Preparation by People with Visual Impairments. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '21)*. Association for Computing Machinery, New York, NY, USA, 1–11. doi:10.1145/3441852.3471215

[34] Franklin Mingzhe Li, Michael Xieyang Liu, Shaun K. Kane, and Patrick Carrington. 2024. A Contextual Inquiry of People with Vision Impairments in Cooking. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–14. doi:10.1145/3613904.3642233

[35] Franklin Mingzhe Li, Michael Xieyang Liu, Yang Zhang, and Patrick Carrington. 2022. Freedom to Choose: Understanding Input Modality Preferences of People with Upper-body Motor Impairments for Activities of Daily Living. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility*. ACM, Athens Greece, 1–16. doi:10.1145/3517428.3544814

[36] Franklin Mingzhe Li, Kaitlyn Ng, Bin Zhu, and Patrick Carrington. 2025. Exploring Object Status Recognition for Recipe Progress Tracking in Non-Visual Cooking. *arXiv preprint arXiv:2507.03330* (2025).

[37] Franklin Mingzhe Li, Kaitlyn Ng, Bin Zhu, and Patrick Carrington. 2025. OSCAR: Object Status and Contextual Awareness for Recipes to Support Non-Visual Cooking. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, 1–6. doi:10.1145/3706599.3720172

[38] Franklin Mingzhe Li, Akihiko Oharazawa, Chloe Qingyu Zhu, Misty Fan, Daisuke Sato, Chieko Asakawa, and Patrick Carrington. 2025. More than One Step at a Time: Designing Procedural Feedback for Non-visual Makeup Routines. *arXiv preprint arXiv:2507.03942* (2025).

[39] Franklin Mingzhe Li, Franchesca Spektor, Meng Xia, Mina Huh, Peter Cederberg, Yuqi Gong, Kristen Shinohara, and Patrick Carrington. 2022. "It Feels Like Taking a Gamble": Exploring Perceptions, Practices, and Challenges of Using Makeup and Cosmetics for People with Visual Impairments. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–15. doi:10.1145/3491102.3517490

[40] Franklin Mingzhe Li, Franchesca Spektor, Meng Xia, Mina Huh, Peter Cederberg, Yuqi Gong, Kristen Shinohara, and Patrick Carrington. 2022. "It feels like taking a gamble": Exploring perceptions, practices, and challenges of using makeup and cosmetics for people with visual impairments. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–15.

[41] Franklin Mingzhe Li, Ashley Wang, Patrick Carrington, and Shaun K. Kane. 2024. A Recipe for Success? Exploring Strategies for Improving Non-Visual Access to Cooking Instructions. In *The 26th International ACM SIGACCESS Conference on Computers and Accessibility*. ACM, St. John's NL Canada, 1–15. doi:10.1145/3663548.3675662

[42] Franklin Mingzhe Li, Lotus Zhang, Maryam Bandukda, Abigale Stangl, Kristen Shinohara, Leah Findlater, and Patrick Carrington. 2023. Understanding visual arts experiences of blind people. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–21.

[43] Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. 2022. Learning to Answer Questions in Dynamic Audio-Visual Scenarios. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 19086–19096. doi:10.1109/CVPR52688.2022.01852 ISSN: 2575-7075.

[44] Mingzhe Li, Mingming Fan, and Khai N Truong. 2017. BrailleSketch: A gesture-based text input method for people with visual impairments. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*. 12–21.

[45] Yunzhi Li, Franklin Mingzhe Li, and Patrick Carrington. 2023. Breaking the "Inescapable" Cycle of Pain: Supporting Wheelchair Users' Upper Extremity Health Awareness and Management with Tracking Technologies. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–17. doi:10.1145/3544548.3580660

[46] Ziming Li, Shannon Connell, Wendy Dannels, and Roshan Peiris. 2022. SoundVizVR: Sound Indicators for Accessible Sounds in Virtual Reality for Deaf or Hard-of-Hearing Users. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '22)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3517428.3544817

[47] Georgianna Lin, Jin Yi Li, Afsaneh Fazly, Vladimir Pavlovic, and Khai Truong. 2023. Identifying Multimodal Context Awareness Requirements for Supporting User Interaction with Procedural Videos. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–17. doi:10.1145/3544548.3581006

[48] Kate Lister, Tim Coughlan, Francisco Iniesto, Nick Freear, and Peter Devine. 2020. Accessible conversational user interfaces: considerations for design. In *Proceedings of the 17th International Web for All Conference*. ACM, Taipei Taiwan, 1–11. doi:10.1145/3371300.3383343

[49] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics* 12 (2024), 157–173. doi:10.1162/tacl_a_00638

[50] Xingyu Liu, Patrick Carrington, Xiang 'Anthony' Chen, and Amy Pavel. 2021. What Makes Videos Accessible to Blind and Visually Impaired People?. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3411764.3445233

[51] Xingyu "Bruce" Liu, Ruolin Wang, Dingzeyu Li, Xiang Anthony Chen, and Amy Pavel. 2022. CrossA11y: Identifying Video Accessibility Issues via Cross-modal Grounding. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology (UIST '22)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3526113.3545703

[52] Yiming Liu, Shengxin Jia, Chun Ki Yiu, Wooyoung Park, Zhenlin Chen, Jin Nan, Xingcan Huang, Hongting Chen, Wenyang Li, Yuyu Gao, Weike Song, Tomoyuki Yokota, Takao Someya, Zhao Zhao, Yuhang Li, and Xinge Yu. 2024. Intelligent wearable olfactory interface for latency-free mixed reality and fast olfactory enhancement. *Nature Communications* 15, 1 (May 2024), 4474. doi:10.1038/s41467-024-48884-z Publisher: Nature Publishing Group.

[53] Yiming Liu, Chun Ki Yiu, Zhao Zhao, Wooyoung Park, Rui Shi, Xingcan Huang, Yuyang Zeng, Kuan Wang, Tsz Hung Wong, Shengxin Jia, Jingkun Zhou, Zhan Gao, Ling Zhao, Kuanming Yao, Jian Li, Chuanlu Sha, Yuyu Gao, Guangyao Zhao, Ya Huang, Dengfeng Li, Qinglei Guo, Yuhang Li, and Xinge Yu. 2023. Soft, miniaturized, wireless olfactory interface for virtual reality. *Nature Communications* 14, 1 (May 2023), 2297. doi:10.1038/s41467-023-37678-4 Publisher: Nature Publishing Group.

[54] Yi Lu, Jing Nathan Yan, Songlin Yang, Justin T. Chiu, Siyu Ren, Fei Yuan, Wenting Zhao, Zhiyong Wu, and Alexander M. Rush. 2024. A Controlled Study on Long Context Extension and Generalization in LLMs. *CoRR* abs/2409.12181 (2024). arXiv:2409.12181 doi:10.48550/ARXIV.2409.12181

[55] Jacob C. Lucas, Zack Arambula, Alexandra M. Arambula, Katherine Yu, Nathan Farrokhian, Linda D'Silva, Hinrich Staecker, and Jennifer A. Villwock. 2022. Olfactory, Auditory, and Vestibular Performance: Multisensory Impairment Is Significantly Associated With Incident Cognitive Impairment. *Frontiers in Neurology* 13 (July 2022). doi:10.3389/fneur.2022.910062 Publisher: Frontiers.

[56] Martin Merkt, Sonja Weigand, Anke Heier, and Stephan Schwan. 2011. Learning with videos vs. learning with print: The role of interactive features. *Learning and Instruction* 21, 6 (Dec. 2011), 687–704. doi:10.1016/j.learninstruc.2011.03.004

[57] Weiqing Min, Shuqiang Jiang, Linhu Liu, Yong Rui, and Ramesh Jain. 2019. A Survey on Food Computing. *ACM Comput. Surv.* 52, 5 (Sept. 2019), 92:1–92:36. doi:10.1145/3329168

[58] Hein Min Htike, Tom H. Margrain, Yu-Kun Lai, and Parisa Eslambolchilar. 2021. Augmented Reality Glasses as an Orientation and Mobility Aid for People with Low Vision: a Feasibility Study of Experiences and Requirements. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–15. doi:10.1145/3411764.3445327

[59] Mohammadreza Mirzaei, Peter Kán, and Hannes Kaufmann. 2020. EarVR: Using Ear Haptics in Virtual Reality for Deaf and Hard-of-Hearing People. *IEEE Transactions on Visualization and Computer Graphics* 26, 5 (May 2020), 2084–2093. doi:10.1109/TVCG.2020.2973441

[60] Zheng Ning, Brianna L Wimer, Kaiwen Jiang, Keyi Chen, Jerrick Ban, Yapeng Tian, Yuhang Zhao, and Toby Jia-Jun Li. 2024. SPICA: Interactive Video Content Exploration through Augmented Audio Descriptions for Blind or Low-Vision Viewers. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–18. doi:10.1145/3613904.3642632

[61] OpenAI. 2023. OpenAI Text-to-Speech Model. https://platform.openai.com/docs/guides/text-to-speech

[62] Sharon Oviatt. 1999. Mutual disambiguation of recognition errors in a multimodel architecture. In *Proceedings of the SIGCHI conference on Human factors in computing systems the CHI is the limit - CHI '99*. ACM Press, Pittsburgh, Pennsylvania, United States, 576–583. doi:10.1145/302979.303163

[63] Shishir G. Patil, Don Kurian Dennis, Chirag Pabbaraju, Nadeem Shaheer, Harsha Vardhan Simhadri, Vivek Seshadri, Manik Varma, and Prateek Jain. 2019. GesturePod: Enabling On-device Gesture-based Interaction for White Cane Users. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology (UIST '19)*. Association for Computing Machinery, New York, NY, USA, 403–415. doi:10.1145/3332165.3347881

[64] Yash Prakash, Akshay Kolgar Nayak, Shoaib Mohammed Alyaan, Pathan Aseef Khan, Hae-Na Lee, and Vikas Ashok. 2024. Improving Usability of Data Charts in Multimodal Documents for Low Vision Users. In *International Conference on Multimodel Interaction*. ACM, San Jose Costa Rica, 498–507. doi:10.1145/3678957.3685714

[65] Oliver Schneider, Jotaro Shigeyama, Robert Kovacs, Thijs Jan Roumen, Sebastian Marwecki, Nico Boeckhoff, Daniel Amadeus Gloeckner, Jonas Bounama, and Patrick Baudisch. 2018. DualPanto: A Haptic Device that Enables Blind Users to Continuously Interact with Virtual Worlds. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology (UIST '18)*. Association for Computing Machinery, New York, NY, USA, 877–887. doi:10.1145/3242587.3242604

[66] Amarjot Singh, Ketan Bacchuwar, and Akshay Bhasin. 2012. A Survey of OCR Applications. *International Journal of Machine Learning and Computing* (2012), 314–318. doi:10.7763/IJMLC.2012.V2.137

[67] Joel Snyder. 2005. Audio description: The visual made verbal. *International Congress Series* 1282 (Sept. 2005), 935–939. doi:10.1016/j.ics.2005.05.215

[68] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. MovieQA: Understanding Stories in Movies Through Question-Answering. 4631–4640. https://openaccess.thecvf.com/content_cvpr_2016/html/Tapaswi_MovieQA_Understanding_Stories_CVPR_2016_paper.html

[69] Lily M. Turkstra, Tanya Bhatia, Alexa Van Os, and Michael Beyeler. 2025. Assistive technology use in domestic activities by people who are blind. *Scientific Reports* 15, 1 (March 2025), 7486. doi:10.1038/s41598-025-91755-w Publisher: Nature Publishing Group.

[70] Stephanie Valencia, Mark Steidl, Michael Rivera, Cynthia Bennett, Jeffrey Bigham, and Henny Admoni. 2021. Aided Nonverbal Communication through Physical Expressive Objects. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '21)*. Association for Computing Machinery, New York, NY, USA, 1–11. doi:10.1145/3441852.3471228

[71] Tess Van Daele, Akhil Iyer, Yuning Zhang, Jalyn C Derry, Mina Huh, and Amy Pavel. 2024. Making Short-Form Videos Accessible with Hierarchical Video Summaries. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–17. doi:10.1145/3613904.3642839

[72] Yujia Wang, Wei Liang, Haikun Huang, Yongqi Zhang, Dingzeyu Li, and Lap-Fai Yu. 2021. Toward Automatic Audio Description Generation for Accessible Videos. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3411764.3445347

[73] R. S. Woodworth and E. L. Thorndike. 1901. The influence of improvement in one mental function upon the efficiency of other functions. (I). *Psychological Review* 8, 3 (1901), 247–261. doi:10.1037/h0074898 Place: US Publisher: The Macmillan Company.

[74] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2021. Just Ask: Learning To Answer Questions From Millions of Narrated Videos. 1686–1697. https://openaccess.thecvf.com/content/ICCV2021/html/Yang_Just_Ask_Learning_To_Answer_Questions_From_Millions_of_Narrated_ICCV_2021_paper.html?ref=https://githubhelp.com

[75] YouTube. [n.d.]. YouTube for Press. https://blog.youtube/press/ Accessed: 2025-04-07.

[76] Yuhang Zhao, Edward Cutrell, Christian Holz, Meredith Ringel Morris, Eyal Ofek, and Andrew D. Wilson. 2019. SeeingVR: A Set of Tools to Make Virtual Reality More Accessible to People with Low Vision. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3290605.3300341