OmniActions: Predicting Digital Actions in Response to Real-World Multimodal Sensory Inputs with LLMs

Jiahao Nick Li
Reality Labs Research, Meta & UCLA
Toronto, Canada
ljhnick@g.ucla.edu

Yan Xu
Reality Labs Research, Meta
Redmond, United States
yanx@meta.com

Tovi Grossman
University of Toronto
Toronto, Canada
tovi@dgp.toronto.edu

Stephanie Santosa
Reality Labs Research, Meta
Toronto, Canada
ssantosa@meta.com

Michelle Li
Reality Labs Research, Meta
Toronto, Canada
michelleli@meta.com

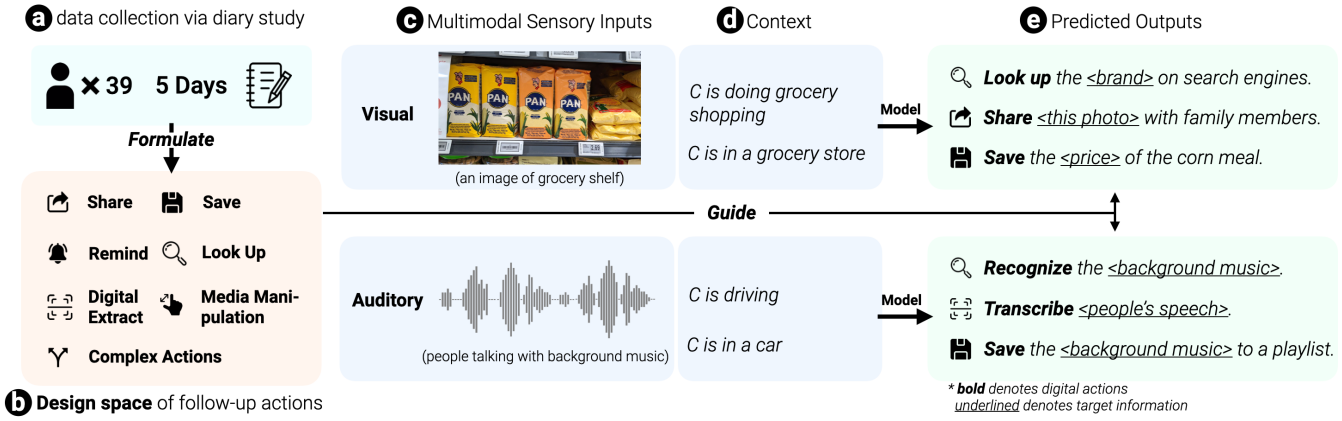


Figure 1: OmniActions contributes: (1) a design space of digital follow-up actions (b) derived from data collected during a five-day diary study with 39 participants (a), and (2) a pipeline that takes multimodal sensory data (c) and contextual information (d) as inputs, and predicts what digital actions users might take and on which specific information in the input they might take these actions (e). The action prediction is guided by the design space.

ABSTRACT

The progression to "Pervasive Augmented Reality" envisions easy access to multimodal information continuously. However, in many everyday scenarios, users are occupied physically, cognitively or socially. This may increase the friction to act upon the multimodal information that users encounter in the world. To reduce such friction, future interactive interfaces should intelligently provide quick access to digital actions based on users' context. To explore the range of possible digital actions, we conducted a diary study that required participants to capture and share the media that they intended to perform actions on (e.g., images or audio), along with their desired actions and other contextual information. Using this data, we generated a holistic design space of digital *follow-up* actions that could be performed in response to different types of multimodal sensory inputs. We then designed OmniActions, a pipeline powered by large language models (LLMs) that processes multimodal sensory inputs and predicts follow-up actions on the target information grounded in the derived design space. Using the empirical data collected in the diary study, we performed quantitative evaluations on three variations of LLM techniques (intent classification, in-context learning and finetuning) and identified the most effective technique for our task. Additionally, as an instantiation of the pipeline, we developed an interactive prototype and reported preliminary user feedback about how people perceive and react to the action predictions and its errors.

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642068

CCS CONCEPTS

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

KEYWORDS

digital follow-up actions, predictive interface, large language models, dataset, pervasive augmented reality, diary study

ACM Reference Format:
Jiahao Nick Li, Yan Xu, Tovi Grossman, Stephanie Santosa, and Michelle Li. 2024. OmniActions: Predicting Digital Actions in Response to Real-World Multimodal Sensory Inputs with LLMs. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 22 pages. https://doi.org/10.1145/3613904.3642068

1 INTRODUCTION

The progression towards "Pervasive Augmented Reality (AR)" envisions easy access to information in different modalities such as text, images, or audio, anytime and anywhere [25]. However, in many everyday scenarios within the real world, users are occupied physically, cognitively or socially, which may limit the use of typical AR inputs such as hand gestures and speech. This can present significant friction in interacting further with the information they encounter in the world. For example, a driver noticing a movie billboard faces increased friction in (1) identifying the movie's name from the billboard and (2) searching for more details about the movie, due to the cognitive and physical demands of driving. This motivates in the need for future interfaces to intelligently reduce friction in interacting with information [4].

Interactions with real-world information generally involve two steps: (1) retrieving desired information (*e.g.*, select the text on the billboard) and (2) performing corresponding *follow-up actions* (*e.g.*, searching for more details on Google). We envision that future interfaces should be designed to simultaneously process multimodal sensory inputs, analogous to *human sensory perception*, and proactively suggest follow-up actions on the target information. This vision represents a more *generalized* approach than existing approaches like iOS' text-in-a-photo action suggestions[1], Google Lens[2], or Shazam's song recognition[3], which recognize one specific modality of sensory inputs (*e.g.*, structured text, images, or audio) and map it to hard coded predefined actions (*e.g.*, detecting an address and launching a navigation app). However, to implement this more generalized vision, two main limitations need to be addressed: *(i)* existing systems cannot predict follow-up actions on aggregated data from multiple modalities and *(ii)* there is a limited understanding of the range of actions users intend to perform during real-world scenarios when using multiple modalities. The latter is crucial for guiding the design of such systems, as it ensures that their output is grounded in a known action space, thus enabling the actions to be executable by the system.

Prior work has explored the design space of mobile and in-situ information needs [13, 17], *i.e.*, *when* and *how* users need *what* types of information. However, there is a limited understanding of the *action needs* users have in-situ. To bridge this gap, we ran a formative workshop followed by a diary study to collect and identify the actions people might take when interacting with multimodal information. In contrast to collecting and reflecting on already captured data in smartphones, the diary study prompted participants to *actively* capture fresh data immediately, *i.e.*, the actions they intended to take whenever they encountered new multimodal information. This approach mirrored the way users interact with information in AR settings, simulating an "always-on" audio-visual sensor. The collected data (*i.e.*, visual inputs such as scenes, physical objects, texts, and auditory inputs such as acoustic sounds, human speech) were then documented as images or text descriptions for further analysis. Over the course of five days, 39 participants contributed 382 data entries. The collected data was then used to inform the creation of a design space of possible follow-up actions that should serve as a blueprint for the design of possible follow-up actions that future interactive systems could incorporate (Figure 2e).

The design space was then used to inform the design of a prototype called OmniActions, containing a pipeline which enables the simultaneous processing of multimodal sensory inputs and subsequent generation of follow-up action predictions on target information (Figure 2f). Powered by a large language model (LLM), OmniActions *(1)* converts multimodal sensory inputs into structured text via existing models (*e.g.*, visual language models for image captioning) and then *(2)* leverages the explicit reasoning of the LLM [29] on the structured text to *(3)* predict target information (*e.g.*, the visible text) and follow-up actions (*e.g.*, share with another person) grounded in the design space (Figure 2g). To demonstrate the effectiveness of our pipeline and explore the LLMs' capabilities to support such real-world tasks, we conducted an evaluation using the empirical data collected from the diary study and compared multiple techniques of using LLMs. We employed three variants of using LLMs: conventional intent classification, in-context learning with Chain-of-Thoughts (CoT) prompting, and fine-tuning with CoT prompting. The results show that our approach yields competitive performance. For instance, in-context learning with CoT prompting using the latest LLM (*i.e.*, `GPT-4`) achieved a high accuracy (94.3%) when predicting the top three possible general actions. As an instantiation of the pipeline, we also developed an interactive smartphone prototype for user interaction. We conducted an in-lab feedback session with 5 participants to collect initial subjective feedback about the system and insights to improve the design and user experiences with the interactive prototype.

The contributions of this research are thus:

- A design space of follow-up actions that can be performed in response to multimodal sensory inputs. This design space was derived from the diary study data and surfaced 7 general and 17 specific categories of follow-up actions.
- A novel pipeline, OmniActions, that provides generalized predictions of follow-up actions for real-world multimodal sensory inputs. OmniActions leverages the explicit reasoning of LLMs (CoT) on structured text converted from multimodal data to ground the predicted actions in the design space.
- An evaluation of the approach enabled by empirical data collected from the diary study using different techniques (i.e., in-context learning and fine-tuning). The results showed competitive performance of the proposed approach. Additionally, the evaluation provided insights into LLMs' capabilities to support real-world tasks.
- An interactive smartphone prototype that predicted users' target information and suggested follow-up actions. User feedback highlighted the system's potential and the design space's comprehensiveness.

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

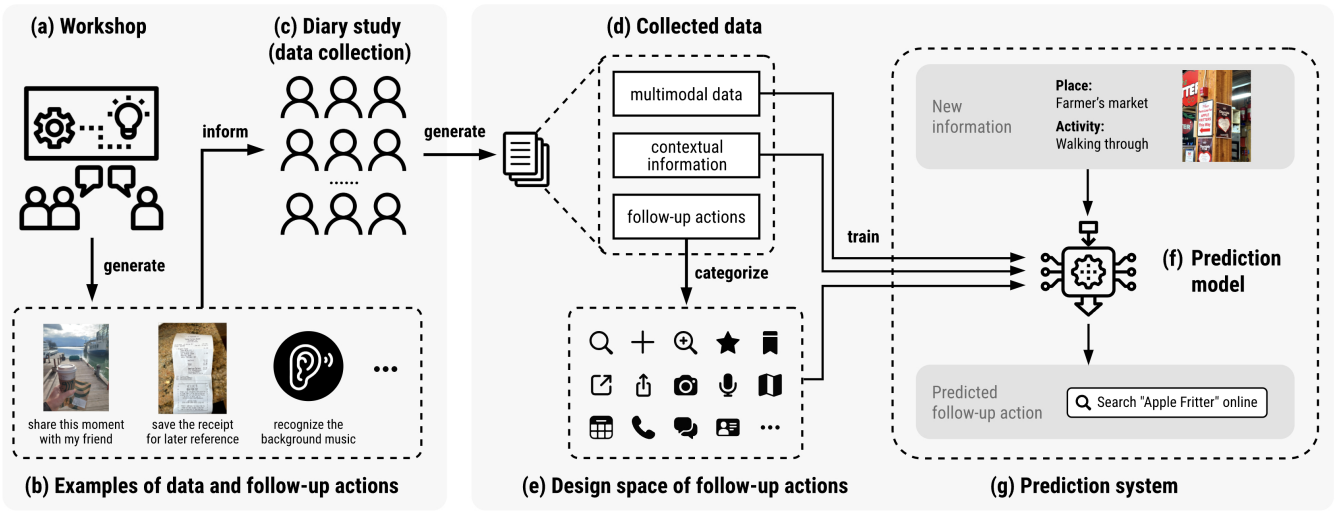


Figure 2: The development process for OmniActions. (a) An internal workshop was conducted to (b) generate informative examples of situations when users may take using multimodal information. (c) The examples were used to inform and inspire the participants during a diary study that (d) collected data 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 to train a prediction model that was (g) integrated within OmniActions to predict multiple follow-up actions given multimodal information.

2 RELATED WORK

The present research was inspired by prior work on users' mobile information needs, multimodal information interaction techniques, and the use of large language models to augment interaction.

2.1 Mobile Information Needs

Information needs, defined as "*any information that is required for a task, or to satisfy the curiosity of the mind, regardless of whether the need is satisfied or not*" [17], is closely related to how users interact with real-world information. Researchers have conducted various diary studies [3, 9, 11, 13, 14, 17, 28, 35, 59] to understand users' information needs under different contexts, including while using mobile phones [11, 14, 28], seeking information within a social network [17] or being on-the-go [9, 59]. While this presents similar use cases as what we expect to encounter in pervasive AR systems, existing research majorly focuses on *what* types of *information* users require, and *how* their *contexts* affect their needs. However, there is a notable gap in understanding the next phase of addressing *actions needs*: *what* types of *actions* users might take once their information needs have been met. Perhaps most related is prior work by Church *et al.*, which explored the types of searches (*i.e.*, informational, geographical, or personal information) associated with different contexts [14]. The scope of these follow-up actions, however, was limited to searching for target information, rather than to a broader range of actions that could be performed with the information. To bridge this gap, OmniActions aims to understand what *actions* users might take once they have access to the information they need. We envision the potential for OmniActions to enable rich contextual understanding in future AR scenarios, therefore, we focus specifically on the real-world information that can be perceived by the sensors on an AR device when using different modalities.

2.2 Multimodal Information-Based Interaction Techniques

To predict follow-up actions while encountering new information in the wild (e.g., music, noise, visible text, objects, etc.), it is crucial that systems are able to understand the context of one's environment and the information that is available to users. One way to obtain such an understanding is to directly retrieve information that is embedded in barcodes, fiducial markers [22], human faces [2], or objects during fabrication processes [19, 20, 38]. Researchers have also explored retrieving "raw" information such as visible text [54, 66], physical objects [23, 51], multimodal scenes [65], human speech (e.g., Google API[4]), and music (e.g., Shazam). Nevertheless, to understand users' intent based on the information in their physical environments, multimodal information must be monitored and processed in a way that a system can make predictions using it.

Lifelogging digitally tracks a person's daily experiences and is one way to process multimodal information [26, 36]. Prior work has used lifelogging to enhance human memory by retrieving moments through natural language [21, 57] or monitor one's health by analyzing logged data [37]. However, lifelogging does not specifically focus on predicting a user's intent and to predict follow-up actions, which requires the categorization of the design space. Several lifelogging datasets have been collected, including the Aria dataset [42], Ego4D [24], and other video datasets [52, 53]. These datasets could be used to investigate desired follow-up actions, but they contain redundant data when such actions are not required. To specifically

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

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

2.3 Large Language Models in HCI

Artificial Intelligence (AI) has been widely used in the Human-Computer Interaction (HCI) community, with LLMs experiencing a surge of usage in recent years [1, 16, 27, 31–33, 48, 49, 60–62]. LLMs' abilities to understand common knowledge and reason within a given context have been leveraged for interactive code support [60], social computing [47, 48] and accessibility support [30]. For example, Visual Caption employed a fine-tuned language model to predict user intent during visual inquiries using the last two sentences [41], while SayCan extracted and leveraged knowledge priors within LLMs to reason about, and execute, robot commands [1]. LLMs have also been used to enhance recommender systems that utilize contextual information to recommend items [7, 34]. For example, GPT-3 [6] was used to augment movie recommendation systems [67].

Most of this prior work relied on the capture of one's explicit intent [10], wherein users or agents interacted with a system via direct prompts. OmniActions unlocks a new interaction method with LLMs by embracing a more implicit intent, focused on the user's current visual input (e.g., multimodal information such as environmental understanding or recognized text) and contextual information. Coupled with the Chain-of-Thoughts prompting, this enables OmniActions to deliver explainable predictions of target information and follow-up actions.

3 FORMATIVE WORKSHOP

We ran a formative workshop to obtain a preliminary understanding about the multimodal information triggers people came across in everyday life and their follow-up actions. The outcomes of the workshop were clusters of the actions participants took with multimodal information triggers. The learnings on the workshop informed our method choices, question design, and example generation for the next study to collect data from general population.

3.1 Procedure

We recruited 10 participants within our institution through group email invitations. The participants included HCI researchers, UX designers, and student interns, all of whom worked within the domain of AR and XR. Their expertise would provide insights on how people may interact with information in the physical world. The participants volunteered to join the workshop and they were not paid. The workshop consisted of three parts and lasted one hour in total. Participants were invited to use a FigJam[5] whiteboard for synchronous collaboration.

3.2 Process

The organizer of the workshop first introduced the goal and agenda of the workshop to the participants. Then they shared two examples with the participants, including a parking ticket and an audio file of some background music, and their related context and follow up actions. During Part 1, participants were asked to share their own media, context, and follow-up actions. During Part 2, participants reflected on other participants' media and came up with their follow-up actions. In Part 3, participants collaboratively clustered similar actions (Figure 3).

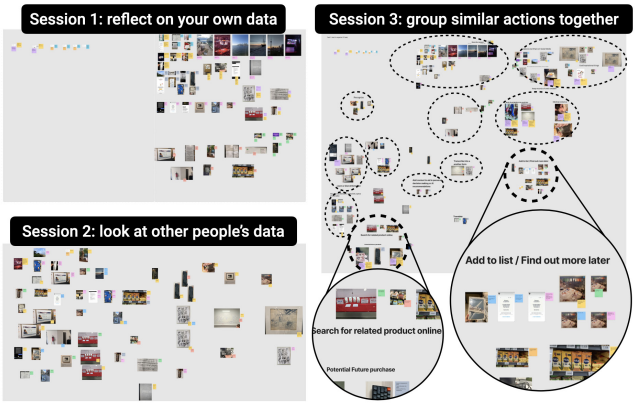


Figure 3: Screenshots from the formative workshop where participants shared data in Session 1, reviewed other participants' data in Session 2, and grouped similar actions in Session 3.

3.2.1 Part One. "Browse past media, share those that you did follow-up actions with". During this part, participants had 20 minutes to browse their personal media storage and upload the ones that they took actions with to shared Google drive and the FigJam board. For each shared media item, participants were asked to recall the record the following information: *(i)* what target they acted on (*e.g.*, the menu of a boba shop), *(ii)* what action they took (*e.g.*, save to the album for future reference) and *(iii)* contextual information such as the location or their activity, which is useful in the next part. For audio and video uploads, the participants described them textually on the FigJam board. 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 at the scene, what actions you would take on the information?" In this part, we aimed to get third-person perspective on what the possible actions could be given the media. Contextual information from part one helps other participants to imagine the scenarios. Participants had 20 minutes to browse examples shared by other participants and to type their imagined follow-up actions for the target information. An additional 104 follow-up actions were proposed, with 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." In Part 3, participants had 15 minutes to collaboratively cluster and label all 170 examples from Part 2, using an affinity diagram.

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

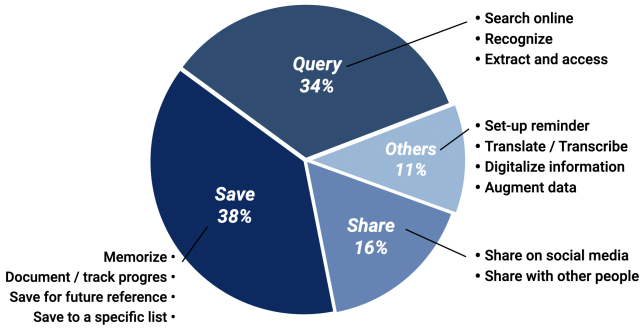


Figure 4: Frequencies of the 13 follow-up actions generated during the workshop (n = 170) that were grouped into 4 categories.

3.3 Results

After the workshop, two researchers coded, filtered, and clustered the 170 follow-up actions independently. The participant-generated clusters were also referenced in this process. This process was inductive, meaning that they coded actions mentioned by the participants, rather than starting with an existing set of actions. The results from each researcher and the clusters from participants were compared. The researchers discussed and resolved the discrepancies in the clusters' boundary, naming, and granularity. As a result, they identified 13 types of actions that were grouped into four categories (i.e., share, save, query, and others; Figure 4). Representative examples from these categories were used as learning materials for participants in our subsequent diary 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 diary study, where we asked participants to share the textual description of their audio rather than the audio itself. We present more details in the next section.

4 DATA COLLECTION VIA A DIARY STUDY

While the workshop provided an initial glimpse of the type of multimodal information and follow-up actions users would desire, we wanted to formalize the findings with in-situ experiences from participants external to our institution. The use of a diary study methodology would enable participants to log data whenever needs arose [59], making it an ideal choice to examine desired follow-up actions when one encounters new information. We leveraged this methodology to answer the following research question:

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

We adopted the *snipped-based diary technique* proposed by Brandt *et al.* [5] to collect data about users' follow-up actions with multimodal information. As opposed to reflecting on captured data (*e.g.*, images in the album) at a fixed time of day, our participants were asked to log data whenever they encountered information in the world they wished to take action upon. This simulates the "always-on" feature of an AR platform where users can interact with AR interfaces anytime and anywhere. We collected the data including *(i)* the target information they wished to take action on, *(ii)* the desired follow-up actions and *(iii)* contextual information such as their goals, locations, and activities. Contextual information was important to collect as it could affect the choice of follow-up actions [8, 39, 55]. 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 contextual information would increase a system's ability to accurately understand users' goals and follow-up actions. We incorporated this information into a predictive model later on in our research process.

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 participants were between the ages of 18 to 69 years old, were proficient in English, and had a smartphone to take photos. Each participant was compensated $50 USD after they completed the diary study for their time.

4.2 Procedure

The diary study consisted of two phases, i.e., an introductory phase and a diary phase. During the introductory phase, participants were shown examples from the workshop that represented several of the categories of media and actions that the workshop participants identified. Note that to avoid bias due to the categorization that resulted from the workshop, participants were only shown the exemplar media and follow-up actions.

During the diary phase, participants were instructed to submit 2 entries each day for five days. These entries needed to reflect genuine participant needs that occurred in the moment. The diary phase began in the middle of the week and extended over the weekend to capture the different types of needs that may occur throughout a week. For each diary entry, participants were required to answer questions about their entry (Figure 5). These included:

Media Containing the Information (Q1, Q2). Although we aimed 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 requirements of our institution. Therefore, if participants wanted to share audio or video, they were asked to provide a text description of the data or screenshots of the videos instead (*e.g.*, "This is the background music I heard in the cafe").

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"* [56]. 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

[6] https://dscout.com/

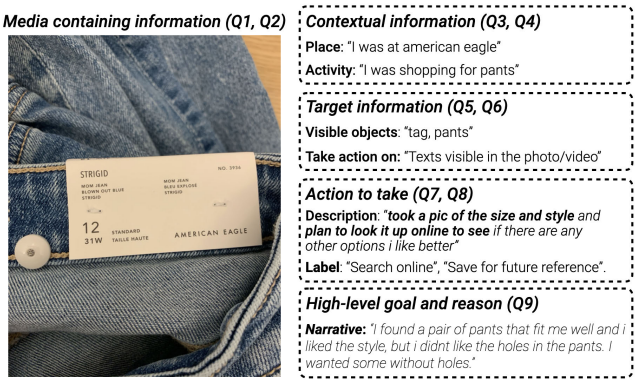


Figure 5: An example diary entry from the diary study.

which information the participant wanted to perform follow-up actions for. For example, participants could be interested in only the text visible in an image or the entire scene. 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 potential user interactions 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 then categorize these actions. Additionally, they could select categories corresponding to these actions using the action categories identified in the workshop. Participants also had the option to create new categories by selecting 'other' if there were actions that did not fit within the existing categories. Note that we minimized potential bias by asking participants to detail their intention and desired actions in their own words on a first page before being shown and asked to choose from the action types on the next page. 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 for doing so.

4.3 Data Summary

During the study, two participants did not finish the number of required data entries (one only submitted 7 and the other only 5) and they were compensated $5 per submitted entry. This resulted in 382 data entries in total. The ratio of collected visual to audio data was approximately 2:1. We collected 254 visual data examples (i.e., 193 photos and 61 videos with visual selected as the target information type) and 128 audio data examples (i.e., 48 videos with audio as the target information type and 80 text descriptions of audio). Participants reported wanting to take action on 55 full scenes (40 photos / 15 videos), 120 individual objects (96 photos / 24 vidoes), 79 pieces of text (57 photos / 22 videos), 51 speech clips (20 videos / 31 audio only), and 77 sound clips (28 videos / 49 audio only). Additionally, participants shared 17 (i.e., 10 visual, 7 audio) follow-up actions which did not fit within any of the categories identified during the workshop.

4.3.1 Contexts of the Captured Data. We coded and summarized the contexts when people came across multimodal information that they intended to take follow-up actions based on survey answers in Q5 and Q6. Figure 6 shows the diversity of location and contextual activities people had. We consider our dataset to be representative to a day in the life, based on the comparison to the American Time Use Survey (ATUS, from U.S. Bureau of Labor Statistics) [44]. The diversity of the contextual activities included all activity categories mentioned in the 2022 ATUS survey [44] except "sleeping" (not applicable to our study), "caring for non-household members", or "organizational, civic, and religious activities". The latter two categories together accounted for 0.5 hours per day per person on average. Most (77%) of the in-situ capture about people's follow-up actions needed had other contextual activities, out of which 24% were low-demanding activities and 39% were high-demanding activities that require full body motion or high cognitive focus, and 13% involved both types of contextual activities. This showed the pervasiveness of multitasking situations where people's physical and cognitive bandwidth were already used for other activities. Therefore, it was important to reduce the friction for people to use follow-up actions.

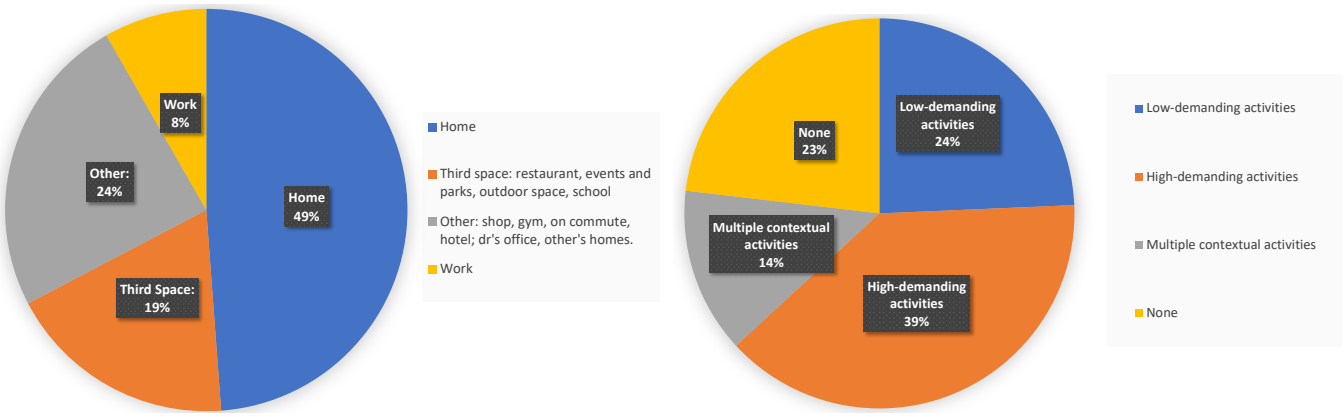
5 DESIGN SPACE OF FOLLOW-UP ACTIONS

Following the diary study, a researcher and research assistant collaboratively reviewed the diary entries, coded the data, and compared and consolidated the codes through iterations. The resulting action space consisted of 7 **general categories** of follow-up actions, including *share*, *save*, *remind*, *look up*, *digital extract*, *media manipulation*, and *complex actions*. These categories were further divided into 17 **specific categories** (Figure 7).

For the general categories, (1) *Share* refers to actions that users employ 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); (2) *Save* refers to the actions used to store information; (3) *Remind* refers to actions that created an alert or notice to remember something later such as setting a reminder after seeing a flight schedule on a screen or noting oneself of the date of a specific event (particularly useful for managing tasks, appointments, or important events); (4) *Look up* refers to actions that searched for specific information or details; (5) *Digital extract* refers to actions taken to obtain and utilize information from multiple sources; (6) *Media manipulation* refers to actions that altered or modified media content to achieve a specific outcome, and (7) *Complex actions* involve processing data from multiple sources. Figure 7 lists the definition of the 17 specific categories; please refer to Appendix C.2 for more detailed explanation.

5.1 Analysis of Diary Data Using the Design Space

We conducted a post-study analysis on the diary study data using the categories within the design space (Figure 8). The *share* (45.9%), *save* (47.4%), and *look up* (32.1%) actions were most common general



(a) The location distribution of our dataset.

(b) The activity distribution of our dataset.

Figure 6: In (a), third space refers to the places outside of home or work where people have the potential opportunity to socialize and engage with the community [45]. In (b), the low-demanding activities include: Sedentary leisure activities (i.e. watching TV, browsing social media, browsing news, drawing, reading), Eating/drinking, Waiting, Sedentary housework (i.e. checking emails, online payments, online shopping, personal care); The high-demanding activities include: Interacting with someone, Physical housework (i.e. cleaning, cooking, organizing, maintaining, getting mails, gardening), Full-body movement activities (i.e. walking, working out, playing), Focused activities (i.e. driving, studying, working), Shopping in a store, Preparing with time pressure, Exploring and navigating environment.

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."

Figure 7: Design space of follow-up actions for multimodal information that emphasizes general and specific categories of actions.

actions while the remainder of the actions (i.e., *remind* (4.5%), *media manipulate* (2.8%), *digital extract* (12.3%), *complex actions* (2.1%)) were less common (Figure 8a). Figure 8b shows the frequencies of each specific action. Within the data, we also observed that participants tend to take multiple actions in succession. For example, participants *remembered* a memorable moment and then *shared* it with family members. Specifically, 183 diary entries had only one action, 147 had two actions, 44 had three actions and 8 had four actions. An example with four aggregated specific actions is illustrated in Appendix D.

Participants also used different patterns of follow-up actions when interacting with data from different modalities (Figure 8c). The overall frequency of specific follow-up actions when the target was visual versus audio were similar, although there appears to be a

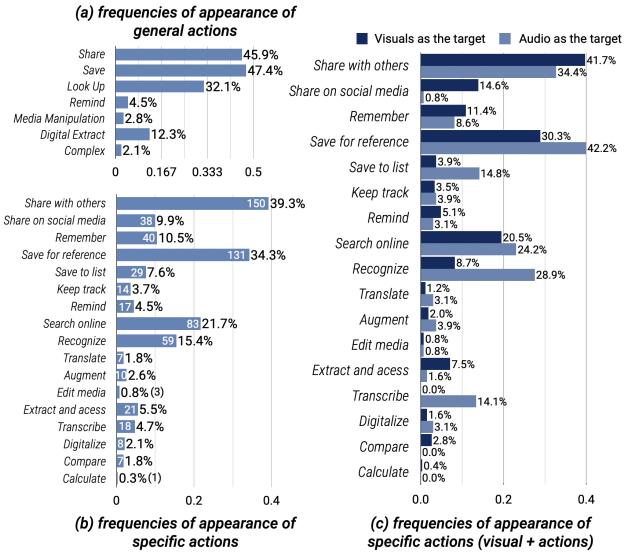


Figure 8: (a) The frequencies of the general actions. (b) The frequencies of the specific actions (with number). (c) The frequencies of the specific actions on visual and audio. Frequency was computed as the number of appearances divided by the total number of diary entries.

difference when *sharing on social media*, *saving to a list*, *recognizing* and *transcribing*. These variations align with typical real-world interactions. For example, people often share visual content (*e.g.*, a breathtaking landscape or an unusual statue) on social media, while it is less common to post specific sounds (*e.g.*, an abnormal noise) in an environment. Additionally, as described earlier, *transcribing* is exclusive to audio. Furthermore, our data showed a trend where individuals *recognized* and *saved* music to their playlists upon hearing a song they enjoyed. This reflected how people interact with, and respond to, real-world audio data, which also leads to a higher frequency of *saving for reference* actions in similar scenarios.

6 OMNIACTIONS PIPELINE

To reduce users' frictions to access follow-up actions triggered by the multimodal information in the world, we create OmniActions. The pipeline of OmniActions senses and processes different multimodal information, and predicts the *target information* and *follow-up actions* grounded in the action space, which is based on the empirical data. Moreover, by reasoning with multimodal and contextual information, this pipeline aims to enhance explainability and model performance.

To achieve this, OmniActions consists of three steps (Figure 9):

(1) OmniActions converts raw multimodal data (i.e., visual and audio data) into structured text by leveraging existing models.
(2) OmniActions then performs intermediate explicit reasoning on the structured text via Chain-of-Thoughts (CoT) prompting. The training data for this prompting was grounded in the data from the diary study.
(3) Finally, OmniActions predicts the *target information* (i.e., the whole scene, physical objects, text, sounds, or speech) and the *follow-up actions* grounded in the design space using a large language model (LLM).

6.1 Converting Multimodal Data into Structured Text

For a model to process information in multiple modalities simultaneously and perform predictions, it is essential to convert the multimodal data into a unified representation format (*e.g.*, a textual representative or a joint embedding space). This would enable a model to identify and learn from patterns in the multimodal input. To enable explicit reasoning for prediction, OmniActions converted multimodal data into a textual representation. Specifically, OmniActions leveraged existing models to convert both visual and audio data into structured text before performing CoT prompting-based reasoning steps. All the converted data for each entry was stored in JSON format for explicit reasoning. Note that our pipeline aims to demonstrate potential using currently available data and could be adapted to broader range of modalities in the future.

6.1.1 Visual Information. Aligning with the findings from our diary study, OmniActions supports three aspects of visual information: the overall scene, physical objects, and any visible text. For the overall scene, OmniActions leverages recent advancements in multimodal learning frameworks that have shown competitive performance in describing a scene with text. In this implementation, we used an open-source, state-of-the-art image captioning model, InstructBLIP [15], with the prompt of *"Write a short description for the image."*. For the physical objects and visible text, OmniActions used the Detectron2 object detection model [64] to detect the objects and Google Cloud Vision[7]) to recognize the text.

6.1.2 Audio Information. OmniActions classified the type of acoustic sounds via YAMNet[8] and used speech-to-text models to transcribe human speech. As our institution would not permit the collection of personal identifiable information, we were unable to collect human speech data during the diary study. As a result, the evaluation of our model's capabilities does not incorporate transcribed speech.

6.1.3 Explicit Contextual Information. As context affects the actions people perform using the information they have available to them, OmniActions leveraged the data collected during the diary study, i.e., where participants were and what were they doing when encountering the information. However, such contextual information may not always be available in practice, and thus this is optional to include in our pipeline. We examined the impact of the contextual information on the prediction performance in Sec. 7.4.

6.2 Generating Chain-of-Thoughts Prompts

Traditional classification methods typically rely on trained models like black boxes. To enhance explainability, a model should explain the rationale behind its predictions for certain follow-up actions. Ideally, this reasoning should be as close to a user's reasoning as

[7]https://cloud.google.com/vision/docs/ocr
[8]https://github.com/tensorflow/models/tree/master/research/ audioset/yamnet

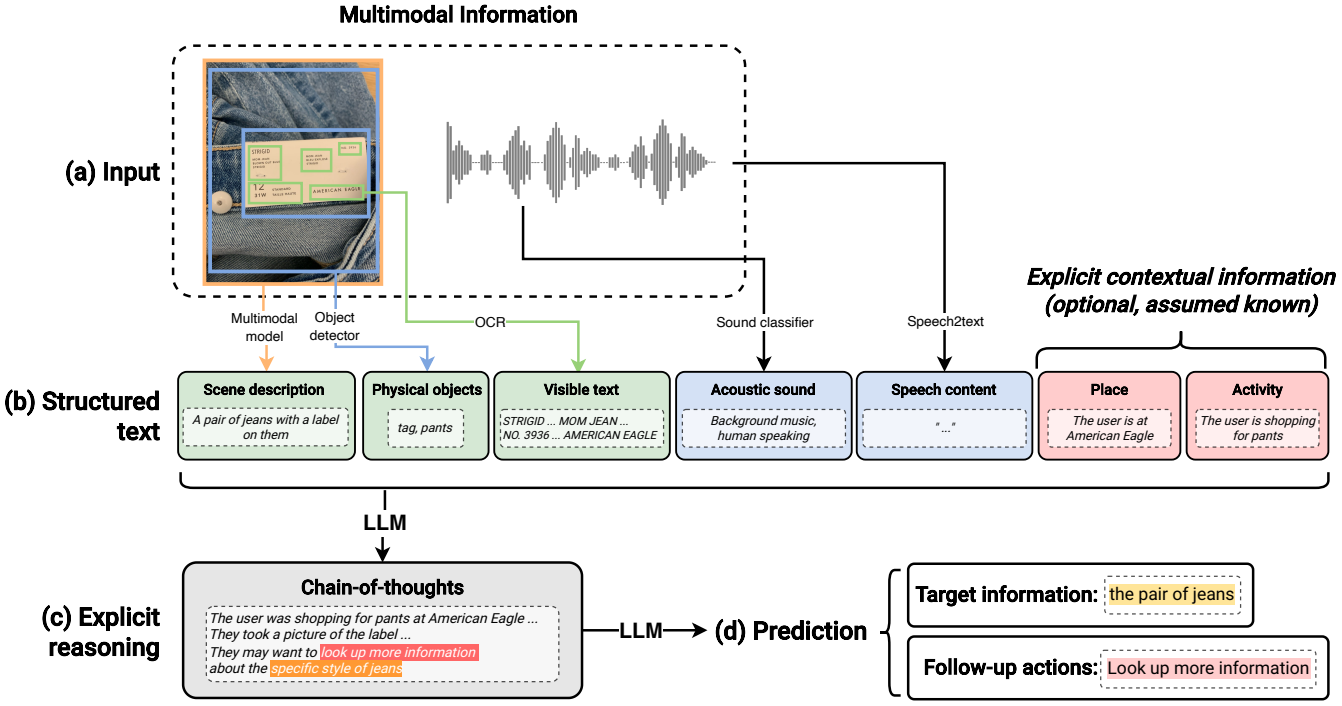


Figure 9: OmniActions processes multimodal information (a) by converting it into structured text using existing models (b). It processes visual data using multimodal models, object detectors, and OCR models and processes audio data via sound classifiers and speech-to-text models. Then, OmniActions performs an explicit reasoning using Chain-of-Thoughts prompting (c) and predicts target information and follow-up actions (d).

possible. This is especially important when there are multiple actionable information items captured and the user's intention is not clear from the sensor data itself. For example, in Figure 5, the person captured an image with multiple texts (including the brand name, the jean's name and the size etc.), but the user only intends to search more information about the specific jean's sizes, rather than the brand name. Such reasoning could be instrumental for subsequent interactions, such as deciding which target information to search. OmniActions addresses this by introducing CoT prompting [63] as an intermediate reasoning step through the prompting and training process (Figure 9c).

One of the challenges is the generation of CoT prompts. Prior work mostly leveraged zero-shot prompting (i.e., using prompts such as *"let's think step-by-step"*) or researcher-crafted prompts for in-context learning. However, these approaches rely on either common sense reasoning or researcher reasoning, which may not represent how our participants reasoned within their context.

To address this, we leveraged the data collected during the diary study to generate CoT prompts in empricial data. During 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. We convert these reasoning from first-person perspective to third-person perspective for the CoT prompts. In the above example, the participant shared their reasoning in the survey (Figure 5):

> *"I found a pair of pants that fit me well and I liked the style, but I didn't 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."*

The above data were used to generate the CoT reasoning as follows:

> *"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 the LLM to generate the CoT prompts for the model as the ground truth label for each data point collected during the diary study. Specifically, the prompt consisted of the list of actions with the respective description (Figure 7) ground truth action label and the participants' responses for their goals and reasons. The template used to generate the CoT prompts is in Appendix A.1.

7 IDENTIFY THE BEST PERFORMANT LLM TECHNIQUE

7.1 LLM Techniques and Implementation

With the OmniActions pipeline, we aim to predict the intended action on multimodal information. Recent LLMs' advancements has shown various techniques' competitiveness for new tasks, such as

in-context learning and fine-tuning. To identify the best-performant among the state-of-the-art LLM techniques for OmniActions and draw insights in exploring LLMs' capabilities in addressing the target task, we use the empirical data collected from the diary study to evaluate the performance of the pipeline using different techniques.

Specifically, we employed three different LLM techniques to predict the intended actions: *(i)* intent classification, *(ii)* in-context learning with chain-of-thoughts prompting, and *(iii)* fine-tuning with chain-of-thoughts training data. We first discuss the rationale for choosing these methods and then explain them in detail.

7.1.1 Conventional Intent Classifier. Prior research in Natural Language Processing (NLP) has explored numerous methods of classifying text-based data for different tasks, including intent classification [40] or sentiment analysis [46]. One key advantage of this is the potential use of smaller models (*e.g.*, BERT [18] or LSTM [50]) for lower cost and faster execution.

To maintain consistency in our comparison, we fine-tuned a pre-trained LLM (`davinci` from OpenAI) to perform the intent classification. The `davinci` model has a smaller size compared to other GPT-3.5 models that support fine-tuning and it outputs `logprobs`, which provide confidence scores for different action predictions, enabling us to rank the top-n likely actions, similar to traditional classification models. As shown in Figure 10, to prepare the training data we formatted the structured text into a tuple as the input for each data entry and use the target label (i.e., target information or the action) as the output. We then used this data to fine-tune the LLM in the legacy prompt-completion[9] way. Specifically, we used 75% of the data entries from the diary study for training and the rest for evaluation.

7.1.2 In-Context Learning with CoT. In-context learning, also known as few-shot prompting, is a popular method for adapting LLMs to new tasks [6]. This technique provides a few examples illustrating the task, specifying both the input format and expected output, without changing the model's parameters (*i.e.,* gradients) for new tasks. This is the key benefit that it does not require a large amount of data for training, thus making it potentially more adaptable to new tasks.

To enhance the explainability of the prediction, we provided exemplar data to instruct the LLM to produce intermediate reasoning (CoT) prior to the final action prediction. We used both `GPT-3.5-turbo` and `GPT-4` as the model for the few-shot prompting method. As shown in Figure 10, besides the converted structured text as the input, we also provide task descriptions and several examples illustrating the exemplar input and output. Specifically, the *task description* defines the role of the system and leverages the definition of the predicted labels from the design space (*e.g.*, definition of specific actions in Figure 7). For the prediction of follow-up actions, Since our design space consists of 17 specific categories, we include 9 data entries which cover all the categories in the prompt, and the rest 373 data entries are used for evaluation. For detailed prompts, please refer to Appendix A.3.

[9]https://platform.openai.com/docs/guides/legacy-fine-tuning

7.1.3 Fine-Tuning with CoT. Different from in-context learning, fine-tuning an LLM would change the model's parameters to specialize it for the target task. This was accomplished by feeding additional training data into a pre-trained model, updating the model's gradients, *i.e.*, *fine-tuning*. The key benefit of this approach is that it enables the model to be exposed to a broader range of examples, and could thus potentially identify and learn more intricate patterns for better performance. However, the drawback is its reliance on a large amount of training data.

As shown in Figure 10, for each data entry, we provided the structured text and the task description as the input and used the generated CoT and target label as the output. We used 75% of the data entries for training and the rest for evaluation. As GPT-4 did not publicly support fine-tuning at the time of this paper's preparation[10], we used `GPT-3.5-turbo` for the fine-tuning approach.

7.2 Performance Evaluation - Accuracy

The two tasks: *(i)* predicting the target information and *(ii)* predicting the follow-up actions, were performed in parallel and thus we evaluate them separately.

7.2.1 Accuracy When Predicting Target Information. Target information prediction is a *multi-class classification* task, where the target modality was 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), the 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. Specifically, we implemented a *three-class classification* (scenes, objects, and text) for visual information, and a *two-class classification* (speech and sounds) for audio.

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

Approach	Visual	Audio
Intent classification	70.6	**92.3**
In-context learning (w/ CoT)	62.3	90.1
Fine-tuning (w/ CoT)	**70.7**	90.9

We measured the accuracy of the the three techniques. For *intent classification* and *finetuning*, we used 75% of the data entries for training and the remaining 25% for testing. For *in-context learning*, we used five data entries from the training set representing each target information modality as the few-shot examples and tested on the rest data (377 entries). The results showed that all the approaches could achieve competitive performance when predicting the target information (Table 1).

7.2.2 Accuracy When Predicting Follow-Up Actions. As users may perform multiple actions using the same information, the prediction of follow-up actions is a *multi-label classification* task, meaning

[10]as of December 11th, 2023

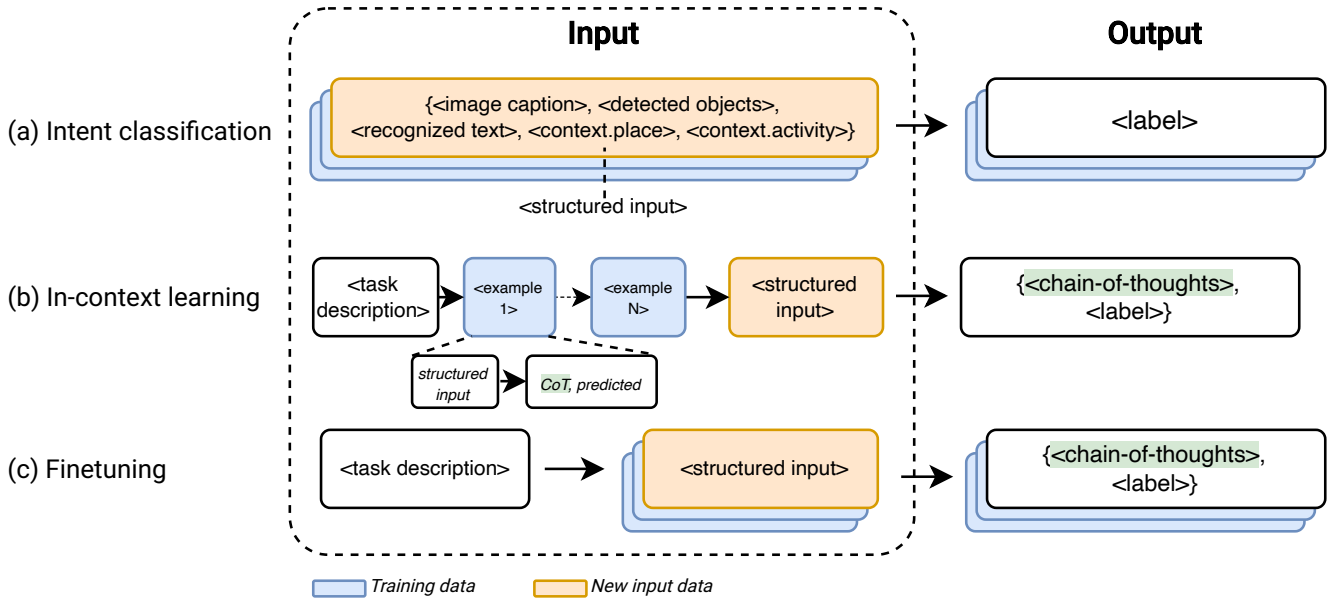


Figure 10: Data preparation and processing for each technique. *Intent classification* and *finetuning* used input-output pairs for training, while *in-context learning* required only a few task examples.

each data entry may contain multiple ground truth labels. Thus, we evaluated the model's accuracy when predicting the top-N most likely predictions (N = 1, 2, 3). It is important to note that, in the current setup, the accuracy of predicting the follow-up actions is not affected by the target information prediction as these two evaluations are conducted in parallel. We used the *full-match* metric to represent the accuracy of the prediction (i.e., the ratio of correct predictions to the minimum of ground truth labels or predictions), to demonstrate the alignment between the predictions and ground truth labels. The accuracy was calculated using a sample average:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\frac{C_i}{\min(G_i, P_i)} \quad (1)$$

where N was the total number of test data samples, C_i represented the number of correct predictions for the i-th data sample, G_i represented the number of ground truth labels for the i-th data sample, and P_i represented the number of predictions made for the i-th data sample.

Besides the three approaches, we also calculated the accuracy of a model when it always predicted the top-N most frequent actions as it might achieve high accuracy due to imbalanced distribution of the data. However, this does not make such a model *good*, as it will never be able to predict actions other than the most dominant ones. Please refer to Appendix Figure 1a for the confusion matrix of this approach.

Results. The results shown in Table 2 demonstrated that in-context learning with the latest LLM (GPT-4) outperformed all other approaches. Notably, it achieves very high accuracy on general actions when predicting the top three possibilities (94.3%) and marked an improvement of 11.6% on specific actions over the next best-performant approach: fine-tuning with GPT-3.5 (from 60.1% to 67.1%). Additionally, the results show that finetuning works better on specific actions (13.8% improvement) than on general actions (6.3% improvement) when predicting top-3 likely actions using the same model (GPT-3.5). This is likely due to the dominance of certain categories in general actions and data-driven approach like finetuning is more sensitive to the data distribution. For detailed data, please refer to Appendix Table 4.

7.3 Confusion Matrices of Predicting Follow-up Actions

Besides the overall prediction accuracy, it is also important to analyze the error – how does the model behave when predicting an incorrect label. We generated confusion matrices for the approaches to visualize the model behavior when predicting the top-3 actions (Figure 11). Specifically, we visualized the confusion matrices of the best-performant approach (*i.e., in-context learning using GPT-4*) in this section. Due to the imbalanced distribution of the data, we normalized the confusion matrix by the number of appearances of each label. For details on creating these matrices and matrices for other approaches, please refer to Appendix B. Note that since we only have one data entry for the *Calculate* action in the specific category and we have included that in the prompt, there is no data entry for this action in the evaluation set in this approach.

Results. The result shows a competitive performance using the in-context learning approach when sufficient examples are provided to cover the diversity of the actions. This highlights the importance of *data diversity* and the potentials for expanding the action space as interaction platforms and techniques evolve. Regarding the data distribution, even without explicit awareness of it, the model performs

Table 2: Overall accuracy (%) when predicting follow-up actions using the full-match metrics.

	Predicting General			Predicting Specific		
Approach / Num of Predictions	1	2	3	1	2	3
Top-n dominant categories	47.4	61.3	78.1	39.3	45.3	54.8
Intent classification	46.0	61.1	83.1	41.7	40.6	54.3
Finetuning (GPT-3.5)	57.7	67.2	84.9	**48.1**	50.2	60.1
In-context learning (GPT-3.5)	57.9	65.2	78.6	36.4	40.1	46.3
In-context learning (GPT-4)	**60.3**	**69.9**	**94.3**	44.4	**52.9**	**67.1**

*All approaches except *intent classification* adopt chain-of-thoughts.

*Top-3 general actions (in order): Save, Share, Look up. Top-3 specific actions: Share with others, Save for reference, Search online.

*In-context learning (GPT-4) is tested on 373 data entries.

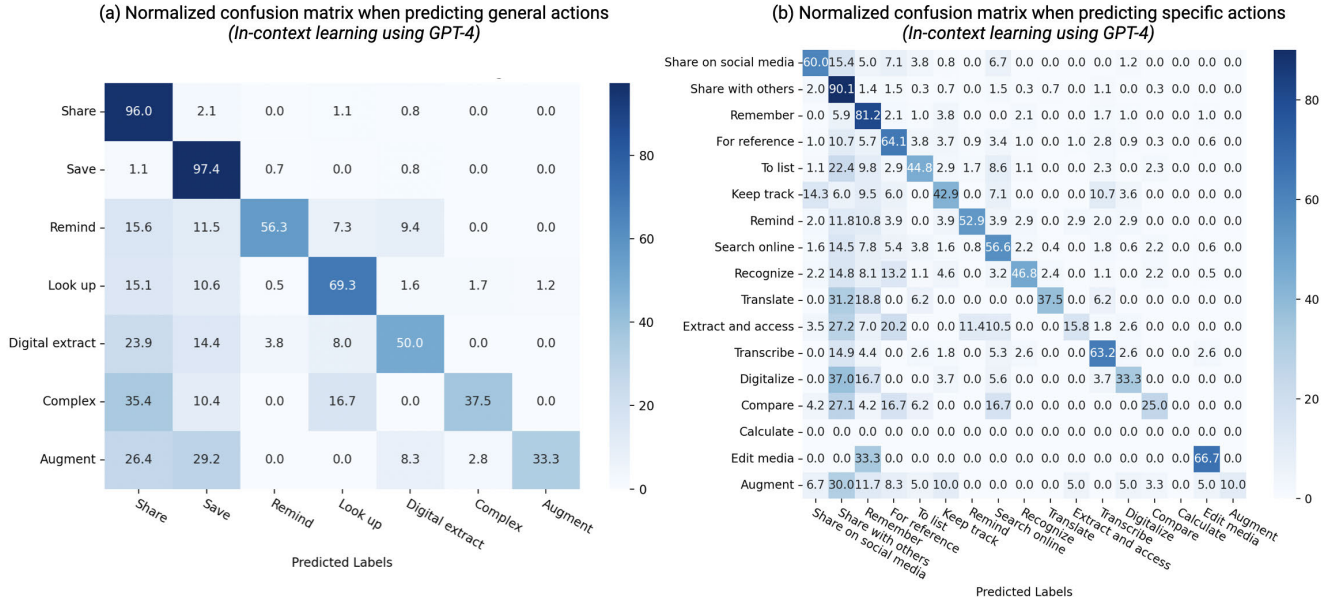


Figure 11: Confusion matrices using in-context learning (GPT-4) to predict the top-3 actions.

better on the dominant ones (*e.g.*, actions in the general *share* and *save* categories), while it performs worse on the less dominant ones (*e.g.*, specific actions like *extract and access* or *compare*). This shows an alignment between the data collected from the general users and the world knowledge that the model was trained on. To increase the model's performance on less dominant categories, soliciting more data for certain actions might be necessary. A future direction could involve collecting more high-quality data, which can be used to enrich the prompts for the in-context learning approach or employed for finetuning the model.

Table 3: Accuracy (%) for the in-context learning approach with and without explicit contextual information while predicting three specific actions.

	W/O Context	Location Only	Activity Only	Full Context
Audio only	47.5	47.7	59.7	**60.0**
Visual only	55.1	59.1	67.5	**70.8**
All data	52.5	55.2	64.9	**67.1**

7.4 Ablation to Understand Explicit Contextual Information and Modalities

The role of contextual information in the model's performance was another crucial aspect to consider. In our evaluation, we utilized data from the diary study assuming that the context was known, however, contextual information might not be readily available in practical scenarios. To understand its impact, we conducted an ablation test using the best-performing approach (*i.e.*, in-context learning with GPT-4), focusing on the two types of contextual information considered. We then computed the accuracy to assess the

impact (Table 3). Furthermore, we also examined how the model performs on visual and audio data separately to gain insights whether contextual information are important for certain modalities.

The result shows that the model's performance was improved by 27.8% when the full context was provided compared to when no context was provided. Within the contextual information, the activity information contributed more to the model's performance than the location information (23.6% improvement for activity and 5.1% for location), especially for audio data (25.7% improvement). Besides the contextual information, the result also shows that the model performs generally better on visual data than audio data (70.8% vs. 60.0%). This might be due to the richer content inherent in visual data, which contains more implicit contextual information. Recent research has shown multimodal models' capabilities in answering questions about the context from visual information ******citedai2023instructblip, thus future work may leverage such multimodal models to extract explicit contextual information before a prediction task.

8 A MOBILE PROOF-OF-CONCEPT PROTOTYPE WITH OMNIACTIONS SERVICE

To give an example about how OmniActions' pipeline serve applications, we developed an interactive prototype (*i.e.*, an Android app), which passes the multimodal input to OmniActions and then executes the predicted follow-up actions.

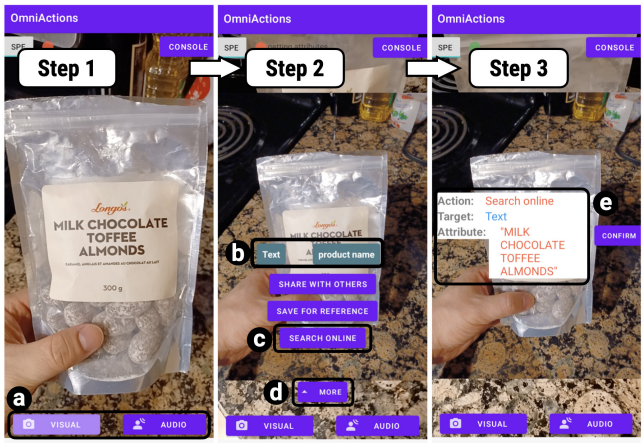


Figure 12: *OmniActions*'s user interface, wherein (a-e) a user could search for the product name on the bag of chocolate by selecting the follow-up actions suggested by the system.

8.1 Workflow

In this workflow, a user is searching for the product name of the chocolate online (Figure 12). First, the user clicks the visual or audio button to specify the modality of information they are interested in. As the user clicked the visual button (a), the system performs a *target modality* prediction and *follow-up action* prediction. The system then predicts the target as text (b) and recommends three actions. If the user finds that the suggested actions do not fit their needs, they can click the *more* button to see other actions in the design space (d). The user then selects the target attribute of the text ("product name") (b) and selects the *Search Online* action (c). After selection, a pop-up window visualizes the user's intent to search for the product name ("MILK CHOCOLATE TOFFEE ALMONDS") online (e). As the system does not currently detect all the context automatically, the user can manually specify a place and activity in the console (Figure **??**) for better prediction performance. Additionally, the user can toggle between predicting general actions and specific actions to view the raw results to increase explainability in the console view as well.

8.2 Implementation

The OmniActions prototype had two modules, a continuous detection module and a trigger-based detection module. The continuous detection module classified the sounds and transcribed speech (if present) in real-time and stored the classified sounds and speech transcription from the previous five seconds for further processing. The trigger-based module captioned the captured images to provide a description, detected objects within the captured images, and used OCR to identify and extract text in the images. Once a user triggered the system, OmniActions processed all the information into a tuple format so it could 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 using API level 33 and was 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.

8.3 Preliminary User Feedback

We used a think-aloud protocol [43] to understand how users perceive and use the prototype. Specifically, we are interested in people's reactions to the proactive interface and the prediction errors.

8.3.1 Setup and Method. Five participants with either programming or product development experience were recruited from a our institution to participate in the study. The participants volunteered to join the study and they were not paid. The study took place in a lab designed to resemble a cafe, which enabled everyday life scenarios such as viewing a menu and interacting with a book on a bookshelf. During the study, a researcher first walked the participants through the basic functionality by demonstrating an example. Participants were then asked to complete six predefined tasks and verbalize their thoughts while doing so. These include tasks such as "*save* the promocode on the gift card for future reference" or "*share* the menu in a cafe with your friends". Lastly, participants used the system to complete additional free-form tasks of users' choices (for as many times as they wanted), where they decided what follow-up actions they'd like to do. Using the *think-aloud* protocol, participants were asked to verbalize their intention on the actions they were taking and then used the system to complete the free-form tasks. After using the prototype system, participants

completed a questionnaire containing 7 point Likert-based usability questions, as well as open-ended questions designed to gather qualitative feedback. the study took between 30-40 minutes to complete. We recorded audio during the study for later transcription and qualitative analysis.

8.3.2 Results. All participants successfully completed the predefined tasks without any assistance. Participants thought the system was easy to use (M = 4.8 ,σ=1.30), they were fond of it (M = 5.6, σ=1.34), and they thought it had potential and promise (M = 5.8, σ=1.64). As the participants experienced the proactive action prediction, they commented on how they could use OmniActions for their everyday tasks in the future. P3 stated, "*having this might fundamentally change the interaction of future AR interfaces*". OmniActions was positively received due to its ability to reduce friction by predicting the actions (P1, P2, P4).

Note that the system did not always predict the users' intended actions correctly. In these cases, the "more" function to quickly view other potential actions was used. For example, P3 commented that the *"comprehensive overview of available actions was very useful."*. This showed the importance to have mechanisms to handle the scenarios where AI predictions didn't match users' intention. However, some users noted that visiting "more" actions could increase the cognitive load as there were many options to read and choose from. Participants (P1, P5) found it overwhelming to go through all the potential actions. To address this challenge, some participants suggested using hierarchical sub-menus (P1, P3, P5) or having fewer options while treating some actions as add-ons (P2). The hierarchical sub-menus could be supported by the prediction of the general actions (which had high accuracy) and then specific actions.

Participants also shared areas of improvement for the prototype. One confusion area is the different interpretation of the wording for actions. For example, "*I thought Save-to-list is saving something important to me while Save-for-reference is something that is not important*" (P3). P2 also 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. Overall, participants were enthusiastic about OmniActions, saw its value for end-users and developers, and provided suggestions for its improvement.

9 DISCUSSION

In this section, we reflect on the design and evaluation of OmniActions. Our insights shed light on the design and implementation of proactive interfaces for AR use cases. We will also discuss the limitations of our current data and method, and a future direction to address these limitations.

9.1 Action Space for Everyday Information Encounters

As far as we know, our work was the first to identify the set of actions people tend to take on the information they encounter during everyday tasks. The diary study method enabled us to capture moments of action needs in-situ, covering the majority of everyday activity types as context. In half of the cases, these activities involved high physical/social/cognitive effort, raising the importance to reduce the friction to any additional interactions. These everyday life scenarios captured in our dataset overlap with those in the Pervasive AR vision, where people use AR anytime and anywhere [25]. Taking the lens of the Jobs-to-be-done [12], each action was "hired" to address a human need, such as staying connected, getting emotional support, reducing memory load, and gaining more understanding, etc. While technology may be fast-evolving, human needs remain relatively stable.

We understand, however, that actions shared by the participants in the current study are limited to actions they are familiar with on their current devices, specifically, phone-based actions. We expect these actions will be different when AR platforms are widely adopted and the ecosystems of actions on these platforms thrive. This kind of socio-technical co-evolution has been witnessed when we look into the literature about how people handled information needs with mobile phones decades ago. Back in 2008, people addressed information needs using the web, map, and calling on the phone, as well as through physical means (e.g. printing, asking something) [59]. In contrast, our dataset shows a greater diversity of actions people can take than before, thanks to the fast evolution and wide adoption of smartphones. Given the accelerating pace of technology, we can also expect an increased number of capabilities and variety of actions supported by future AR platforms with an always-on sensor stack and increased computational intelligence. For example, the actions will be more adaptive to users' contexts with multi-sensor streams, more proactive with better prediction of users' intention from eye tracking, and more tailored to users' preferences and goals with the first-person perspective cameras etc. These new computing platforms will be able to provide users with different actions tailored to the individual to address their everyday needs in response to information triggers in the world. For the future systems that predict user's follow-up actions, developers will need to update their data over time to reflect the evolution of the actions, like any AI system would do.

9.2 Actions Prediction with Multimodal Information

With OmniActions, we created a pipeline that predicts follow-up actions and target information by turning the multi-modal information into structured texts to LLM. Among several state-of-art LLM techniques, we identified the most performant one (in-context learning with chain-of-thoughts, using enough examples to cover the diversity of the actions) that reaches competitive accuracy in prediction accuracy.

Compared to multimodal LLMs (e.g. GPT-4v) where raw multimodal information was used directly as input, our approach has more transparency and explanability. We could evaluate how much each of the contextual factors contribute to end-to-end prediction

performance by including/excluding it. We could better leverage chain-of-thoughts because the reasoning involves multiple contextual factors. We generate the chain-of-thoughts prompts from user data rather than a researcher's common sense. For all three LLM techniques we evaluated (intent classification, fine-tuning, and in-context learning), their performance relied on the dataset quality. Therefore, it is critical to collect up-to-date and relevant data that covers a wide range of the action space. As we mentioned in the above section, when the computing platforms like AR evolves, the action space will change, and the data needs to be updated.

In our work, the collection of data from the diary study and the use of the data in prediction were two separate steps. We envision integrating the data with online training. Users wear a lifelogging system throughout the day (*e.g.*, RayBan Stories[11]); it captures how people act upon what information over time and trains/prompts the model with the data. Users could also later reflect on the data, identify important information they missed, and label potential actions related to it. This way the pipeline will gain up-to-date and personalized data iteratively with the user.

9.3 Handling Predictions Errors

Like many other AI-based predictions, our system makes errors. With our mobile prototype that leverages OmniActions to surface actions, we got valuable feedback about users' reactions and suggestions when the prediction did not match their intention. It is critical to have mechanisms to recover from error (the "Offer Simple Error Handling" rule [58]), however, we observed in the user feedback sessions that presenting a "more" button to list the rest of the actions may increase people's cognitive load. One way to reduce the cognitive load in error handling might be to leverage the higher-level grouping of actions, which achieved a high accuracy (94%) in the general action prediction. This would then funnel users to the right categories of actions from which they could process a smaller set of sub-actions.

10 CONCLUSION

In this paper, we presented OmniActions, which predicts follow-up actions when users encounter multimodal information. To inform the design of OmniActions, we conducted a five-day diary study to understand of the design space of follow-up actions. Through the study, we identified 7 general categories of actions (i.e., *share*, *save*, *remind*, *look up*, *digital extract*, *media manipulation*, and *complex actions*) and 17 specific follow-up categories of actions.

We then developed the OmniActions pipeline and prototype to predict follow-up actions for multimodal information powered by an LLM. The system harnessed the reasoning capabilities of LLMs by introducing intermediate reasoning steps (*i.e.*, CoT prompting). We evaluated three state-of-art LLM techniques, and the results indicated that integrating CoT prompting significantly improved the system's performance. Specifically, the model attained 94% accuracy when predicting top three general actions when using in-context learning with CoT prompting. We then conducted a user study to understand users' feedback towards the action prediction and its errors. The findings demonstrated the potential of OmniActions and provided valuable insights into possible enhancements for systems alike.

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

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] Alia Amin, Sian Townsend, Jacco van Ossenbruggen, and Lynda Hardman. 2009. Fancy a drink in canary wharf?: A user study on location-based mobile search. In *Human-Computer Interaction–INTERACT 2009: 12th IFIP TC 13 International Conference, Uppsala, Sweden, August 24-28, 2009, Proceedings, Part I 12*. Springer, 736–749.

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

[5] Joel Brandt, Noah Weiss, and Scott R Klemmer. 2007. txt 4 l8r: lowering the burden for diary studies under mobile conditions. In *CHI'07 extended abstracts on Human factors in computing systems*. 2303–2308.

[6] 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.

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

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

[9] Li Chen and Luole Qi. 2010. A diary study of understanding contextual information needs during leisure traveling. In *Proceedings of the third symposium on Information interaction in context*. 265–270.

[10] 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).

[11] Mauro Cherubini, Rodrigo De Oliveira, Anna Hiltunen, and Nuria Oliver. 2011. Barriers and bridges in the adoption of today's mobile phone contextual services. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*. 167–176.

[12] Clayton M Christensen, Taddy Hall, Karen Dillon, and David S Duncan. 2016. Know your customers' jobs to be done. *Harvard business review* 94, 9 (2016), 54–62.

[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] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. arXiv:2305.06500

[16] 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).

[17] David Dearman, Melanie Kellar, and Khai N. Truong. 2008. An Examination of Daily Information Needs and Sharing Opportunities. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work* (San Diego, CA, USA) *(CSCW '08)*. Association for Computing Machinery, New York, NY, USA, 679–688. https://doi.org/10.1145/1460563.1460668

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL]

[19] 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.

[20] 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.

[21] 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.
[22] 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.
[23] Ross Girshick. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 1440–1448.
[24] 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.
[25] Jens Grubert, Tobias Langlotz, Stefanie Zollmann, and Holger Regenbrecht. 2016. Towards pervasive augmented reality: Context-awareness in augmented reality. *IEEE transactions on visualization and computer graphics* 23, 6 (2016), 1706–1724.
[26] 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.
[27] 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.
[28] Annika M Hinze, Carole Chang, and David M Nichols. 2010. Contextual queries express mobile information needs. In *Proceedings of the 12th international conference on Human computer interaction with mobile devices and services*. 327–336.
[29] Jie Huang and Kevin Chen-Chuan Chang. 2022. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403* (2022).
[30] Mina Huh, Yi-Hao Peng, and Amy Pavel. 2023. GenAssist: Making Image Generation Accessible. *arXiv preprint arXiv:2307.07589* (2023).
[31] 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).
[32] 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).
[33] Vladimir Kirilyuk, Xiuxiu Yuan, Peggy Chi, Alex Olwal, Ruofei Du, et al. 2023. Visual Captions: Augmenting Verbal Communication with On-the-fly Visuals. (2023).
[34] 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.
[35] Daijiro Komaki, Takahiro Hara, and Shojiro Nishio. 2012. How does mobile context affect people's web search behavior?: A diary study of mobile information needs and search behaviors. In *2012 IEEE 26th International Conference on Advanced Information Networking and Applications*. IEEE, 245–252.
[36] 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.
[37] 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.
[38] 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.
[39] 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.
[40] Bing Liu and Ian Lane. 2016. Attention-based recurrent neural network models for joint intent detection and slot filling. *arXiv preprint arXiv:1609.01454* (2016).
[41] 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.
[42] 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.
[43] Jakob Nielsen. 1994. *Usability engineering*. Morgan Kaufmann.
[44] U.S. Bureau of Labor Statistics. 2023. AMERICAN TIME USE SURVEY - 2022 RESULTS. https://www.bls.gov/news.release/atus.t12.htm. [Online; accessed 10-Dec-2023].
[45] Ray Oldenburg. 1999. *The great good place: Cafes, coffee shops, bookstores, bars, hair salons, and other hangouts at the heart of a community*. Da Capo Press.
[46] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. *arXiv preprint cs/0205070* (2002).
[47] 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).
[48] 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.
[49] 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).
[50] Suman Ravuri and Andreas Stolcke. 2015. Recurrent neural network and LSTM models for lexical utterance classification. In *Sixteenth annual conference of the international speech communication association*.
[51] 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).
[52] Luca Rossetto, Ivan Giangreco, and Heiko Schuldt. 2015. OSVC-Open Short Video Collection 1.0. *Technical Report* CS-2015-002 (2015).
[53] 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.
[54] 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.
[55] 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.
[56] Bill N Schilit and Marvin M Theimer. 1994. Disseminating active map information to mobile hosts. *IEEE network* 8, 5 (1994), 22–32.
[57] 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.
[58] Ben Shneiderman. 2005. Shneiderman's eight golden rules of interface design. *Retrieved july* 25 (2005), 2009.
[59] 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.
[60] Bryan Wang, Gang Li, and Yang Li. 2022. Enabling Conversational Interaction with Mobile UI using Large Language Models. *arXiv preprint arXiv:2209.08655* (2022).
[61] Bryan Wang, Yuliang Li, Zhaoyang Lv, Haijun Xia, Yan Xu, and Raj Sodhi. 2024. LAVE: LLM-Powered Agent Assistance and Language Augmentation for Video Editing. https://api.semanticscholar.org/CorpusID:267740556
[62] 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).
[63] 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.
[64] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. 2019. Detectron2. https://github.com/facebookresearch/detectron2.
[65] 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).
[66] 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.
[67] 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).

A PROMPT TEMPLATES

A.1 Chain-of-Thoughts Prompts

{"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: Identify 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
Digitize: 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 In-Context Learning Prompts to Predict Target Information

Predicting **visual** target information:

You are an assistant that predicts the target information that users take follow-up actions on when they encounter multimodal information using chain-of-thoughts analysis.
The target information include three categories: scene, object, text:
scene: users would like to take actions on the whole visual content
object: users would like to take actions on specific physical objects they see
text: users would like to take actions on visible text in the scene
Output the prediction result in a JSON dict, where applicable: "chain-of-thoughts", "prediction"

Predicting **audio** target information:

You are an assistant that predicts the target information that users take follow-up actions on when they encounter multimodal information using chain-of-thoughts analysis.
The target information include two categories: sound, speech:
sound: users would like to take actions on acoustic sound they hear
speech: users would like to take actions on someone's speech
Output the prediction result in a JSON dict, where applicable: "chain-of-thoughts", "prediction"

A.3 In-Context Learning Prompts to Predict 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:
[CATEGORIES]: [DEFINITION] (refer to Figure 7)
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: Identify 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
Digitize: 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 CONFUSION MATRICES FOR ALL APPROACHES

To compute the confusion matrices for each action category, for each data instance, we need to count both the corrected and incorrect predictions for the ground truth label. However, since we are forcing the model to predict the top-3 likely actions, this would introduce unavoidable *errors* which do not reflect the model's performance. To account for this, we only count the error when there exists at least one groud truth label that is not correctly predicted by the model.

The confusion matrices for the following approaches: (1) only predicting top-3 dominant actions, (2) intent classification, (3) finetuning GPT-3.5, (4) in-context learning with GPT-3.5 are shown in Appendix Figure 1a to 2b.

Table 4 shows the improvement from in-context learning to finetuning using the same model (GPT-3.5). The results indicate that the finetuning method is sensitive to the distribution of training data. Notably, in the case of general actions, the dominant categories are excessively predominant (>30%) accounting compared to other categories (<15%). Conversely, in specific actions, the data is more evenly spread across various non-dominant categories. Consequently, given the current data distribution, finetuning demonstrates better performance with specific actions than with general actions.

Table 4: Improvement (%) on each action category from in-context learning to finetuning.

Predicting General Actions	In-context learning	Fine-tuning	**Improvment**
Share*	82.7	96.7	**+16.9**
Save*	78.7	96.9	**+23.1**
Remind	6.2	0	-100
Look up*	66.9	93.4	**+39.6**
Digital Extract	56.4	17.9	-68.2
Complex	12.5	0	-100
Augment	40.0	20.0	-50.0
Predicting Specific Actions			
Share on social media	78.4	4.5	-94.3
Share with others*	44.9	89.5	**+99.3**
Remember	70.2	47.8	-31.9
For reference*	26.8	74.2	**+176.9**
To list	19.2	58.8	**+206.2**
Keep track	28.6	11.1	-61.2
Remind	6.2	0	-100
Search online*	64.6	70	**+8.4**
Recognize	25.9	56.7	**+118.9**
Translate	37.5	25	-33.3
Extract and access	11.1	8.3	-25.2
Transcribe	61.1	16.7	-72.7
Digitalize	16.7	0	-100
Compare	14.3	0	-100
Calculate	0	0	0
Edit media	0	0	0
Augment	10.0	33.3	**+233.0**

Bolded denotes positive improved categories.

C GENERATING THE DESIGN SPACE

C.1 Survey Questions for the Diary Study

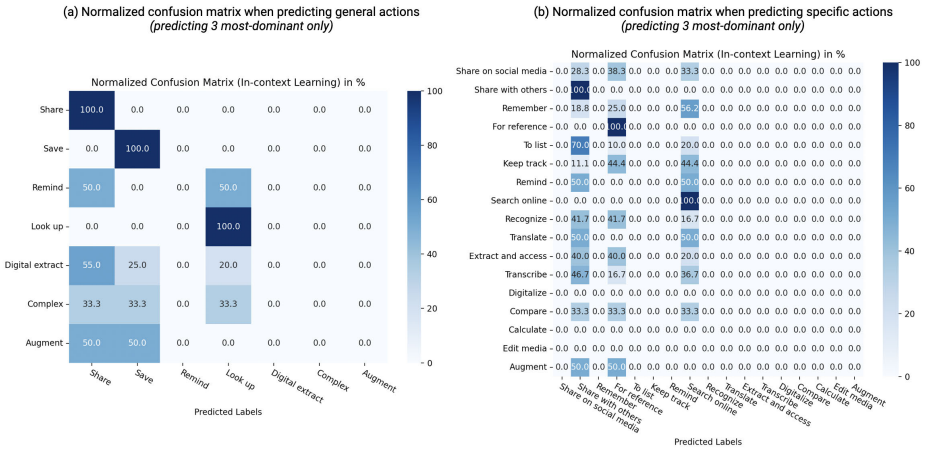
The survey questions are listed in Table 6.

C.2 Definition of Specific Follow-Up Action Categories

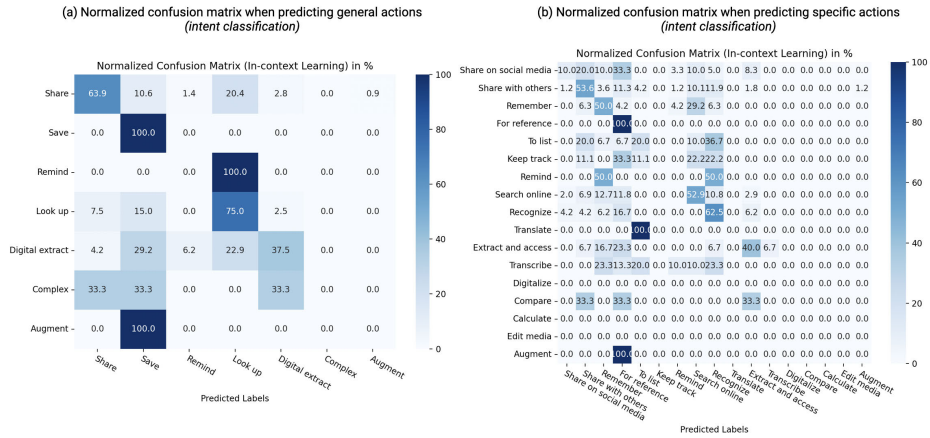
C.2.1 ***Share***.

Sharing with Others. When *sharing with others*, future systems could 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 could suggest multiple hashtags to use.



(a) Confusion matrix for only predicting the dominant actions.



(b) Confusion matrix for intent classification.

Figure 1: Confusion matrices for predicting dominant only and intent classification.

C.2.2 ***Save***.

Remember. This refers to actions where users wished to cherish a specific moment to retrieve it in the future. *Remember* often occurred when participants mentioned words such as "funny", "memorable", etc. or alongside other *share* actions.

Save for Reference. This refers to actions where users stored information with the specific goal of using it later. Participants mentioned various types of *later usages*, including using it for a later purchase, saving a gift card to avoid losing it, and so on. By automatically incorporating metadata into the information (*e.g.*, when, where, and 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 added 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, etc.) and store the information in a list.

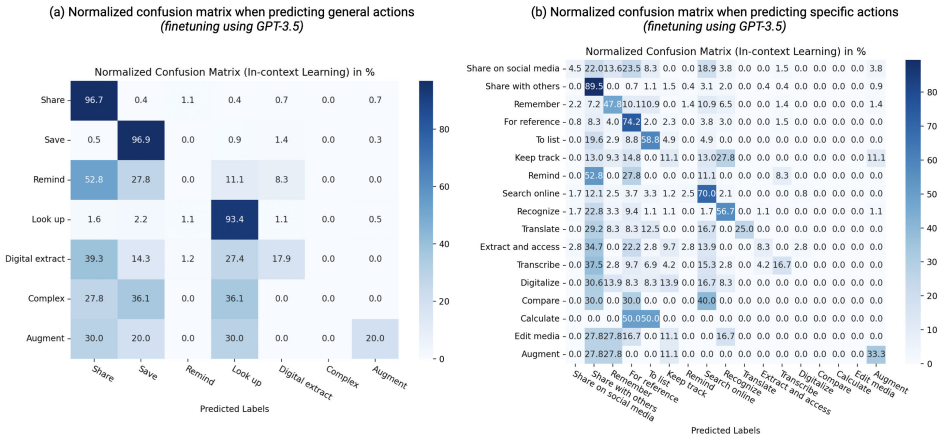
Keeping Track of Progress. Participants captured information to record their performance or progress towards specific goals such as recording the progress of their bulking (or cutting) while working out or playing the piano. Different from *saving to a list*, this information tended to be similar yet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

C.2.3 ***Look Up***.

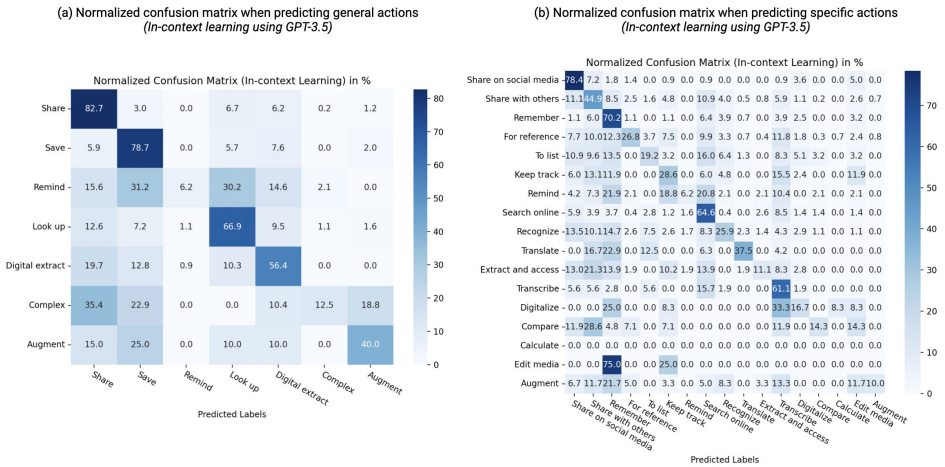
Search Online. Users conducted online searches to acquire additional information related to their intent, utilizing a variety of search tools (*e.g.*, Google).

Recognize. Users also identified information using specific tools, e.g., product searching (*e.g.*, using Google Lens or Images) or recognizing music (*e.g.*, Shazam).

Translate. In the context of text or speech, *translate* refers to the actions that sought the meaning of text or speech in a different



(a) Confusion matrix for finetuning with GPT-3.5.



(b) Confusion matrix for in context learning with GPT-3.5.

Figure 2: Confusion matrices for finetuning and in-context learning.

language, enabling one to better understand and communicate across language barriers.

C.2.4 ***Digital Extract***.

Extract and Access. These actions extracted information from the physical world and directly took action on it based on its type. For example, systems could enable users to 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 converted audio into text. This included transcribing a lecture or transcribing the lyrics from a song that was playing.

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

C.2.5 ***Media Manipulation***.

Augment Media. Augment refers to actions that enhanced 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 were 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 an image for her slides. These editing actions ranged from simple adjustments, such as cropping or resizing, to more complex alterations, such as color grading or adding visual effects.

C.2.6 ***Complex Actions***.

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 additional information and present it simultaneously for the user to compare.

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

Figure 3: An example of the collected data with four follow-up actions.

D DATA WITH AGGREGATED ACTIONS

Participants tends to perform multiple actions on the information they encounter. Figure 3 shows an example of the collected data with four follow-up actions. In this example, the participant took a picture of their rabbit as they think the rabbit might be ill. Since the rabbit will run away if they get too close, the participant decided to take a picture of the rabbit first from afar to (1) zoom in for clearer view (*augment*) and (2) share the picture with a veterinarian (*share with others*). They would also save the picture for future reference (*for reference*) and could possibly search online for more information if the veterinarian is not available (*search online*).

Table 5 shows performance of the model on data with and without aggregated actions.

Table 5: Accuracy (%) on data with and without aggregated actions (predicting top-3 actions using in-context learning)

Num of actions in data	1	2	3	4	>2	All
General Actions	98.7	91.2	68.6	87.5	85.5	94.3
Specific Actions	73.7	64.1	50.3	79.2	61.1	67.1

Table 6: 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]