OmniActions: Understanding and Predicting Follow-Up Actions On Multimodal Information Using Large Language Models

ANONYMOUS AUTHOR(S)

SUBMISSION ID: 7980

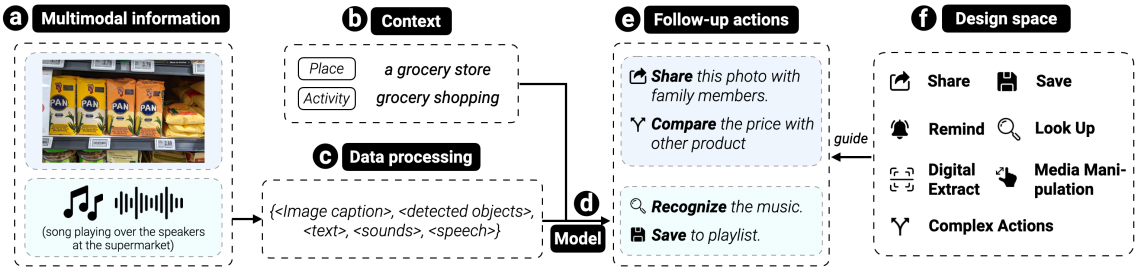

Fig. 1. OmniActions predicts possible follow-up actions on multimodal information. To do so, OmniActions first processes the multimodal information (a) into a standardized format (c). Then, by incorporating contextual information (b), the system feeds the processed data into a fine-tuned large language model (d), which predicts potential follow-up actions on the information (e). The actions are selected from a design space based on the data collected from a five-day diary study with 39 participants (f).

Users constantly engage with a diverse range of multimodal information in physical environments. With the emergence of ubiquitous computing devices such as smartphones and smart glasses, there is an opportunity to not only capture such information, but to also perform follow-up actions in real-time, such as sharing a moment with friends or recognizing music. To minimize the friction that users experience while performing such follow-up actions, one could harness an intelligent interface that proactively suggests follow-up actions to perform given the multimodal information in one's environment. To explore the range of follow-up actions, we conducted a diary study that prompted participants to capture and share the media that they wished to take action on (e.g., images or audios), along with their desired follow-up actions and other contextual information. With the data, we generated a holistic design space of follow-up actions that could be performed with different types of multimodal information. We then developed OmniActions, an interactive prototype that recommends follow-up actions for the information that users encounter in their environment. OmniActions leverages a large language model, fine-tuned using the collected data, to proactively predict follow-up actions. We report on the results from an in-lab qualitative evaluation that collected feedback on the proposed design space and insights to improve the interactive system.

CCS Concepts: • **Human-centered computing** → **User studies**; **Interactive systems and tools**; *Interaction techniques*.

Additional Key Words and Phrases: follow-up actions, predictive interface, large language models, AI agent, dataset, augmented reality, diary study, open coding

ACM Reference Format:

Anonymous Author(s). 2018. OmniActions: Understanding and Predicting Follow-Up Actions On Multimodal Information Using Large Language Models. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 29 pages. https://doi.org/XXXXXXX.XXXXXXX

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2018 Association for Computing Machinery.

Manuscript submitted to ACM

1 INTRODUCTION

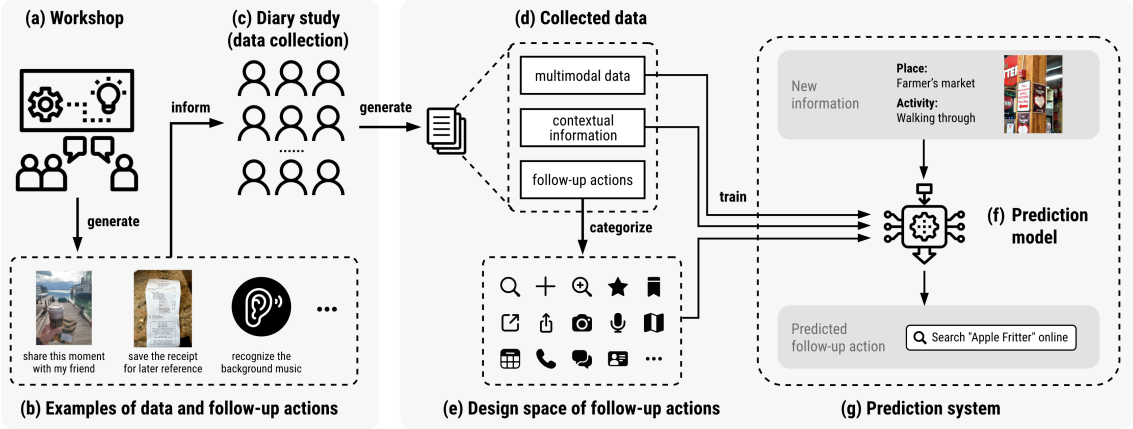

Fig. 2. The development process for OmniActions. (a) An internal workshop was conducted to (b) generate informative examples of situations when users may take using multimodal information. (c) The examples were used to inform and inspire the participants during a diary study that (d) collected data when participants wished to take action using multimodal data. (e) The follow-up actions submitted by participants were then analyzed and categorized into a design space. (f) The collected data included contextual information that was used employed to train a prediction model that was (g) integrated within OmniActions to predict multiple follow-up actions given multimodal information.

Multimodal information refers to information that is presented in different modalities such as text, image, audio, and video [6]. Users constantly engage with a diverse range of such information in different ways depending on their context, intentions, and goals. While users rely on traditional user interfaces to execute tasks in many contexts, this could take many clicks in launching an app, navigating to the desired action and parameterizing it. Additionally, there exist circumstances that necessitate more rapid access to such follow-up actions [4]. These include scenarios where users are on-the-go or multitasking, where manual input becomes infeasible (*e.g.*, users' hands are occupied while cooking), where voice commands might be deemed inappropriate (*e.g.*, in a formal meeting), or where handheld devices are absent, such as in Augmented Reality (AR) systems [31].

Interactive interfaces, regardless of their form, can mitigate this by proactively suggesting follow-up actions for multimodal information so that users do not need to locate an application from their library or navigate through menus to find target actions. ?to find or discover new information. As opposed to existing products such as iOS text-in-a-photo action suggestion[1], Google Lens[2], or Shazam's song recognition[3], which typically recognize one specific modality of information (*e.g.*, structured text, image or audio) and map to hardcoded predefinded actions (*e.g.*, detect an address and launch navigation), these proactive interfaces need to have the ability to adapt to diverse contexts and varying user objectives. This necessitates an inherent ability to comprehend information across multiple modalities and to generate a more generalized follow-up action prediction. To realize such functionality, two essential tasks need to be accomplished: *(i)* understand the design space of potential follow-up actions across multiple modalities to inform the future development

[1]https://support.apple.com/en-us/HT212630
[2]https://lens.google/
[3]https://www.shazam.com/

of proactive interactive systems and *(ii)* develop models to predict such actions adaptively in response to the diverse information encountered.

While researchers have conducted several studies to investigate mobile information needs in daily life [13], as well as the intent behind the usage of mobile applications [5, 10, 14, 19] and web searches [34, 49], there remains a gap in understanding the potential actions users might take when encountering new information in physical environments. Several systems have also been proposed to address specific interaction and action needs such as retrieving real world information (*e.g.*, image search [9]) or visualizing contextual information using augmented reality [3]. Few systems predict users' goals or suggest follow-up actions based on them. Furthermore, it is unknown how to interact with such multimodal information (*e.g.*, sound) due to the increased complexity of analyzing and interpreting the various modalities involved.

To provide a comprehensive understanding of potential follow-up actions for visual information (*e.g.*, scenes, physical objects, texts) and auditory information (*e.g.*, acoustic sounds, human speech), we conducted a diary study, wherein participants contributed 382 data entries that recorded information that they wished to take action on (Figure 2cd). The participants in the diary study were guided by examples generated from an internal workshop, ensuring proper understanding of the definition and discovering potential follow-up actions (Figure 2ab). The collected data informed a design space of follow-up actions that can serve as a guideline for the design of future interactive systems (Figure 2e).

To utilize the design space and demonstrate its potential, we developed a system called OmniActions, which processes new information such as visible text, physical objects, the whole scene, acoustic sounds and human speech as input. The system then outputs the target information (*e.g.*, the visible text) and appropriate follow-up actions from the design space (*e.g.*, share with another person) (Figure 2g). The system is powered by a large language model (LLM) by first converting the multimodal information into a text format (*e.g.*, by interpreting the scene via an image captioning model) and then harnessing the reasoning capabilities of LLMs [27] to predict users' follow-up actions. This reasoning capabilities enables the LLMs to articulate the rationale behind its action prediction. Evaluating on the data collected in the diary study, we compared multiple techniques of using LLM and the result indicates that few-shot prompting with chain-of-thought using the latest model (`GPT-4`) performs the best (90.7%) to fully match users' desired actions. The system also features an interactive prototype developed for user interaction. We conducted an in-lab feedback session with 5 participants to collect subjective feedback about the system and insights to improve design and user experiences with the interactive system.

The contributions of this research are thus:

- A design space derived from the diary study data that emphasizes 7 general and 17 specific categories of follow-up actions that could be used as a guideline for developers and designers to design future interactive systems.
- A novel approach of using LLM to reason with real-world multimodal data to predict the follow-up actions. Our approach is tested via three methods: few-shot prompting, fine-tuning via chain-of-thoughts and intent classification. We validated these methods using data collected from the diary study and the results shows that our approach yields competitive performance.
- An interactive prototype that predicts users' target information and suggests follow-up actions. User feedback showed the system's potential and the design space's comprehensiveness.

2 RELATED WORK

The present research was inspired by research on detecting and categorizing user intent, proposed techniques to interact with multimodal information, and prior work using large language models to augment interaction.

2.1 Understanding and Categorizing User Intent

As the automatic detection of user intentions and goals are crucial to the utility of predictive systems, it is imperative that systems have methods to identify and categorize such intentions. One popular method of categorization has been to survey existing examples in the literature to develop design spaces that identify and organize goals and intentions along various dimensions. For example, while using gaze-based interaction, Pfeuffer *et al.* identified the importance of continuum transitions, task transitions, and information level transitions [48]. While creating interactive AR-based toys, Zhu *et al.* identified factors that were common to physical and AR toys including spatial placement, manipulation, and collision [65]. During the design of linkage-based objects, the motion type and force profile of linkages were identified as being important [38, 39]. Although these design spaces help guide the design of software to support specific use case outcomes, they cannot be extended to other tasks as they are dependant on niche-specific goals that differ by use case.

Others have used diary-style research methodologies to understand user intentions. For example, Church *et al.* conducted a three-month diary study to gather insights about user information needs in mobile and non-mobile settings, and developed a taxonomy of ways information needs were met [13]. Zhang, Capra, and Li, on the other hand, analyzed diary responses from participants who work on creative projects and identified six types of information and seven intentions to use such information that exist while searching for information online [63]. Some researchers have also employed in-situ, lab-based studies to understand user intentions, for example during mobile web search interactions [25, 26, 34], mobile app recommendation usage [5], and reality-based information retrieval [9]. Interestingly, Carrascal and Church combined both of these methodologies to conduct an in-wild and a lab-based study to understand mobile app usage and search recommendations [10]. Although this prior research did not synthesize the findings into a design space or taxonomy, the use of diary-style techniques has thus demonstrated that it is valuable to directly probe users about how their goals are reflected in the actions they choose to take. The present research specifically focuses on follow-up actions for multimodal information, which differs from these prior foci yet builds on the importance of using diary-style methodologies to understand user intent.

2.2 Multimodal Information-Based Interaction

To predict follow-up actions when encountering new information in the world (e.g., music, noise, visible text, objects, etc.), it is crucial to understand and process such multimodal information. One approach involves directly retrieving information from the physical world embedded in barcodes, fiducial markers [20], human faces [2], and objects during fabrication processes [16, 17, 37]. Researchers have also explored retrieving "raw" information such as visible text [53, 62], physical objects [21, 50], multimodal scenes [61], human speech (e.g., Google API[4]), and music (e.g., Shazam). Nevertheless, to understand users' intent based on information in physical environments, multimodal information must be monitored and processed in a way that a system can make predictions on.

Lifelogging offers a method for processing such multimodal information by digitally tracking a person's daily experiences [23, 35]. Example applications include enhancing human memories by retrieving moments through natural language [18, 56], or health monitoring by analyzing logged data [36]. However, lifelogging does not specifically focus on predicting user intent and in order to predict follow-up actions, which requires the categorization of the design space. Several lifelogging datasets have been collected, including the Aria dataset [43], Ego4D [22], and other video datasets [51, 52]. These datasets could be used to investigate desired follow-up actions, but they contain considerable redundant data when such actions are not required. To specifically explore follow-up actions on multimodal information,

[4]https://cloud.google.com/speech-to-text

we conducted a diary study prompting participants to log data whenever they needed to act on their captured information. Building on prior research in processing multimodal information, we used this data to develop a system capable of predicting follow-up actions.

2.3 Large Language Models in HCI

AI has been widely used in the Human-Computer Interaction (HCI) community, with LLMs experiencing a surge in usage in recent years [1, 15, 24, 29, 30, 32, 46, 47, 58, 59]. LLMs have demonstrated their capabilities in understanding common knowledge and reasoning within context in a broad range of research in HCI community including interactive code support [58], social computing [45, 46] and accessibility support [28]. For instance, Visual Caption employs a fine-tuned language model to predict user intent in visual inquiries based on the last two sentences [42], while SayCan extracts and leverages knowledge priors within LLMs to reason about and execute robot commands [1]. LLMs can further enhance recommender systems, which utilize contextual information to recommend items [8, 33]. For example, GPT-3 [7] has been used to augment movie recommendation systems [64].

However, most of the prior work relies on explicit intent [12], where users or agents interact with the system via direct prompts. OmniActions unlocks a new interaction method with the LLM by embracing a more implicit intent, centered on the user's current visual input (e.g., multimodal information like environmental understanding or recognized text) and contextual information. Coupled with the chain-of-thoughts mechanism, this enables OmniActions to deliver explainable predictions on both the target information and the follow-up actions.

3 FORMATIVE WORKSHOP

We ran a formative workshop to gain a preliminary understanding of the multimodal information triggers and design space for related actions. The outcome of the workshop include: 1) the clusters of the actions people took with multimodal information triggers; 2) iterations on the research methods and questions we use to collect data from general population later; 3) examples of data.

3.1 Procedure

We recruited 10 participants within our institution through group email invitations. The participants included HCI researchers, UX designers, and student interns. The workshop consisted of three parts, lasting one hour in total. Participants were invited to use a FigJam[5] whiteboard for synchronous collaboration.

3.2 Methodology

The organizer of the workshop first walked through two exemplary information triggers and their context and follow up actions, including an image of a parking ticket and an audio file of background music. In part 1, participants were then asked to share their own media, context, and follow-up actions; in part 2, the participants reflected on other participants' media and come up with their follow-up actions; and in part 3, participants collaboratively clustered similar actions (Figure 3).

3.2.1 Part One. 'Browse past media and recall what you wanted to do with the information". During this session, participants were given 20 minutes to browse their personal storage, upload media to the Google drive and the FigJam board. For each shared media item, participants were asked to recall and record the following information: *(i)* the context

[5]https://www.figma.com/figjam/

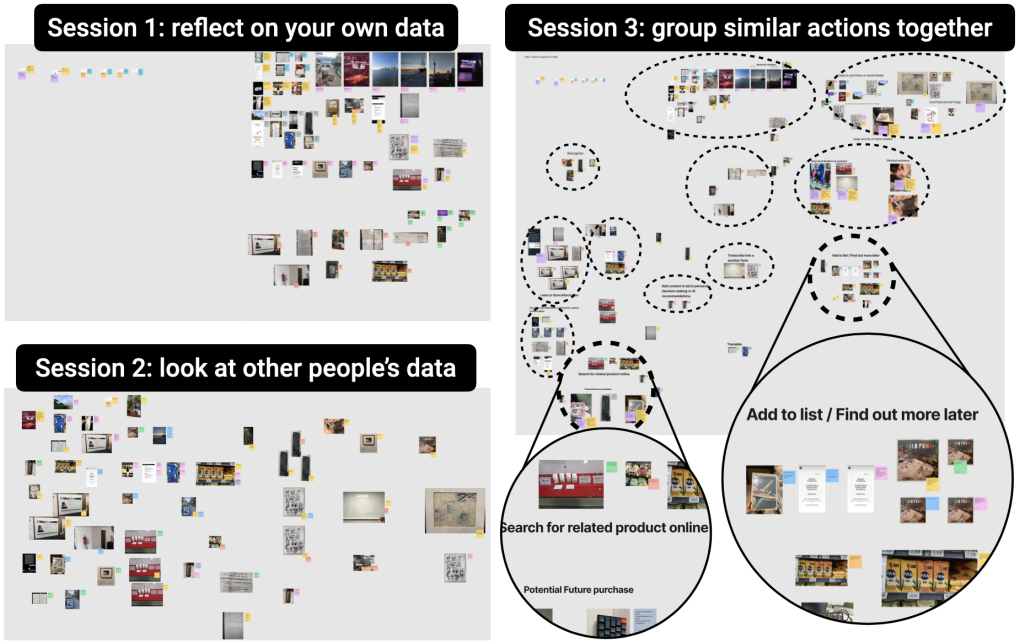

Fig. 3. Screenshots from the formative workshop where participants shared data in Session 1, reviewed other people's data in Session 2, and grouped similar actions in Session 3.

when they captured the image including the place and activity, *(ii)* the target information (*e.g.*, the whole scene of the image), and *(iii)* what they wanted to do with the information afterwards. For audio and video, an additional description of the media was required. Participants shared a total of 66 examples (i.e., 6 video/audio clips and 60 images) and 66 follow-up actions.

3.2.2 Part Two. " Imagine if you were the person, what actions you would take on the information?" Participants were given another 20 minutes to browse examples shared by other participants and to input follow-up actions for the target information. The goal of this session was to provide a third-party opinion about the captured data to potentially collect more diverse follow-up actions. An additional 104 follow-up actions were proposed for a total of 170 follow-up actions between session one and two.

3.2.3 Part Three. "Now group together those actions that are similar." For the last 15 minutes, participants collaboratively clustered all 170 examples from Part 2, using affinity diagram as the method. The participants also labeled each of the clusters.

3.3 Results

After the workshop, two researchers clustered, filtered, and organized the 170 follow-up actions independently. Then the results from each researcher were compared and any discrepancies were discussed and resolved until a consensus was reached. This process is inductive, meaning that we openly extract and record any action mentioned by the participants rather than starting with an existing set of actions. As a result, we identified 13 types of actions that were grouped into

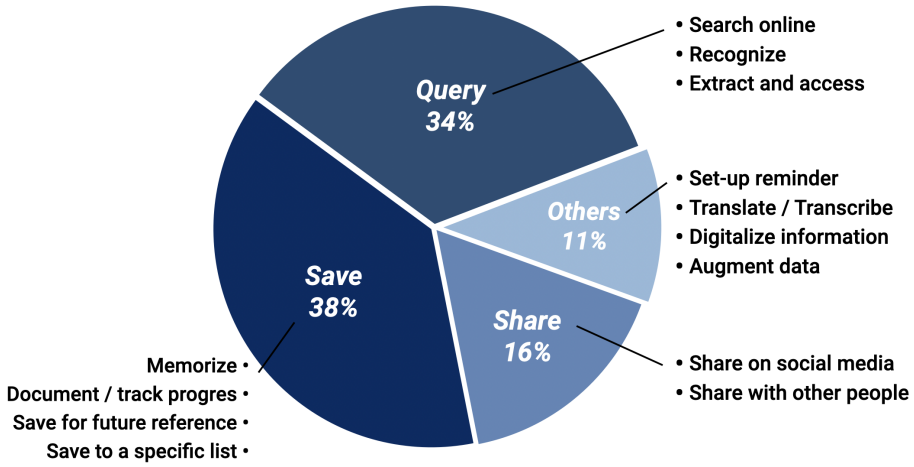

Fig. 4. Frequencies of the 13 follow-up actions generated during the workshop (n = 170) that were grouped into 4 categories.

four categories (i.e., share, save, query, and others; Figure 4). Representative examples from these categories were used as learning materials for participants in the subsequent study.

One important observation was that participants seldom captured or shared audio. This might be due to the fact that audio contains temporal information that is hard to capture (*e.g.*, an abnormal sound that occurs intermittently). This finding informed the design of the user study, where we asked users to share the textual description of their audio rather than the audio itself. We present more details in the next section.

4 DATA COLLECTION WITH A DAIRY STUDY

While the workshop provides an initial glimpse of type of multimodal information and follow-up actions, we needed to formalize the findings with in-situ experiences of external participants. Diary studies allow participants to log data whenever needs arise [57], making it an ideal choice for examining desired follow-up actions when encountering new information. We leverage the diary study to answer the following research question:

RQ: Which follow-up actions do general users wish to take when they encounter new multimodal information in a physical environment?

We conducted a diary study where participants were asked to report on the target information and their desired follow-up actions, in addition to contextual information such as their goals, reasons for actions, locations and activities. Contextual information was important to collect as it affects the choice of follow-up actions [11, 40, 54]. For example, looking at a shampoo bottle in a drug store has a different desired follow-up action than looking at the same bottle at home (*e.g.*, comparing the price to a similar product versus ordering another bottle on Amazon). Therefore, we hypothesized that contexts would increase the accuracy of understanding users' goals and follow-up actions. We incorporated them into a predictive model later during our research process.

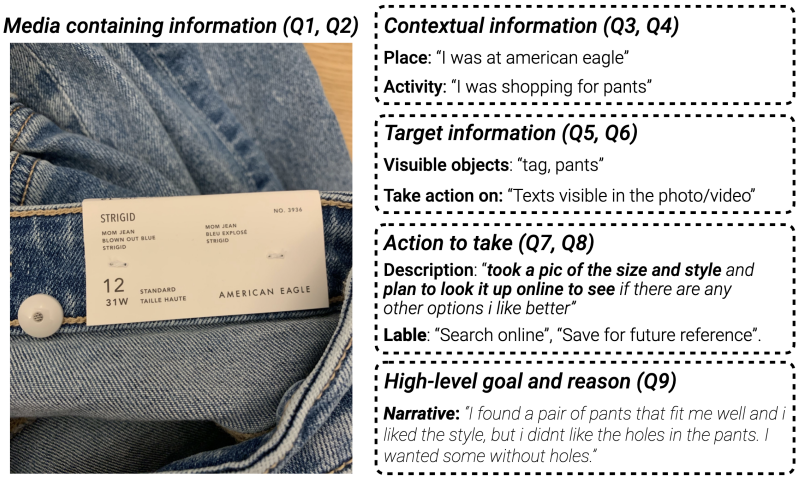

Fig. 5. An example diary entry of the diary study.

4.1 Participants

Thirty-nine participants (i.e., 16 male, 22 female and 1 non-binary) were recruited from the dscout user research platform[6]. All the participants were between the ages of 18 to 69 years old who were proficient in English and had a smart phone to take photos. Each participant was compensated with 50 US dollars after they completed the diary study.

4.2 Procedure

The diary study consists of two phases, an introductory phase and the diary phase. In the introductory phase, participants were shown examples from the workshop that represented several of the categories that were identified. Note that participants were only shown the exemplar media and follow-up actions. The categorization of the actions was hidden to avoid bias.

In the diary phase, participants were instructed to submit 2 entries each day for five days. These entries needed to be about the genuine needs participants had at that moment. The diary phase began in the middle of the week and extended over the weekend to capture the different types of needs in a week. For each diary entry, participants were required to answer questions about the media that contained the information, relevant contextual information, the target information modality, actions to be taken, and their high-level goals and reasons for taking these actions. Depending on the information modality (visual or audio), the survey included different questions to accommodate different attributes of the information. Specifically, a data entry included the following aspects:

Media Containing the Information (Q1, Q2). Although we aim to collect multimodal information, we were not allowed to collect audio or video data from participants that could contain potentially identifiable personal information due to the legal requirement from our institution. Therefore, if participants wished to share audio or video, they were asked to provide a text description of the data (or screenshots for videos) instead (*e.g.*, "This is the background music I heard in the cafe").

[6]https://dscout.com/

Contextual Information (Q3, Q4). Context was first introduced by Schilit *et al.* as *"locations, identities of nearby people and objects, and changes to those objects"* [55]. In our use case to predict follow-up actions, we identified how the location and the user's activity would affect how users would interact with the encountered information.

Target Information (Q5, Q6). Since we were investigating follow-up actions for multimodal information, it was essential to know which information the participant was interested in taking follow-up action about. For example, participants could be interested in the text visible in an image or the whole scene to share with friends. Participants were thus asked to identify the objects visible in the image or the sounds that could be heard (Q5). This provided additional context to achieve a better understanding of the user interaction with the information.

Actions to be Taken (Q7, Q8). Participants were asked to use natural language to describe the actions they intended to take and categorize these actions. Additionally, they can select the categories of the actions based on the action categories identified in our formative workshop. Participants had the option to create new categories. Note that we minimize the bias by having participants detail their intention and desired actions in their own words on a first page before being shown and ased to choose from the action types on the next page. The participants selected categories that were later used as a reference point during the iteration towards the final design space presented in the following sections.

High-Level Goal and Reasons (Q9). To better understand why participants intended to take certain follow-up actions, we asked participants to share their high-level goals and reasons.

4.3 Data Analysis

In total, we collected 382 diary entries from participants. An example entry is shown in Figure 5. The ratio of collected visual to to audio data was approximately 2:1. We collected 254 visual data examples (i.e., 193 photos and 61 videos with visual as the target) and 128 audio data examples (i.e., 48 videos with audio as the target and 80 text descriptions of audio). Participants reported wanting to take action on 55 full scenes, 120 objects, 79 pieces of text, 51 clips of speech, and 77 clips of sounds. Additionally, participants shared 17 (10 visual, 7 audio) follow-up actions which did not fit any of the existing categories.

5 DESIGN SPACE OF FOLLOW UP ACTIONS

During the diary study, after participants described the follow-up actions they wish to take on the target information, they labeled their response using the categories that were generated during the internal workshop. If there were actions that cannot be labeled with the existing list of action categories, users can select "other". From the users' natural language description of the follow-up actions (especially those marked as "other"), we identified more action categories than what we originally found in the workshop. We also verified the existing categories of actions also apply to the diary data. This work was done by a researcher and a research assistant. They collaboratively reviewed the diary entries, manually labeled the data, and refined the design space through a process of comparison and analysis. The resulting design space consisted of 7 **general categories** of follow-up actions, including *share*, *save*, *remind*, *look up*, *digital extract*, *media manipulation* and *complex actions*, that were broken down into 17 **specific categories** (Figure 6).

5.1 Follow-Up Action Categories

5.1.1 ***Share.*** *Share* refers to the actions that people use to make information available to others (i.e., sending information to friends or family or posting the information on a social media platform such as Instagram or Facebook).

General category	Specific category	Definition	Exemplar usage
Share	Share with others	*Send to specific entity(s)*	Share with family members, friends, etc
	Share on social media	*Share/upload on social platforms*	Post on instagram/Facebook/etc
Save	Remember	*Cherish a specific experience/moment for later recall*	e.g., "I want to capture the moment for him as it is memorable."
	Save for reference	*Store information for later usage or consultation*	e.g., "I took a picture of the product to purchase it later."
	Save to list	*Add information to a designated, organized collection*	Add song to playlist / add artwork to favorites album
	Keep track of progress	*Record the development of a task or goal*	e.g., "record my son's progress at painting"
Remind	Set up reminder	*Make an alert or notice to remember something later*	Save an event to the calendar
Look Up	Search online	*Search for more information online related to specific goals*	e.g, "I plan to look up more info on the event online."
	Recognize	*Identify the information using specific tools*	Shazam background music/ google lens to search product
	Translate	*Translate text/speech from one language to another*	e.g., "translate the Korean to English"
Digital Extract	Extract and access	*Extract and utilize information from sources*	Extract and access QR codes, URLs, addresses, etc
	Transcribe	*Convert audio to text*	e.g., "transcribe professors speech"
	Digitalize	*Transform information to a digital format for easier access*	Scan documents to digital copies.
Media Manipulation	Augment visual/audio	*Enhance images or sounds to improve overall experience*	Zoom in the photo / filter the noise / etc
	Edit media	*Modify media files to accomplish a specific tasks*	e.g., "I want to trim the video to post it on TikTok later"
Complex	Compare	*Compare similarity and difference between two sets of info*	Compare the price between two different products
	Calculate	*Perform mathematical operations to solve a problem/task*	e.g., "I want to add the calories to see if it fits my goal today."

Fig. 6. Design space of follow-up actions for multimodal information that emphasizes general and specific categories of actions.

Sharing with Others. When *sharing with others*, future systems may leverage additional contextual information such as recommending people who have recently expressed their love for dogs when a user takes a photo of their dog.

Sharing on Social Media. When *sharing on social media*, future systems may suggest multiple hashtags to use. Having the ability to distinguish between these two actions would enable more tailored user experiences in future applications.

5.1.2 ***Save***. *Save* refers to the actions used to store information. We identified four types of *save* actions:

Remember. This refers to an action where users wish to cherish a specific moment so that they will be able to retrieve the memory in the future. *Remember* often occurs when participants mention words such as "funny", "memorable", etc. We also noticed that *remember* often occurs alongside other *share* actions.

Save for Reference. This refers to the actions where users store information with the specific goal to use it later. Participants mentioned various types of *later usage*, including using it as a reference for a later purchase, saving a gift card to avoid losing it, etc. By automatically incorporating metadata into the information (*e.g.*, when, where, what type of object), future systems could enhance user experiences by enabling quick and efficient retrieval of the information when needed.

Save to a List. These actions add information to a designated collection, e.g., music to a playlist. Future systems could leverage this action by identifying the category of the information (*e.g.*, painting, music, groceries) and store it in a list.

Keeping Track of Progress. Users capture information to record their performance or progress towards specific goals. Participants shared experiences such as recording the progress of their bulking (or cutting) while working out or playing the piano. Different from *saving to a list*, the information tends to be similar yet sequential in nature, enabling users to observe and evaluate their growth over time, which could be potentially enabled by future systems.

5.1.3 ***Remind***. *Remind* refers to actions that create an alert or notice to remember something later such as setting a reminder after seeing a flight schedule on a screen or noting the date of a specific event. These actions can be particularly useful for managing tasks, appointments, or important events.

5.1.4 ***Look Up***. *Look up* refers to actions that search for specific information or details. Three types of *look up* actions were identified in the collected data:

Search Online. Users perform actions to conduct online searches to acquire additional information related to their intent, utilizing a variety of search tools (*e.g.*, Google).

Recognize. This refers to actions taken to identify information using specific tools. Participants noted two types of recognition of information: product searching (*e.g.*, using Google Lens or Images) or recognizing music.

Translate. In the context of text or speech, *translate* refers to the actions that seek the meaning of text or speech in a different language, enabling one to better understand and communication across language barriers.

5.1.5 ***Digital Extract***. *Digital extract* refers to actions taken to obtain and utilize information from multiple sources. We identified three types of digital extractions:

Extract and Access. This refers to actions taken to to extract information from the physical world and directly take action on it based on its type. For example, users could directly scan and access the content of a QR code, take a picture of a contact card and directly make a phone call, or extract an address from text and navigate to it.

Transcribe. Mostly applying to audio, *transcribe* refers to actions that convert audio into text. This includes transcribing a lecture or transcribing the lyrics from a song that is playing.

Digitize. This refers to actions that transform various forms of information, such as physical documents or audio, into a digital format for easier access, storage, or sharing. Mostly common *digitize* actions scanned physical information to create a digital copy for easier access and sharing. Digitizing audio, for instance, involves converting voice recordings into digital files, which could then be added to various media, such as TikTok videos.

5.1.6 ***Media Manipulation***. *Media manipulation* refers to actions that alter or modify media content to achieve a specific outcome.

Augment Media. Augment refers to actions that enhance images or sounds to improve overall experiences. For example, participants wanted to zoom in to see the details of an object or isolate music from noise for precise recognition.

Edit Media. This refers to actions that are taken to modify media files for specific tasks. For example, a participant wanted to trim a video to share it on social media. Another participant wanted to crop the image for her slides. These editing actions can range from simple adjustments, such as cropping or resizing, to more complex alterations, such as color grading or adding visual effects.

5.1.7 ***Complex Actions***. We also identified two complex actions that involve processing data from multiple sources.

Compare. Compare refers to actions that compared similarities and differences between two sets of information. One participant, for example, wanted to compare the price of two similar products. This would require a system to retrieve other information and present it together for the user to compare.

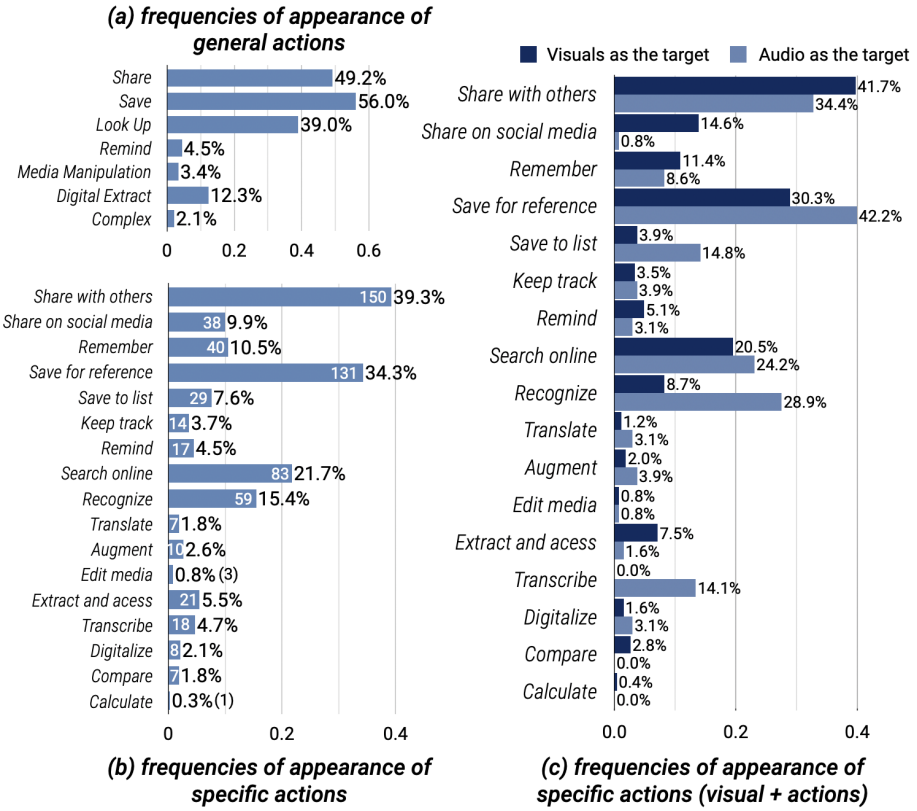

Fig. 7. (a) The frequencies of general actions. (b) The frequencies of specific actions (with number). (c) The frequencies of specific actions on visual and audio.

Calculate. While only mentioned by one participant, *calculate* actions involve performing mathematical operations to solve a problem or a task, e.g., calculating if the calories one consumed exceeded one's daily limit while cutting weight.

5.2 Analysis of Diary Data Using Design Space

We conducted a post-analysis on the diary study data using the categories within our design space. The first pattern we observed was that participants tend to take multiple actions alongside others. For example, participants *remembered* a memorable moment and then *shared* it with family members. Two hundred and eighteen diary entries had only one general action specified by participants, 145 had two general actions, and 19 had three or more actions. For the specific actions, 183 diary entries had only one action specified, 147 had two actions specified, and 52 had two or more actions.

Figure 7 shows the frequencies of each action (general and specific) when participants encounter new information (number of appearance divided by the number of diary entries). The *share* (49.2%), *save* (56.0%), and *look up* (39.0%) actions were most common primary general actions while the remainder of the actions (i.e., *remind* (4.5%), *media manipulate* (3.4%), *digital extract* (12.3%), *complex actions* (2.1%)) were less dominant (Figure 7a). Figure 7b shows the frequencies of each specific action.

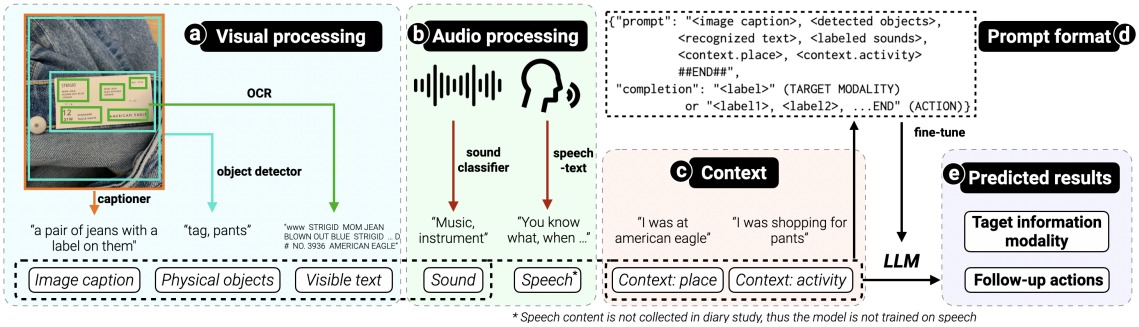

Fig. 8. OmniActions processes visuals by performing image captioning, object detection and text OCR (a), and audio by sound classification and speech recognition (b). Incorporating the context (c), the processed data is format into a tuple (d) and fed into the LLM for prediction (e).

We also observed that participants had different patterns of follow-up actions when interacting with data from different modalities (Figure 7c). The overall frequency of specific follow-up actions when the target was visual versus audio were similar, although there appears to be a difference when *saving to a list*, *recognizing* and *transcribing*. As described earlier, *transcribing* is inherently exclusive to audio. *Recognizing* and *saving to a list* was common for audio as many participants wanted to save music they heard to a playlist.

6 OMNIACTIONS MODEL

6.1 Method

As discussed above, the interactive system should adapt to various contexts to provide generalized follow-up action predictions. By utilizing the design space generated from the diary study—representing a holistic understanding of potential follow-up actions on multimodal information—we can guide the development of future systems to predict actions grounded in the design space. As an instantiation, we developed a system named OmniActions to process multimodal information and predict the **target information** (i.e., the whole scene, physical objects, text, sounds or speech) and the **follow-up actions** (from the design space). To achieve the design goal, two key questions arise:

(1) How should the system understand and process real-world information such that it can be used for a model to predict outputs?
(2) How can we enhance the reasoning ability of the model to optimize the system's performance?

For *(1)*, as OmniActions is designed to predict follow-up actions from a design space by interpreting information from the physical world, the goal is to process multimodal data in a way that enables the model to understand real-world information and discover patterns to predict the actions. Given the challenge of processing image data as pixel values in tandem with other information types such as recognized text or acoustic sound, OmniActions overcomes it by converting the multimodal data into a language representation and leveraging an LLM to execute the prediction.

For *(2)*, traditional classification methods typically rely on trained black boxes. To augment explainability, the model should elucidate the rationale behind its prediction of certain follow-up actions. For example, for the data in Figure 5, when a user captures an image of a label intending to search for more information about its size, in addition to the model's predicted action of "Search online", it should understand that the user is likely interested in information related to the jeans rather than the store's name. Such reasoning can be instrumental for subsequent interactions, such as deciding which

portion of text to search. OmniActions addresses this by introducing chain-of-thoughts [60] as an intermediate reasoning step through the prompting and training process. The details are discussed in the following subsections.

Leveraging the data collected from the diary study, we developed and investigated three different methods of using an LLM to predict the follow-up actions on the data together with the evaluation. These methods consist of *(i)* conventional intent classification (without chain-of-thoughts), *(ii)* few-shot prompting with chain-of-thoughts and *(iii)* finetuning with chain-of-thoughts. Given that chain-of-thoughts are not directly collected in the data, we will first discuss the preparation of the chain-of-thoughts for prompting and training. Then we will detail the three methodologies and present a quantitative assessment of their performance.

6.2 Generating Chain-of-Thoughts

In the diary study, we collected participants' high-level goals and reasons (Sec. 4.2 (Q9)) to understand the rationale behind their intended follow-up actions. Although this closely aligns with the concept of chain-of-thoughts, when predicting follow-up actions, the model's chain-of-thoughts ought to be framed from a third-person viewpoint. In other words, the model should only generate chain-of-thoughts grounded solely in available information, such as the user's current viewpoint or the known context. For instance, when a user took a picture of the label with the following intent (Figure 5):

> "I found a pair of pants that fit me well and i liked the style, but i didnt like the holes in the pants. I wanted some without holes. So i took a pic of the size and style and plan to look it up online to see if there are any other options i like better."

a model's chain-of-thoughts to predict the follow-up actions should be from a third-person perspective:

> "The user was shopping for pants at American Eagle and found a pair they might like. They took a picture of the label, which includes the style and size of the jeans. They may want to look up more information about the specific style of jeans, such as reviews or other colors available."

We prompted an LLM to generate the chain-of-thoughts for the model as the ground truth label for each data point we collected in the diary study. Specifically, the prompt consists of the list of actions with respective description (Figure 6) ground truth action label, and the participants' response of goals and reasons. The prompt template used for generating chain-of-thoughts is listed in Appendix A.1.

6.3 LLM Usage for Prediction

In this section we discuss the three different methods of using LLMs to predict the follow-up actions.

6.3.1 Conventional classifier. LLMs can be fine-tuned to execute classification tasks on text-based data, such as sentiment classification. Given that OmniActions translates multimodal information into linguistic representations, this technique is viable for detecting patterns and training a model dedicated to action prediction (with each action functioning as a classification label) and target information prediction (with each type of information as a label). While being the most cost-efficient method, one drawback is that, especially for action prediction, the model becomes a black box post-training, obscuring the rationale behind the prediction. To address this drawback, we investigated two other approaches by incorporating chain-of-thoughts during the prompting or training process, which will be discussed in Sec. 6.3.2 and 6.3.3.

Procedure: The first step to fine-tuning the model was to format the data into a tuple (Figure 8d) by conducting the processing steps on the multimodal information. We finetuned the model separately for target information prediction and

action prediction. The model was finetuned on the `davinci` model by OpenAI[7] and 75% of the data entries from the diary study were used for training and the rest is used for evaluating the performance.

6.3.2 Few-shot prompting with chain-of-thoughts. Prompting is another typical approach of using LLMs. For enhanced explainability, we designed the prompt to instructs the LLM to produce intermediate reasoning (chain-of-thoughts) prior to the final action prediction.

Procedure: We used both `GPT-3.5-turbo` and `GPT-4` as the model for the few-shot prompting method. While `GPT-4` outperforming reasoning capability [41], it does not support finetuning[8]. To clearly illustrate the performance differences between few-shot prompting and finetuning, we opted to also utilize `GPT-3.5-turbo` in the evaluation. While the input of each data entry remains the same as in Figure 8, additional information is provided including the role and task of the model and the list of actions with their respective description. Furthermore, the output contains not only the prediction label but also the chain-of-thoughts. Following the guidelines recommended by Brown *et al.* [7], we pick two data entries as the few-shot examples for each prediction and this approach is evaluated on the rest 380 data entries. Please refer to the full prompt in Appendix A.2.

6.3.3 Fine-tuning with chain-of-thoughts. Similar to the few-shot method with chain-of-thoughts, we investigated a technique that fine-tunes the model with the data. Essentially, it is similar to few-shot prompting but with a larger sample size beyond the token limit for prompting.

Procedure: The format is the same as the few-shot approach and only `GPT-3.5-turbo` supports finetuning.

6.4 Model Evaluation

6.4.1 Predicting The Target Information to Take Actions On. The target modality of information could be one of five modalities: the whole scene (*e.g.*, capture the whole moment or share a view with friends), physical objects (*e.g.*, recognizing a specific product and search online), text visible in a visual (*e.g.*, save a promo code on a gift card), speech (*e.g.*, transcribe the teacher's lecture) or acoustic sound (*e.g.*, recognize background music). As 80 diary entries were audio-only and there was only a text description of the audio without any visual information, we decided to separate the target modality prediction for visual (i.e., scenes, objects, and text) and audio (i.e., speech and sounds) information.

Other than testing with the intent classification method, we also tested it with few-shot prompting the LLM without chain-of-thoughts. The result shows that the intent classification method achieved an accuracy of 70.6% when predicting the target modality from visual information and 92.3% when predicting the target modality from auditory information, which outperformed the direct prompting method (Table 1).

Table 1. Accuracy (%) when predicting the target information.

Model	Visual	Audio
Few-shot prompting (w/o Chain-of-Thoughts)	60.8	84.6
Target Information Classification	**70.6**	**92.3**

[7]https://platform.openai.com/docs/guides/fine-tuning
[8]as of September 14th, 2023

Table 2. Accuracy (%) when predicting actions using the full-match metrics.

# num of Pred	**Intent classification**	**Fine-tuning w/ COT** (GPT-3.5)	**Few-shot prompting w/ COT** (GPT-3.5)	**Few-shot prompting w/ COT** (GPT-4)
Predicting General Actions				
1	46.0	59.0	58.1	**61.7**
2	61.1	64.7	66.8	**70.6**
3	83.1	80.1	79.7	**90.7**
Predicting Specific Actions				
1	**41.7**	39.7	30.0	39.1
2	40.6	**46.8**	35.0	45.9
3	54.3	55.3	36.7	**59.2**

*Note: Few-shot prompting is tested on 380 data entries.

6.4.2 Predicting Follow-Up Actions. Due to the fact that the users may take multiple actions upon the same information, we used the metric of *full-match* (i.e., accuracy was calculated by dividing the number of correctly predicted results by the minimum of either the number of ground truth labels or the number of predictions) to evaluate the model performance. The full-match metric aims to show the degree of alignment between the predictions and ground truth labels of actions.

The result demonstrated that when introducing chain-of-thoughts, the model has a better performance than the conventional intent classification method (Table 2). Notably, it achieved very high accuracy on general action when predicting top three possibilities (90.7%) and exhibited improvement when predicting multiple specific actions (59.2%). The result also illustrates that while few-shot prompting with `GPT-4` surpassed the finetuning method, the distinction is due to the varying reasoning capabilities between `GPT-3.5` and `GPT-4` models. When utilizing the same model (`GPT-3.5-turbo`), the finetuning approach surpassed few-shot prompting. This suggests a potential improvement of the performance once `GPT-4` supports finetuning.

7 OMNIACTIONS PROTOTYPE

To demonstrate the effectiveness of the prediction model and to exemplify a practical application of our design space, we developed an interactive prototype (an Android app) to execute the prediction of follow-up actions.

7.1 Interaction Flow

We walk through an example to demonstrate OmniActions's user experience in which the user intends to search the product name of the chocolate online (Figure 9). Firstly, the user can click the visual or audio button to specify the modality of information they are interested in. In this case, the user clicks the visual button (a) and the system performs the *target modality* prediction and *follow-up action* prediction. In this example, the system predicts the target as text (b) and recommends three specific actions. If the user finds that the suggested actions do not fit their needs, they could click the *more* button to expand other actions in the design space (d). In the example, the user selects the target attribute of the text ("product name") (b) and select the *Search Online* action (c). After selection, a pop-up window visualizes the users intent to search for the product name ("MILK CHOCOLATE TOFFEE ALMONDS") online (e). As the system does not currently detect all the contexts automatically, users can manually specify place and activity in the console (Figure 10) for better prediction performance. Additionally, users can toggle between predicting general actions versus specific actions to view raw processed results for potential explainability in the console view as well.

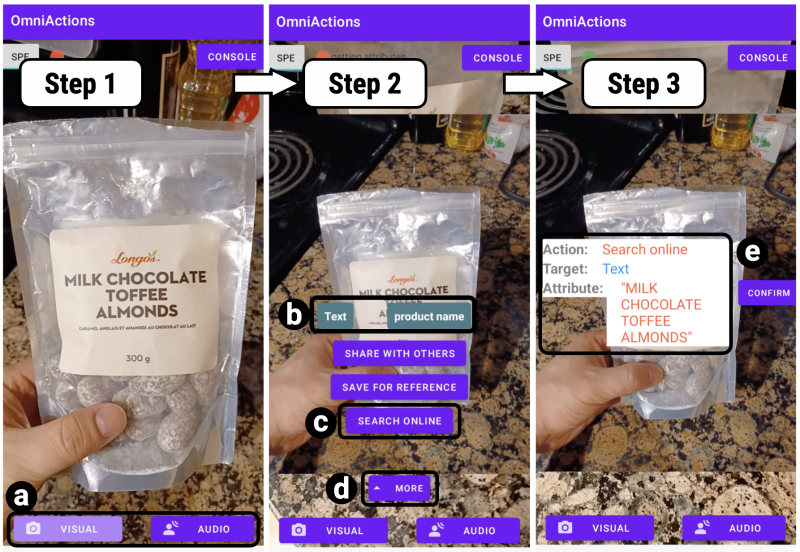

Fig. 9. *OmniActions*'s user interface, wherein (a-e) a user could search the product name on the bag of chocolate by selecting the follow-up actions suggested by the interactive system.

7.2 Implementation

The OmniActions prototype was composed of two primary modules: a continuous detection module and a trigger-based detection module. The continuous detection module was responsible for classifying the sounds and transcribing speech (if present) in real-time and stored the classified sounds and speech transcription from the previous five seconds for further processing. In contrast, the trigger-based module adopts the similar process in Sec. 6.3 by captioning the captured images to provide a description, detecting objects within the captured images, and using OCR to identify and extract text in the images. Once the user triggers the system, OmniActions processes all the information into the tuple format so that it can be used by the fine-tuned model for prediction.

The system was implemented on a Samsung Galaxy A13 5G phone running Android version 13.0. The code was developed in Android Studio with API level 33 and written in the Kotlin programming language. The image captioning on the phone utilized the `blip-image-captioning-base` via the Hugging Face API, the object detection used `MobileNet V1`, and the text recognition used the Google Cloud Vision API. The audio classification used `YAMNet` and the continuous speech-to-text recognition used the Google Cloud Speech API.

7.3 User Feedback

We conducted a user study with a think aloud protocol [44] to assess the usability of the prototype, identify limitations, and gather feedback from end-users on potential improvements to enhance the user experience.

7.3.1 Method. Five participants were recruited from a tech company to participate in the study. Participants had either programming or product development experience. The study took place in a lab designed to resemble a cafe, where we set up everyday life scenarios such as viewing the menu afar and interacting with a book on the bookshelf. During the study, the researcher first described the system and walked through the basic functionality by demonstrating an example.

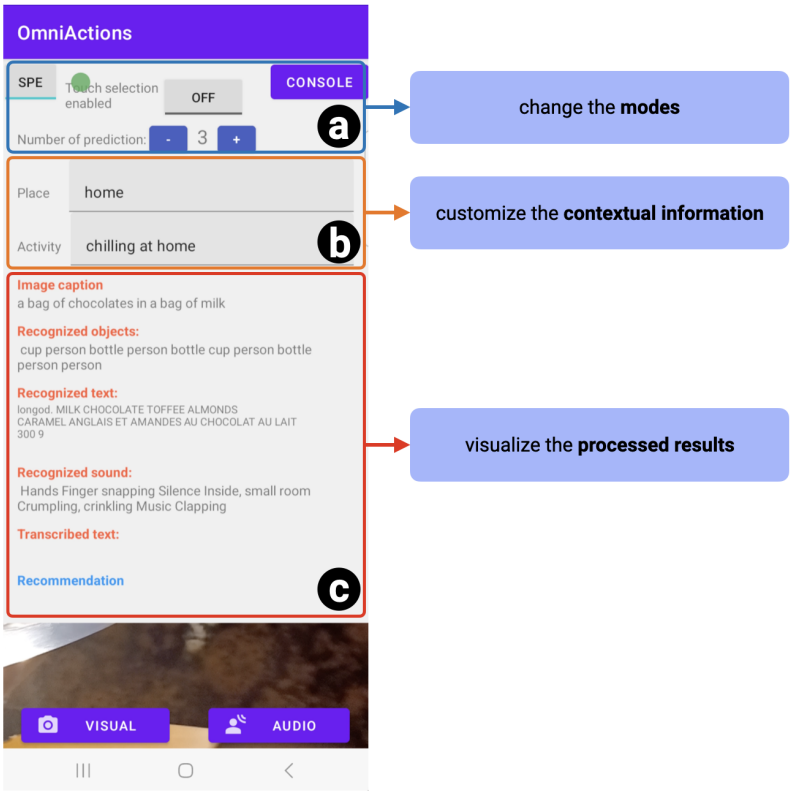

Fig. 10. In the console menu, a user could predict specific or general actions, enable or disable touch selections, and (a) adjust the number of predictions, (b) specify the contextual information, and (c) view the processed data.

Participants were then asked to complete six predefined tasks. During the tasks, participants were asked to verbalize their thoughts as they used the prototype system to complete the tasks. Lastly, participants used the system to complete free-form tasks where they defined their own goal. After experiencing the prototype system, participants completed a questionnaire containing Likert-based usability questions, as well as open-ended questions designed to gather qualitative feedback. It took 30-40 minutes to complete the process of the study. We recorded audio during the study for transcription and analysis.

7.3.2 Results. All participants successfully completed the predefined tasks without any assistance. On a scale of 1 to 7, participants gave an average rating of 4.8 (σ=1.30) for ease of use, an average rating of 5.6 (σ=1.34) for fondness towards the system, and an average rating of 5.8 (σ=1.64) for the system's potential and promise. All the participants commented on the potential of using OmniActions for their everyday tasks in the future. P3 stated, "*having this might fundamentally change the interaction of future AR interfaces*". The OmniActions was positively received due to (1) its ability to reduce friction by predicting the actions (P1, P2, P4); P4 especially praised, "the prediction was awesome!" during his free-form tasks; (2) the comprehensive overview of available actions (P3); (3) the straightforward interface with sufficient functionalities.

Participants, did however, note areas for improvement on the understandability for the wording of actions, "*I thought Save-to-list is saving something important to me while Save-for-reference is something that is not important*" (P3). P2 stated "*As a developer, I see the value of distinction between each actions which help me implement the functions ... but as an end-user, I find it confusing to differentiate between them and understand specific purposes*". P2 also mentioned that "*trying to understand the difference between two suggested similar actions may also increase my cognitive load*". Participants suggested adding content-aware examples to each action to help end-users understand the outcome.

Participants also noted that having too many options to select may cause high cognitive load. In the case of incorrect prediction and expanding more actions, participants found it a bit overwhelming to read through potential actions (P1, P5). To address this challenge, participants suggested using hierarchical sub-menus (P1, P3, P5) or having fewer options while treating some actions as add-ons (P2).

Overall, participants were enthusiastic about OmniActions and saw its value for for end-users and developers, and also provided several useful suggestions that could guide future developments.

8 DISCUSSION

In this section, we discuss the limitations of our research including the study design and data collection, as well as the design of the interactive system. We also provide insights on various future directions to address the limitations and to extend this line of research.

8.1 Specifics of Follow-up Actions

Follow-up actions can be classified into hierarchical categories: general, specific, and parameter-based. Parameters further categorize specific actions. For example, parameters include *"what language"* to translate or *"what social media platform"* to share on. Higher-level prediction may achieve better model performance (as shown in Table 2), while lower-level predictions could offer users a more effortless experience. Additional information should be considered to improve the model performance such as the context and location (which we collected in our diary study) or individual history and preference. This paper focuses on the first two levels (general and specific) as the first step towards understanding the interaction and building prototypes for demonstration while leaving the parameter space as future work.

8.2 Additional Data Collection

In future work, we'll collect additional data to overcome several limitations of our current diary study dataset, including:

(1) **A few categories of follow-up actions have small amount of samples in our dataset**: For example, in our dataset, there were only 7 entries with the follow-up action of translation. But in a specific context, such as travling in a foreign country, translation will be a frequently used action.
(2) **Lack of speech data**: Speech data contains semantic information that can be very informative in deciding follow-up actions. For example, if a user just asked their partner about the song that was playing and did not receive a good answer, they would be likely to take action to recognize the music. Additionally, the content within speech may also be the target information that a user would take actions on. We did not collect speech data due to the anonymity requirements of our study protocol.
(3) **The limitation of mobile phone captured data**: Our participants in the diary study used a mobile phone to capture the multimodal data they act on. If we could also collect other information, such as eye gaze, we could understand users' intention better, thus increasing the accuracy of predictions.

8.3 Aggregating Actions to Predict Goals

One pattern we observed within the collected data was the aggregation of multiple follow-up actions to achieve goals, which was further confirmed in our evaluation study. For example, to take actions using background music, participants may first recognize the music and then search for the name of the song online and then potentially save it to a future playlist. Furthermore, such actions may be taken at different time frames, ranging from immediate actions to those executed at a later time. For example, participants may save a poster for future reference and search for more information about it online when they get home.

Such follow-up actions rely on the understanding of users' goals and more thorough contextual information. Goals can also range from general objectives such as "I want to stay connected with my friends" to specific tasks such as "I want to do some research about this product". The former may involve actions such as sharing a photo with friends to remember funny moments between them, while the latter may involve actions such as recognizing a product using Google Lens, searching for a brand name online, or comparing prices with other similar products. Future interactive systems should encompass the ability to suggest executable items by understanding or predicting users' goals at different levels of abstraction. These systems should also support parameterization of more specific follow-up actions (*e.g.*, share with friend, but I get to choose which friend). These insights could be incorporated into future systems to better predict and support user actions based on goals, contexts, times, and personal objectives, ultimately leading to more effective and personalized user experiences.

8.4 Increasing the Proactivity of Predictions

Our diary study enabled users to upload information they deemed important regardless of whether they were able to capture it as a photo or video or not (i.e., by providing a text description for an intermittent sound). To create a future AI system that is proactive, it is essential to investigate information that users may not initially recognize as important but may prove to be useful at a later time by proactively suggesting follow-up actions. One future direction could be to incorporate a lifelogging system that users wear throughout the day (*e.g.*, RayBan Stories[9]) that captures information intermittently. Users could later reflect on the collected data, identify important information they missed, and label potential actions related to it. Another approach could be to introduce a third party, as we did in the workshop, to provide an objective opinion on the types of actions that could be taken using the embedded information. This would allow for a more comprehensive understanding of the information and its potential uses.

9 CONCLUSION

In this paper, we present OmniActions to predict follow-up actions when users encounter multimodal information. To inform the design of the interactive system, we conducted a five-day diary study with 39 participants, aiming to gain a comprehensive understanding of the design space of follow-up actions. Through the study, we identified 7 general categories (i.e., *share*, *save*, *remind*, *look up*, *digital extract*, *media manipulation*, and *complex actions*) and 17 specific follow-up action categories.

We developed the OmniActions system to predict follow-up actions on multimodal information powered by an LLM. The system harnesses the reasoning capabilities of LLMs by introducing intermediate reasoning steps (chain-of-thoughts). We evaluated three different methods of utilizing LLMs, and the results indicate that integrating chain-of-thoughts significantly improves system performance. Specifically, the model attains a 90.7% accuracy rate when predicting three

[9]https://www.ray-ban.com/usa

general actions via few-shot prompting with chain-of-thoughts in a full-match metric. The system also features an interactive prototype developed for user interaction. We then conducted a user study to assess the prototype's usability, identify limitations, and gather feedback for future improvements. The findings demonstrated the potential of OmniActions and provided valuable insights into possible enhancements for the system.

(8039 words)

REFERENCES

[1] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, et al. 2022. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691* (2022).

[2] Antti Ajanki, Mark Billinghurst, Hannes Gamper, Toni Järvenpää, Melih Kandemir, Samuel Kaski, Markus Koskela, Mikko Kurimo, Jorma Laaksonen, Kai Puolamäki, et al. 2011. An augmented reality interface to contextual information. *Virtual reality* 15, 2 (2011), 161–173.

[3] Antti Ajanki, Mark Billinghurst, Toni Järvenpää, Melih Kandemir, Samuel Kaski, Markus Koskela, Mikko Kurimo, Jorma Laaksonen, Kai Puolamäki, Teemu Ruokolainen, et al. 2010. Contextual information access with augmented reality. In *2010 IEEE International Workshop on Machine Learning for Signal Processing*. IEEE, 95–100.

[4] Daniel L Ashbrook. 2010. *Enabling mobile microinteractions*. Georgia Institute of Technology.

[5] Linas Baltrunas, Karen Church, Alexandros Karatzoglou, and Nuria Oliver. 2015. Frappe: Understanding the usage and perception of mobile app recommendations in-the-wild. *arXiv preprint arXiv:1505.03014* (2015).

[6] Mohammad Ubaidullah Bokhari and Faraz Hasan. 2013. Multimodal information retrieval: Challenges and future trends. *International Journal of Computer Applications* 74, 14 (2013).

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[8] Robin Burke. 2007. Hybrid web recommender systems. *The adaptive web: methods and strategies of web personalization* (2007), 377–408.

[9] Wolfgang Büschel, Annett Mitschick, and Raimund Dachselt. 2018. Here and now: Reality-based information retrieval: Perspective paper. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*. 171–180.

[10] Juan Pablo Carrascal and Karen Church. 2015. An in-situ study of mobile app & mobile search interactions. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 2739–2748.

[11] Guanling Chen and David Kotz. 2000. A survey of context-aware mobile computing research. (2000).

[12] Xiang'Anthony' Chen, Jeff Burke, Ruofei Du, Matthew K Hong, Jennifer Jacobs, Philippe Laban, Dingzeyu Li, Nanyun Peng, Karl DD Willis, Chien-Sheng Wu, et al. 2023. Next Steps for Human-Centered Generative AI: A Technical Perspective. *arXiv preprint arXiv:2306.15774* (2023).

[13] Karen Church, Mauro Cherubini, and Nuria Oliver. 2014. A large-scale study of daily information needs captured in situ. *ACM Transactions on Computer-Human Interaction (TOCHI)* 21, 2 (2014), 1–46.

[14] Karen Church and Barry Smyth. 2009. Understanding the intent behind mobile information needs. In *Proceedings of the 14th international conference on Intelligent user interfaces*. 247–256.

[15] Hai Dang, Sven Goller, Florian Lehmann, and Daniel Buschek. 2023. Choice Over Control: How Users Write with Large Language Models using Diegetic and Non-Diegetic Prompting. *arXiv preprint arXiv:2303.03199* (2023).

[16] Mustafa Doga Dogan, Faraz Faruqi, Andrew Day Churchill, Kenneth Friedman, Leon Cheng, Sriram Subramanian, and Stefanie Mueller. 2020. G-ID: identifying 3D prints using slicing parameters. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[17] Mustafa Doga Dogan, Ahmad Taka, Michael Lu, Yunyi Zhu, Akshat Kumar, Aakar Gupta, and Stefanie Mueller. 2022. InfraredTags: Embedding Invisible AR Markers and Barcodes Using Low-Cost, Infrared-Based 3D Printing and Imaging Tools. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–12.

[18] Lydia Dubourg, Ana Rita Silva, Christophe Fitamen, Chris JA Moulin, and Céline Souchay. 2016. SenseCam: A new tool for memory rehabilitation? *Revue Neurologique* 172, 12 (2016), 735–747.

[19] Denzil Ferreira, Jorge Goncalves, Vassilis Kostakos, Louise Barkhuus, and Anind K Dey. 2014. Contextual experience sampling of mobile application micro-usage. In *Proceedings of the 16th international conference on Human-computer interaction with mobile devices & services*. 91–100.

[20] Sergio Garrido-Jurado, Rafael Muñoz-Salinas, Francisco José Madrid-Cuevas, and Manuel Jesús Marín-Jiménez. 2014. Automatic generation and detection of highly reliable fiducial markers under occlusion. *Pattern Recognition* 47, 6 (2014), 2280–2292.

[21] Ross Girshick. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 1440–1448.

[22] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18995–19012.

[23] Cathal Gurrin, Alan F Smeaton, Aiden R Doherty, et al. 2014. Lifelogging: Personal big data. *Foundations and Trends® in information retrieval* 8, 1 (2014), 1–125.

[24] Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2023. Evaluating Large Language Models in Generating Synthetic HCI Research Data: a Case Study. In *ACM SIGCHI Annual Conference on Human Factors in Computing Systems*. ACM.

[25] Morgan Harvey and Matthew Pointon. 2017. Searching on the go: the effects of fragmented attention on mobile web search tasks. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 155–164.

[26] Morgan Harvey and Matthew Pointon. 2019. Understanding in-context interaction: An investigation into on-the-go mobile search. *Information Processing & Management* 56, 6 (2019), 102089.

[27] Jie Huang and Kevin Chen-Chuan Chang. 2022. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403* (2022).

[28] Mina Huh, Yi-Hao Peng, and Amy Pavel. 2023. GenAssist: Making Image Generation Accessible. *arXiv preprint arXiv:2307.07589* (2023).

[29] Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-Writing with Opinionated Language Models Affects Users' Views. *arXiv preprint arXiv:2302.00560* (2023).

[30] Eunkyung Jo, Daniel A Epstein, Hyunhoon Jung, and Young-Ho Kim. 2023. Understanding the Benefits and Challenges of Deploying Conversational AI Leveraging Large Language Models for Public Health Intervention. (2023).

[31] Tanya R Jonker, Ruta Desai, Kevin Carlberg, James Hillis, Sean Keller, and Hrvoje Benko. 2020. The Role of AI in Mixed and Augmented Reality Interactions. In *CHI2020 ai4hci Workshop Proceedings. ACM*.

[32] Vladimir Kirilyuk, Xiuxiu Yuan, Peggy Chi, Alex Olwal, Ruofei Du, et al. 2023. Visual Captions: Augmenting Verbal Communication with On-the-fly Visuals. (2023).

[33] Khalil Klouche, Tuukka Ruotsalo, Diogo Cabral, Salvatore Andolina, Andrea Bellucci, and Giulio Jacucci. 2015. Designing for exploratory search on touch devices. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. 4189–4198.

[34] Christoph Kofler, Martha Larson, and Alan Hanjalic. 2016. User intent in multimedia search: a survey of the state of the art and future challenges. *ACM Computing Surveys (CSUR)* 49, 2 (2016), 1–37.

[35] Amel Ksibi, Ala Saleh D Alluhaidan, Amina Salhi, and Sahar A El-Rahman. 2021. Overview of lifelogging: current challenges and advances. *IEEE Access* 9 (2021), 62630–62641.

[36] Ju Yeon Lee, Ju Young Kim, Seung Ju You, You Soo Kim, Hye Yeon Koo, Jeong Hyun Kim, Sohye Kim, Jung Ha Park, Jong Soo Han, Siye Kil, et al. 2019. Development and usability of a life-logging behavior monitoring application for obese patients. *Journal of Obesity & Metabolic Syndrome* 28, 3 (2019), 194.

[37] Dingzeyu Li, Avinash S Nair, Shree K Nayar, and Changxi Zheng. 2017. Aircode: Unobtrusive physical tags for digital fabrication. In *Proceedings of the 30th annual ACM symposium on user interface software and technology*. 449–460.

[38] Jiahao Li, Alexis Samoylov, Jeeeun Kim, and Xiang'Anthony' Chen. 2022. Roman: Making Everyday Objects Robotically Manipulable with 3D-Printable Add-on Mechanisms. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–17.

[39] Nianlong Li, Han-Jong Kim, LuYao Shen, Feng Tian, Teng Han, Xing-Dong Yang, and Tek-Jin Nam. 2020. HapLinkage: Prototyping haptic proxies for virtual hand tools using linkage mechanism. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 1261–1274.

[40] David Lindlbauer, Anna Maria Feit, and Otmar Hilliges. 2019. Context-aware online adaptation of mixed reality interfaces. In *Proceedings of the 32nd annual ACM symposium on user interface software and technology*. 147–160.

[41] Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. 2023. Evaluating the logical reasoning ability of chatgpt and gpt-4. *arXiv preprint arXiv:2304.03439* (2023).

[42] Xingyu" Bruce" Liu, Vladimir Kirilyuk, Xiuxiu Yuan, Alex Olwal, Peggy Chi, Xiang" Anthony" Chen, and Ruofei Du. 2023. Visual Captions: Augmenting Verbal Communication With On-the-Fly Visuals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–20.

[43] Zhaoyang Lv, Edward Miller, Jeff Meissner, Luis Pesqueira, Chris Sweeney, Jing Dong, Lingni Ma, Pratik Patel, Pierre Moulon, Kiran Somasundaram, Omkar Parkhi, Yuyang Zou, Nikhil Raina, Steve Saarinen, Yusuf M Mansour, Po-Kang Huang, Zijian Wang, Anton Troynikov, Raul Mur Artal, Daniel DeTone, Daniel Barnes, Elizabeth Argall, Andrey Lobanovskiy, David Jaeyun Kim, Philippe Bouttefroy, Julian Straub, Jakob Julian Engel, Prince Gupta, Mingfei Yan, Renzo De Nardi, and Richard Newcombe. 2022. Aria Pilot Dataset. https://about.facebook.com/realitylabs/projectaria/datasets.

[44] Jakob Nielsen. 1994. *Usability engineering*. Morgan Kaufmann.

[45] Joon Sung Park, Joseph C O'Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442* (2023).

[46] Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2022. Social Simulacra: Creating Populated Prototypes for Social Computing Systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–18.

[47] Savvas Petridis, Nicholas Diakopoulos, Kevin Crowston, Mark Hansen, Keren Henderson, Stan Jastrzebski, Jefrey V Nickerson, and Lydia B Chilton. 2023. AngleKindling: Supporting Journalistic Angle Ideation with Large Language Models. (2023).

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

[49] Robert W Reeder, Adrienne Porter Felt, Sunny Consolvo, Nathan Malkin, Christopher Thompson, and Serge Egelman. 2018. An experience sampling study of user reactions to browser warnings in the field. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–13.

[50] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* 28 (2015).

[51] Luca Rossetto, Ivan Giangreco, and Heiko Schuldt. 2015. OSVC-Open Short Video Collection 1.0. *Technical Report* CS-2015-002 (2015).

[52] Luca Rossetto, Heiko Schuldt, George Awad, and Asad A Butt. 2019. V3C–a research video collection. In *MultiMedia Modeling: 25th International Conference, MMM 2019, Thessaloniki, Greece, January 8–11, 2019, Proceedings, Part I 25*. Springer, 349–360.

[53] Rohit Saluja, Ayush Maheshwari, Ganesh Ramakrishnan, Parag Chaudhuri, and Mark Carman. 2019. Ocr on-the-go: Robust end-to-end systems for reading license plates & street signs. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 154–159.

[54] Bill Schilit, Norman Adams, and Roy Want. 1994. Context-aware computing applications. In *1994 first workshop on mobile computing systems and applications*. IEEE, 85–90.

[55] Bill N Schilit and Marvin M Theimer. 1994. Disseminating active map information to mobile hosts. *IEEE network* 8, 5 (1994), 22–32.

[56] Abigail J Sellen, Andrew Fogg, Mike Aitken, Steve Hodges, Carsten Rother, and Ken Wood. 2007. Do life-logging technologies support memory for the past? An experimental study using SenseCam. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 81–90.

[57] Timothy Sohn, Kevin A Li, William G Griswold, and James D Hollan. 2008. A diary study of mobile information needs. In *Proceedings of the sigchi conference on human factors in computing systems*. 433–442.

[58] Bryan Wang, Gang Li, and Yang Li. 2022. Enabling Conversational Interaction with Mobile UI using Large Language Models. *arXiv preprint arXiv:2209.08655* (2022).

[59] Sitong Wang, Savvas Petridis, Taeahn Kwon, Xiaojuan Ma, and Lydia B Chilton. 2021. PopBlends: Strategies for Conceptual Blending with Large Language Models. *arXiv preprint arXiv:2111.04920* (2021).

[60] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems* 35 (2022), 24824–24837.

[61] Andy Zeng, Adrian Wong, Stefan Welker, Krzysztof Choromanski, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, et al. 2022. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598* (2022).

[62] Fangneng Zhan and Shijian Lu. 2019. Esir: End-to-end scene text recognition via iterative image rectification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2059–2068.

[63] Yinglong Zhang, Rob Capra, and Yuan Li. 2020. An in-situ study of information needs in design-related creative projects. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*. 113–123.

[64] Yuhui Zhang, Hao Ding, Zeren Shui, Yifei Ma, James Zou, Anoop Deoras, and Hao Wang. 2021. Language models as recommender systems: Evaluations and limitations. (2021).

[65] Zhengzhe Zhu, Ziyi Liu, Tianyi Wang, Youyou Zhang, Xun Qian, Pashin Farsak Raja, Ana Villanueva, and Karthik Ramani. 2022. MechARspace: An Authoring System Enabling Bidirectional Binding of Augmented Reality with Toys in Real-time. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–16.

A PROMPTS TEMPLATE

A.1 Prompts for Generating Chain-of-Thoughts

{"role": "system", "content":
"You are an assistant that produces chain-of-thoughts analysis leading to reasons about why users take specific follow-up actions from a third-person perspective. You should operate under the assumption that the goal is not known to you.
Follow-up actions: Share on social media: Share/upload on social platforms
Share with others: Send the info to specific entities
Remember: Cherish a specific experience/moment for later recall
For reference: Store information for later usage or consultation
To list: Add information to a designated, organized collection
Keep track: Record the development of a task or goal
Remind: Make an alert or notice to remember something later
Search online: Search for more information online related to specific goals
Recognize: Indentify the information using specific tools (e.g., song names)
Translate: Translate text/speech from one language to another
Extract and access: Extract and utilize information from sources
Transcribe: Convert audio to text
Digitalize: Transform information to a digital format for easier access
Compare: Compare similarity and difference between two sets of info
Calculate: Perform mathematical operations to solve a problem/task
Edit media: Enhance images or sounds to improve overall experience
Augment: Modify media files to accomplish a specific task
Output in a list of JSON dicts, where applicable: "chain-of-thoughts", "prediction" (the follow-up actions)" }

A.2 Few-shot Prompts for Predicting Follow-up Actions

Predicting **specific** follow-up actions:

{"role": "system", "content":
"You are an assistant that predicts the follow-up actions users will take based on multimodal information input using chain-of-thoughts analysis. Provide up to [NUM_OF_PREDICTION] most likely follow-up actions from the following options (with definition):
Follow-up actions: Share on social media: Share/upload on social platforms
Share with others: Send the info to specific entities
Remember: Cherish a specific experience/moment for later recall
For reference: Store information for later usage or consultation
To list: Add information to a designated, organized collection
Keep track: Record the development of a task or goal
Remind: Make an alert or notice to remember something later
Search online: Search for more information online related to specific goals
Recognize: Indentify the information using specific tools (e.g., song names)
Translate: Translate text/speech from one language to another
Extract and access: Extract and utilize information from sources
Transcribe: Convert audio to text

Digitalize: Transform information to a digital format for easier access
Compare: Compare similarity and difference between two sets of info
Calculate: Perform mathematical operations to solve a problem/task
Edit media: Enhance images or sounds to improve overall experience
Augment: Modify media files to accomplish a specific task
Output in a list of JSON dicts, where applicable: "chain-of-thoughts", "prediction" (the follow-up actions)"
},
{ “role”: “user”, “content”: “<example 1>” },
{ “role”: “assistant”, “content”: “<result 1>” },
{ “role”: “user”, “content”: “<example 2>” }
{ “role”: “assistant”, “content”: “<result 2>” }

Predicting **general** follow-up actions:

{“role”: “system”, “content”:
“You are an assistant that predicts the follow-up actions users will take based on multimodal information input using chain-of-thoughts analysis. Provide up to [NUM_OF_PREDICTION] most likely follow-up actions from the following options (with definition):
(general)
Share
(specific)
Share on social media: Share/upload on social platforms
Share with others: Send the info to specific entities

(general)
Save
(specific)
Remember: Cherish a specific experience/moment for later recall
For reference: Store information for later usage or consultation
To list: Add information to a designated, organized collection
Keep track: Record the development of a task or goal

(general)
Remind
(specific)
Remind: Make an alert or notice to remember something later

(general)
Look up
(specific)
Search online: Search for more information online related to specific goals
Recognize: Indentify the information using specific tools (e.g., song names)
Translate: Translate text/speech from one language to another

(general)
Digital extract

(specific)
Extract and access: Extract and utilize information from sources
Transcribe: Convert audio to text
Digitalize: Transform information to a digital format for easier access

(general)
Complex
(specific)
Compare: Compare similarity and difference between two sets of info
Calculate: Perform mathematical operations to solve a problem/task

(general)
Augment
(specific)
Edit media: Enhance images or sounds to improve overall experience
Augment visual/audio: Modify media files to accomplish a specific task
Output the prediction result in a list of JSON dicts (the length will be the number of prediction), where applicable: "chain_of_thoughts", "prediction"
Output the general category" },
{ "role": "user", "content": "<example 1>" },
{ "role": "assistant", "content": "<result 1>" },
{ "role": "user", "content": "<example 2>" }
{ "role": "assistant", "content": "<result 2>" }

B GENERATING DESIGN SPACE

B.1 Formative study

Figure 3 shows the screenshot of the FigJam board where the participants collaborate with each other. We picked the examples shown in Figure 11 to represent the categories identified in the formative workshop.

B.2 Diary study

The detailed survey questions are listed in Table 3.

(a) Examples picked from the workshop.

#	Type	Action
1	🖼	Share this fun moment on Instagram.
2	🎧	Send it to a garage maintenance person to diagnose the problem.
3	🖼	Keep record of my son's painting progress.
4	🖼	Save this receipt and send for reimbursement later.
5	🖼	Search the keyboard on google lens to buy one.
6	🖼	Search "APPLE FRITTERS" online.
7	🎧	Transcribe the professor's lecture. Save as note for later to review.
8	📹	Zoom in to see the seal more clearly.
9	🖼	Translate some of the dishes to my native language.
10	🖼	Set a reminder for next week's homework due.

#2 is an abnormal clicking sound of a garage door opening.
#7 is an audio of professor talking about the history of the United States.

(b) Actions corresponding to each example.

Fig. 11. Examples covering actions that were shown to participants before the diary mission began.

Table 3. Survey questions that participants were required to answer for each diary entry.

#	Target: **Visual**	Target: **Audio**	Question type
Q1	Upload your photo or a screenshot of your video.	(For video only) Upload a screenshot of your video (audio as the main target).	[File upload]
Q2	Briefly describe the photo. *e.g., "This is a billboard of the movie Dunkirk showing when it will be in theater."*	Briefly describe the audio you captured AND wanted to take follow-up actions with. *e.g., "This is the background music I heard in the cafe."*	[Open-ended]
Q3	Where were you when you captured the data?		[Open-ended]
Q4	What were you doing when you captured the data?		[Open-ended]
Q5	Please list the physical objects visible in the data.	What types of sounds could be heard in the recording? *- Speech / Music / Tools / Environmental noise / ...* *- Others [Force answer]*	[Multi-type]
Q6	What best describes the information you intended to take action on? *- The whole scene / environment / place* *- Objects in the photo/video* *- Text visible in the photo/video* *- Others [Force answer]*	Please choose the audio information you want to take action on: *- [Same as in Q5]*	[Multiple choice]
Q7	In 1-3 sentences, explain what actions you plan to take on the information in the data you shared. *For example: "Save the date to my calendar." If you have multiple actions, please list them all.*		[Open-ended]
Q8	From the list below, which best characterizes your previous response. Select all that apply. *- [Categories from the workshop* *- Others [Force answer]*		[Multiple choice]
Q9	In 1-3 sentences, briefly explain: *(i)* the overall goal(s) of taking the above actions. *(ii)* the reason(s) why you want to take the above actions when you captured the photo/video.		[Open-ended]

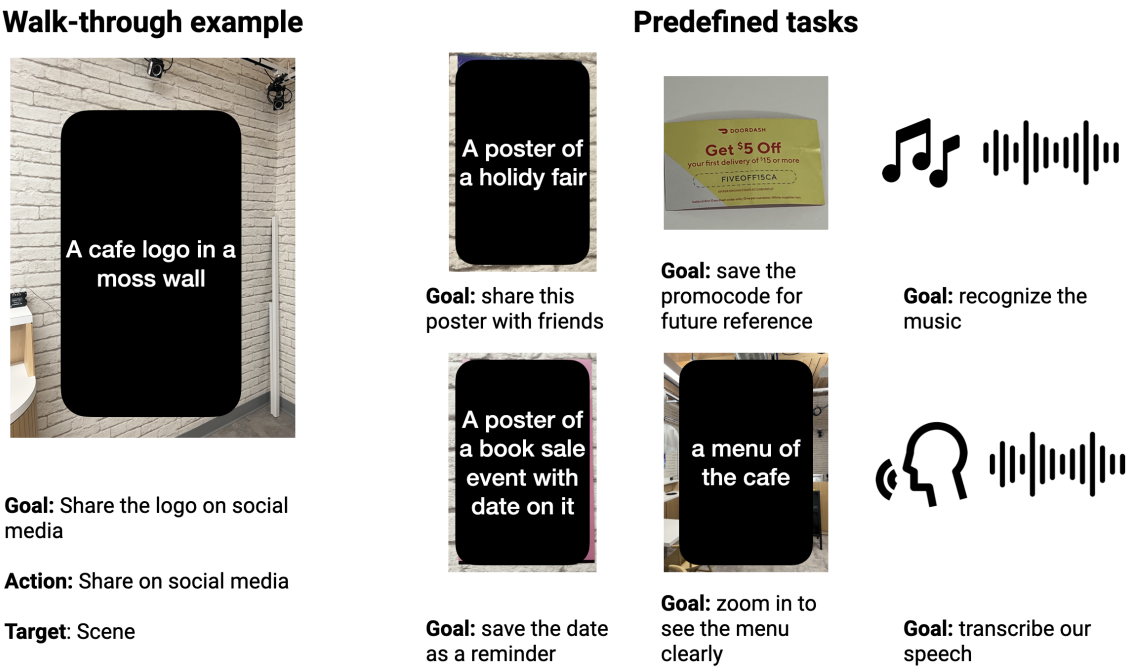

Fig. 12. The six predefined tasks that participants completed during the evaluation study (anonymized due to confidentiality).