

Human I/O: Towards a Unified Approach to Detecting Situational Impairments

Xingyu Bruce Liu*
University of California, Los Angeles
Los Angeles, California, USA
xingyuliu@ucla.edu

Jiahao Nick Li
University of California, Los Angeles
Los Angeles, California, USA
ljhnick@ucla.edu

David Kim
Google Research
Zurich, Switzerland
kidavid@google.com

Xiang ‘Anthony’ Chen
University of California, Los Angeles
Los Angeles, California, USA
xac@ucla.edu

Ruofei Du†
Google Research
San Francisco, California, USA
me@duruofei.com

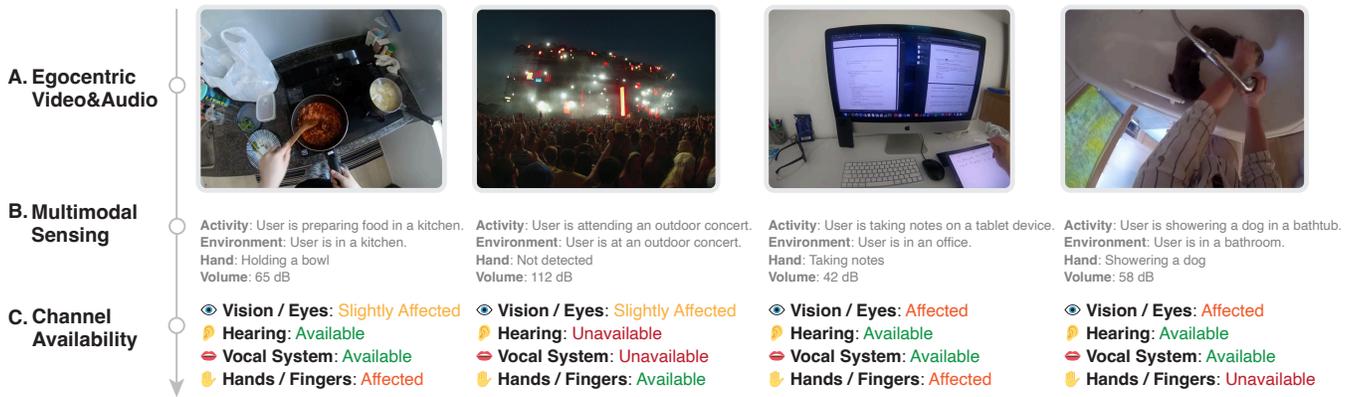


Figure 1: We introduce a new approach to detecting situational impairments based on *the availability of human input/output channels*. We instantiate this idea as Human I/O, a system that (A) captures egocentric video and audio stream; (B) processes input data and generates a description of the context; and (C) predicts the availability of vision, hearing, vocal, and hands channels.

ABSTRACT

Situationally Induced Impairments and Disabilities (SIIDs) can significantly hinder user experience in contexts such as poor lighting, noise, and multi-tasking. While prior research has introduced algorithms and systems to address these impairments, they predominantly cater to specific tasks or environments and fail to accommodate the diverse and dynamic nature of SIIDs. We introduce Human I/O, a unified approach to detecting a wide range of SIIDs by gauging the availability of human input/output channels. Leveraging egocentric vision, multimodal sensing and reasoning with large language models, Human I/O achieves a 0.22 mean absolute error and a 82% accuracy in availability prediction across 60 in-the-wild

*Project was conducted when the first author interned at Google Labs.

† Corresponding author.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

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

© 2024 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-0330-0/24/05.

<https://doi.org/10.1145/3613904.3642065>

egocentric video recordings in 32 different scenarios. Furthermore, while the core focus of our work is on the detection of SIIDs rather than the creation of adaptive user interfaces, we showcase the efficacy of our prototype via a user study with 10 participants. Findings suggest that Human I/O significantly reduces effort and improves user experience in the presence of SIIDs, paving the way for more adaptive and accessible interactive systems in the future.

CCS CONCEPTS

• **Human-centered computing** → **Accessibility systems and tools**; *Mixed / augmented reality*; *HCI theory, concepts and models*.

KEYWORDS

situational impairments, augmented reality, large language models, multimodal sensing, context awareness

ACM Reference Format:

Xingyu Bruce Liu, Jiahao Nick Li, David Kim, Xiang ‘Anthony’ Chen, and Ruofei Du. 2024. Human I/O: Towards a Unified Approach to Detecting Situational Impairments. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*, May 11–16, 2024, Honolulu, HI, USA. ACM, New York, NY, USA

1 INTRODUCTION

Everyone experiences Situationally Induced Impairments and Disabilities (SIIDs). These impairments can arise due to various situational factors, such as noise, lighting, temperature, stress, social norms, etc. For example, one might miss an important phone call in a noisy restaurant, or struggle to reply to a text message when doing dishes. These varied situational contexts in daily lives can cause temporary declines in our physical, cognitive, or emotional capacities, leading to unsatisfactory experiences.

Recently, researchers have developed systems to address SIIDs by enhancing situational awareness of mobile devices. Most systems employ a “sense-model-adapt” design pattern [53], that is, to first build a model to identify a particular situation that causes specific SIIDs, and then curate adaptations tailored to that context. For example, detecting when a person is driving [5], walking [11, 20], inebriated [38], distracted [37], or has rainwater on their touch screen [50].

However, SIIDs are often dynamic and pervasive, making it challenging to scale previous one-off solutions to accommodate users’ changing impairments in real-time, across diverse scenarios. Consider a typical morning routine: when a person is brushing teeth, they may be constrained from engaging with voice assistants; when washing face, they may struggle with reading urgent messages; and when using a hairdryer, they may miss auditory notifications from their phone. Although previous systems have developed models tailored to specific situational impairments, manually designing detection solutions for all possible scenarios and their combinations is impractical and limited in scalability.

In this paper, we propose Human I/O, a new approach that considers SIIDs not as context-specific impairments that require specific detection models, but rather through a unified lens that focuses on the *limited availability of human input/output channels*. For instance, rather than devising individual models for activities like face-washing, tooth-brushing, or hair-drying, Human I/O universally assesses the availability of a user’s vision, hearing, and hand interaction channels. With the recent development of Large Language Models (LLMs), which exhibit open-vocabulary few-shot learning and reasoning capabilities, we see an exciting opportunity to leverage LLMs and introduce a single, unified framework to identify SIIDs. This abstraction broadens our thinking of SIIDs to a comprehensive range of impairments, and allows for the development of an extensible framework that empowers other researchers and developers to continually expand. Our paper focuses on the comprehensive technical framework to detecting SIIDs, deferring the adaptation of SIID for future research.

We first conducted a formative study with 10 participants to understand the scope of modeling SIIDs based on channel availability. These insights emphasize the need for systems to integrate activity, environment, and direct sensing cues for channel availability prediction, and recognize challenges in detecting attentional, affective, and technological SIIDs. Our findings also suggest that systems should provide varying levels of channel availability, rather than a binary scale as previously assumed in most systems. This will better align with users’ needs and allow developers to create tailored strategies based on impairment severity. We iteratively

developed a four-level scale for measuring channel availability: available, slightly affected, affected, and unavailable.

These insights informed Human I/O, a unified system that can automatically detect SIIDs in a wide range of daily activities. Human I/O leverages (1) an egocentric camera and microphone, (2) computer vision and audio analysis models, and (3) the reasoning capabilities of LLMs to detect SIIDs. In Human I/O’s computational pipeline, the system first captures a user’s egocentric view with audio and video streams, providing a first-person viewpoint of the user’s state. The vision and audio models then process the input data, converting it into textual representations. Finally, we leverage LLMs with chain-of-thought reasoning [52] to analyze these textual representations and predict the current availability of human input and output channels.

We evaluated Human I/O on a dataset of 300 clips selected from 60 real-world egocentric video recordings covering 32 distinct scenarios. Human I/O reaches a 0.22 mean absolute error and 82% average accuracy on channel availability predictions, with 96% of predictions deviate by ≤ 1 from actuals. We also deployed our system in the real world and evaluated it with 10 participants, where they experienced four different scenarios with and without Human I/O. Participants found the detection and adaptation for SIIDs significantly reduced their effort level, mental, physical and temporal demands, and improved their user experience.

In summary, we contribute:

- A new approach to detecting SIIDs by modeling the availability of human input/output channels.
- Insights from a formative study that inform the design of our system, highlighting the need for integrating contextual cues, the scope of our proposed approach, and a four-level scale for measuring channel availability.
- The design and implementation of Human I/O, which leverages egocentric vision, multimodal sensing, and large language models to predict channel availability across various daily-life situations. Human I/O is deployed and open-sourced at <https://github.com/google/humano>.
- A technical evaluation of Human I/O’s performance on a diverse set of 60 in-the-wild egocentric videos, and a user study with 10 participants demonstrating its potential in improving user experience in the presence of SIIDs.

2 RELATED WORK

Our work builds upon previous research in situationally aware computing, egocentric vision, reasoning by large language models, activity and environmental sensing.

2.1 Situationally Aware Computing

Previous research in human-computer interaction and accessibility have developed systems to model different types of situational impairments. A large body of work focused on making mobile devices more situationally aware and capable of improving interaction for users experiencing SIIDs. Kane et al. investigated walking user interfaces (WUIs) [20] that adapt their layout based on user movement, demonstrating comparable performance to static interfaces. Goel et al. introduced WalkType [11], an adaptive text entry system that leverages the mobile device’s accelerometer to compensate

for movement while walking, improving typing performance. In another study, they presented ContextType [12], a system that uses hand posture information to enhance touch screen text entry. Mariakakis et al. developed SwitchBack [37], a system built upon Focus and Saccade Tracking to help users resume tasks more efficiently in the presence of distractions. They also explored drunk user interfaces [38], estimating blood alcohol levels using machine learning models trained on performance metrics and sensor data. Tung et al. proposed RainCheck [50], a solution that filters out water-caused touch points on capacitive touchscreens, enhancing interaction accuracy and target selection time.

Although prior art has advanced the field of mobile device usage under various situational impairments, their limitations lie in their narrow focus on specific situations. These efforts, while significant, only address a fraction of all SIIDs. Therefore, there remains a need for a more comprehensive approach to detect a broader range of SIIDs. In this paper, we leverage egocentric vision and large language models to address challenges of detecting SIIDs.

2.2 Egocentric Vision

The concept of using a wearable camera to gather first-person visual data dates back to the 1970s with Steve Mann’s “Digital Eye Glass” invention [36]. Since then, wearable cameras have been employed in various health-related applications within the context of Wearable AI. The Microsoft SenseCam uses a lifelogging camera with fisheye lens and trigger sensors, such as accelerometers, heat sensing, and audio devices, to aid those with poor memory as a result of disease or brain trauma [17]. Kanade and Hebert proposed a prototypical first-person vision system which consists of localization, recognition, and activity recognition components, to provide contextual awareness for caregiving applications [19]. Early computational techniques for egocentric analysis centered on hand-related activity recognition and social interaction analysis, as well as addressing challenges of temporal segmentation [16] and summarization [43] due to the unconstrained nature of video data. Over the past decade, the field has diversified, with emerging research topics including social saliency estimation [47], privacy-preserving techniques, attention-based activity analysis, hand pose analysis, understanding social dynamics and attention, and activity forecasting [23].

In this work, we build upon previous egocentric vision research to develop a more comprehensive approach to situational impairment detection. We chose egocentric vision as a means for implementation since it provides the widest “bandwidth” of detecting a broad range of SIIDs; but the key idea behind our system should be independent of such implementation.

2.3 Reasoning Capabilities of Large Language Models

Recent large language models have demonstrated reasoning capabilities via different approaches including zero-shot learning [24], few-shot learning [2], chain-of-thoughts [52] or incorporating multimodal information [54]. These reasoning abilities are particularly useful for tasks such as mathematical problem-solving [9, 30, 52], image-based question answering [31, 54], understanding human intents [34, 48], etc. They have been applied to a broad range of

research in HCI community recently including interactive coding support [18, 51], social computing [41], and communication augmentation [34]. For example, Social Simulacra uses LLMs to simulate social interactions and behaviors as social computing prototypes [41]. Visual Captions leverages a fine-tuned large language model to proactively suggest relevant visuals in open-vocabulary conversations [34]. InstructPipe [56] employs a node selector, a code writer, and a code interpreter to create AI pipelines from human instructions.

Such reasoning capabilities of large language models make it possible for our system to, in an open-vocabulary manner, predict the availability of human input/output channels based on the detected activity, environment, and other contextual information.

2.4 Activity and Environmental Sensing

A wide variety of sensing technologies and strategies [4, 10, 26–29, 49] have been investigated to achieve detection of human activity and measuring a physical environment. More relevant to our work are sensing approaches that utilize camera-based [1, 21] and audio-based [25, 44] systems. For example, Mo et al. [39] applied deep learning to classify 5 different locations and 12 distinct activities. BodyBeat [44] employs a piezoelectric microphone to detect on-speech body sounds, such as eating noise, breathing, laughter, and coughing. Other works [22, 45, 49] demonstrate promising outcomes when using multiple modalities.

In our paper, we develop a novel framework that enables the reasoning of human input/output channels through the use of computer vision and audio analysis of video and audio streams. Leveraging large language models, our framework is highly adaptable, allowing for easy integration with both existing and forthcoming sensing technologies for activity and environmental monitoring.

2.5 SIIDs as the Availability of Human I/O Channels

Our motivation for using the availability of human input/output channels for detecting SIIDs lies on Dix et al.’s fundamental model of human-computer interaction [6]: humans, similar to computer I/O, receive and send information via different channels. For example, we use vision, hearing, tactile, etc. to receive information coming from the world (input); and use vocal system, eye gaze, hand gestures, etc. to convey information (output). He and Card [3] both describe human as an “*information processing system*” with limited capacity to process information through various channels. In addition, Microsoft’s Inclusive 101 Guidebook¹ provides an overview chart of four types of permanent, temporary, and situational impairments: touch, see, hear, and speak, categorized by human sensory channels. Similarly, CrossA11y [35] divides video accessibility issues into lack of information in the visual and auditory channels.

Building upon these inspirations, we hypothesized that modeling the availability of human input/output channels might provide a more unified approach to detecting SIIDs. We summarize a list of human input/output channels from prior work (Figure 2). In this paper, we focus on channels that are most commonly used in human-computer interaction: vision, hearing, tactile (input), and eyes/gaze, vocal system, hands/fingers (output).

¹Microsoft’s Inclusive 101 Guidebook: <https://inclusive.microsoft.design>

Human Input		Example use cases
 Vision		Read text message, watch video, etc.
 Hearing		Hear notification, phone call, etc.
 Tactile		Haptic feedback, feel temperature, etc.
 Taste		Drink, eat, etc.
 Smell		Smell, etc.
Human Output		
 Eyes / Gaze		Face ID, gaze-based interaction, etc.
 Vocal System		Conversation, voice assistants, etc.
 Hands / Fingers		Touch screen, gesture control, etc.
 Limbs / Movement		Walk, reach, etc.
 Head / Face		Nodding, facial expression, etc.

Figure 2: Human input/output channels with channels most commonly used in human-computer interaction highlighted in black. We designed and implemented Human I/O based on these channels.

3 FORMATIVE STUDY

To validate the feasibility and further explore the scope of modeling SIIDs as the availability of human input/output channels, we conducted a remote whiteboard session with 10 participants. We report on our insights and how they informed our design and implementation of the Human I/O system.

3.1 Procedure

We recruited 10 participants via group email invitations and internal communication channels in Google. Participants had various technical and non-technical backgrounds, including software engineers, researchers, UX designers, visual designers, students, etc. In a 90-minute online brainstorming session, we first introduced and explained what SIIDs are, showed videos of previous systems that can detect and adapt to SIIDs, and presented our initial ideas on identifying SIIDs by estimating the availability of human input/output channels. Participants then brainstormed on a digital whiteboard (Figure 3) based on three prompts: (1) For each input/output channel, what are some situations that make it unavailable? (2) For each input/output channel, when it is unavailable, what are some implicated consequences? (3) For each impairment, to what extent would you like to have an adaptive system to intervene versus overcoming it yourself? We went over each participant’s responses and asked them to explain and elaborate on their examples after brainstorming.

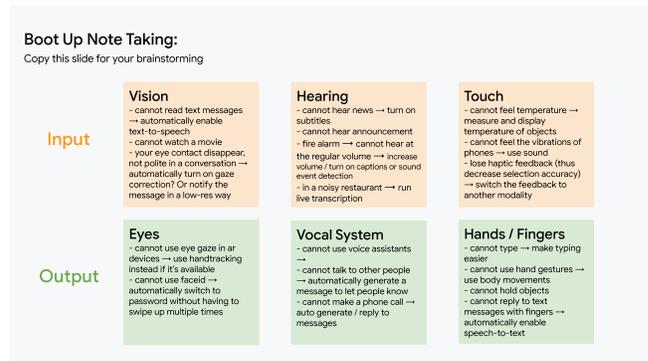


Figure 3: An example brainstorming whiteboard from a participant.

3.2 Findings

Two researchers organized and analyzed participants’ responses with the affinity diagram approach. Informed by the set of themes derived from the grouped notes, we present insights (I1 to I5) around methods to predict channel availability, the scope of our approach in detecting SIIDs, and different levels of channel availability.

3.2.1 Methods to Predict Channel Availability. We first asked participants to brainstorm for each channel, what were some situations that would make it unavailable. Participants brainstormed 82 situations in total. We found that these situations can be broadly detected via three ways (with overlaps) (I1):

Activity-based (46 mentions): Participants identified that sometimes unavailability of a channel is due to the user’s engagement in activities, such as driving, cooking, or attending a meeting.

Environment-based (26 mentions): Environmental factors were recognized as another source of channel unavailability, such as traveling on an airplane or studying in a library.

Channel-based (20 mentions): Participants also noted that some situations could be directly sensed by detecting the state of sensory channels, such as a user wearing headphones, in a loud environment, or holding an object in their hand.

Participants’ identified situations align with observations made by Sears et al. [46], which emphasizes the contribution of both the environment and activity on the existence of impairments and disabilities. In addition, direct sensing of channels emerged as a third theme. By measuring metrics such as the current environmental volume level, or whether the hand is occupied, systems can leverage more direct, lower-level information to predict channel availability.

Furthermore, participants mentioned that some situations can impact more than one channel. For example, playing drums would affect both the user’s ability to hear and their hands for interacting with devices. Certain channels might also be correlated with each other. In particular, the input channel of vision and the output channel of eyes/gaze are almost always available or unavailable together, similarly for tactile and hands/fingers. In Human I/O, we combine human input/output channels into four categories:

- Vision / Eye
- Hearing

- Vocal System
- Hands / Fingers

3.2.2 Scope & Limitations. We conducted a comprehensive review of the generated situations to understand the boundaries of our approach. Specifically, we sought to identify what types of SIIDs can be effectively modeled by evaluating the availability of human input/output channels and the inherent limitations. We compared participants' generated situations to the Situational Factors taxonomy proposed by Wobbrock [53], and reviewed what types of situations can be properly detected. Wobbrock analyzed a decade's work from 2008 to 2018 on situationally aware mobile devices and categorized different kinds of SIIDs into six categories: *Behavioral, Environmental, Attentional, Affective, Social, and Technological*.

Our findings suggest that our approach is capable of identifying situational impairments centered around human sensory abilities, including those induced by *behavioral* (e.g., walking, driving, operating machine), *environmental* (e.g., ambient noise, darkness), and *social* (e.g., conversation, crowd) factors. These situations often have a direct impact on the availability of human input/output channels.

However, the approach exhibits limitations in SIIDs that are **(I2): Intrinsic to Human**: Our approach may not be as effective in identifying impairments related to cognitive states, i.e., *attentional* (divided attention, distraction), *affective* factors (stress, fear, fatigue) factors, which may indirectly influence the availability of input/output channels. There is often not a clear mapping between one's mental states and their ability to use input/output channels.

Technological: Impairments induced by technological constraints, like power shortages, weak Wi-Fi connectivity, or hardware limitations—are not directly discernible through the human input/output channel metrics. Such SIIDs will need supplementary techniques or synchronization with device-centric data.

3.2.3 Levels of Channel Availability. To understand participants' preferences on how they would like to deal with SIIDs, we also asked them to discuss the extent they would desire system intervention versus overcoming the impairment themselves. This provided insights into 106 impairments, along with their preferred levels of system intervention. Some interesting examples include: to delay notifications when people's vision channel is currently engaged in activities like driving or biking, generate automatic replies when people's hands are not available to text, and turn on live captions [33] when the hearing channel is not available.

Participants highlighted that *adaptations may not always be necessary or preferred* **(I3)**, especially if the unavailability is temporary or can be easily overcome. For instance, when holding a remote controller that occupies the hand, instead of switching to voice input to interact with the device, a user might prefer to put the remote down temporarily and use the hands as the input method again. When taking a sip of coffee, users might not need adaptations for the brief moment when their vocal channel is unavailable.

These observations led us to reconsider, that *channel availability (and SIIDs in general) is not binary* **(I4)**, as assumed in many previous situationally aware systems. We should not simply model a channel as "available" or "unavailable", "impaired" or "not impaired". There are many cases in which the channel has a gradient of availability. This could depend on factors such as the difficulty

of the ongoing activity, the duration of unavailability, or the user's ability to overcome the situation.

Moreover, *availability of a channel depends not only on the current situation, but also on the incoming task* **(I5)**. For instance, one participant mentioned, when a user's hands are wet, they might still be usable to perform simple actions such as tapping on the screen to answer an incoming phone call. However, the same hands might be deemed insufficient for more intricate tasks, such as typing to respond to a text message. This relates back to our previous point that SIIDs are non-binary — e.g., the wet hand does not cause a binary impairment, but rather exhibits a variable bandwidth in its channel depending on the task at hand.

3.3 Design Implications

Based on the insights from our formative study, we outline design implications for systems that models SIIDs based on the availability of human input/output channels:

(1) Consider activity, environment, and direct sensing cues for predicting channel availability (I1): To more accurately predict channel availability, systems should take into account a combination of activity-based, environment-based, and directly sensing cues. Systems can provide a more comprehensive understanding of the user's situation.

(2) Acknowledge the limitations (I2): This approach may struggle to identify SIIDs that are attentional, affective, or technological. Designers should be aware of these limitations and consider additional methods for these types of SIIDs.

(3) Predict multiple levels of channel availability (I3, I4, I5): Systems should provide different levels of availability to align users' needs. It is important to note that sometimes users may not want the system to adapt to their situations. Hence, it's essential to provide users with the agency to decide how their SIIDs should be managed. This will also allow developers to design different strategies based on the severity.

For Human I/O, we developed a four-level channel availability based on insights from our formative study. We randomly selected a total of 20 situations proposed by participants during the session, 5 from each of the vision/eye, hearing, vocal system, and hands/fingers channels. Iteratively, two researchers first proposed descriptions of different levels, then coded the 20 situations with the drafted levels, and revised the description of the levels based on the disagreements or ambiguities. Researchers repeated this process for three meetings until full agreement was reached. We identified four levels of channel availability:

- **Available:** The channel is currently not involved in any activity, or constrained by any environmental factors. It takes low to zero effort to use the channel to do a new task.
Example: A user is sitting at their desk with their hands free, eyes not engaged in any task, and no background noise interfering with their hearing or speech.
- **Slightly Affected:** The channel is engaged in an activity or constrained by an environmental factor. Given a new task that requires the channel, users can multitask, easily pause and resume to the current activity, or easily overcome the situation.

Example: A user is holding a remote control, which can be quickly put down to free up their hand for another task.

- **Affected:** The channel is involved in an activity or constrained by an environmental factor. Given a new task, the user may experience inconvenience or require some effort to use the channel.

Example: A user is carrying grocery bags in both hands, making it challenging to use their hands for other tasks without putting the bags down first.

- **Unavailable:** The channel is completely unavailable due to an activity or environmental factor, and the user cannot use it for a new task without substantial changes, significant adaptation or changing the environment.

Example: A user is attending a loud concert, making it impossible for them to hear incoming notifications or carry on a conversation without stepping outside.

We observed that the distinction amongst these four levels hinges on the amount of effort for a user to free up a channel for an interactive task and re-occupy the channel later. To validate the consistency and applicability of these levels, we continued to label all remaining situations independently. We computed the inter-rater reliability and the result shows a high level of agreement between raters, with Cohen's Kappa $\kappa = 0.847$.

4 HUMAN I/O SYSTEM

Following our formative study insights, we developed Human I/O, a system that detects situational impairments based on the availability of human input/output channels.

4.1 Overview

The Human I/O computational pipeline, illustrated in Figure 4, consists of three components: (1) An egocentric camera and microphone capturing video and audio streams of the user's current situation (Figure 4.1). (2) A processing module that processes the video and audio data in one-second intervals using a combination of computer vision, natural language processing, and audio analysis algorithms, and generates a rich set of data, including the user's activity, environment, and direct sensing of specific channels (Figure 4.2). (3) A reasoning module that leverages a large language model to process the contextual information. It employs chain-of-thought prompting to predict the availability of vision/eye, hearing, vocal, and hands/fingers channels. A smoothing algorithm is incorporated at the end to enhance system reliability (Figure 4.3).

We implemented Human I/O as a web application to offer a versatile and accessible platform both for users to learn about their SIIDs in daily lives, and developers to debug and evaluate detection methods (Figure 7). This approach enables connectivity with different cameras and microphones and allows for easy experimentation on different devices, such as mobile phones, tablets and AR glasses. Furthermore, the web app supports testing on both live video streams and pre-recorded videos, providing a flexible environment for evaluation and user studies. A live demo and open-sourced repository of Human I/O can be found at <https://github.com/google/humanoio>.

4.2 Data Capture

In our research setup, Human I/O uses a webcam (Logitech C930e) and its integrated microphone to obtain real-time video and audio streams for data capture. We envision that future implementations of Human I/O would seamlessly integrate with lightweight, all-day AR glasses [40] equipped with an array of sensors, such as cameras, LiDARs, microphones, eye trackers, and inertial measurement units (IMUs). These sensors will enable richer data capture and provide more comprehensive input to enhance the system's capability.

4.3 Processing Module

Human I/O recognizes the user's context by analyzing video and audio data in one-second intervals.

4.3.1 Activity Description. To detect the current activity in an egocentric video, we employ a two-step process. First, we generate an image caption of the current video frame (compressed to 640×480) using the state-of-the-art image captioning model in early 2023, BLIP-2 [31]. Although BLIP-2 generates high-quality and objective descriptions, it occasionally falls short in providing an explicit "activity" description, *i.e.*, what the person wearing the camera is doing. For instance, consider the examples shown in Figure 5.

The BLIP-2 output for a tennis court frame is "a tennis court with lights on at night", while for a frame capturing the inside of a bus, it outputs "a view of the inside of a passenger bus". These captions describe the scenes but do not effectively convey the user's actions.

To address this limitation, we integrate the BLIP-2 model with GPT-3 text-curie-001, a faster version of the GPT-3 model capable of simpler tasks. We use the following prompt structure to guide the GPT-3 model to generate a more accurate activity description (full version in Appendix C):

```
"An egocentric view of User is showing" +
<BLIP-2 output> +
"Describe what User is doing concisely. Answer in the
format of 'User is ...'"
```

By combining the two models, we obtain a refined activity description that better reflects the user's actions. Referring back to the examples in Figure 5.1 and 5.2, the integrated output for the tennis court frame becomes "User is playing tennis at night on a lit court", while for the bus frame, it produces "User is riding on a bus".

4.3.2 Environment Description. To identify user's current environment, we once again combine the outputs of the BLIP-2 and GPT-3 text-curie-001 model. We forward the image caption from BLIP-2 to GPT-3 using the following prompt structure:

```
"An egocentric view of User is showing" +
<BLIP-2 output> +
"What location or environment is User likely to be in?
Answer in the format of 'User is in...'"
```

This improves the quality of environment descriptions. For instance, consider an egocentric video frame showing a person washing dishes (Figure 5.3). The original BLIP-2 caption states, "A person washing dishes in a sink". With the integration of the GPT-3 model, the output is refined to, "User is in a kitchen". For a video frame displaying a person playing a computer game (Figure 5.4), the initial BLIP-2 caption reads, "A person playing a video game on a

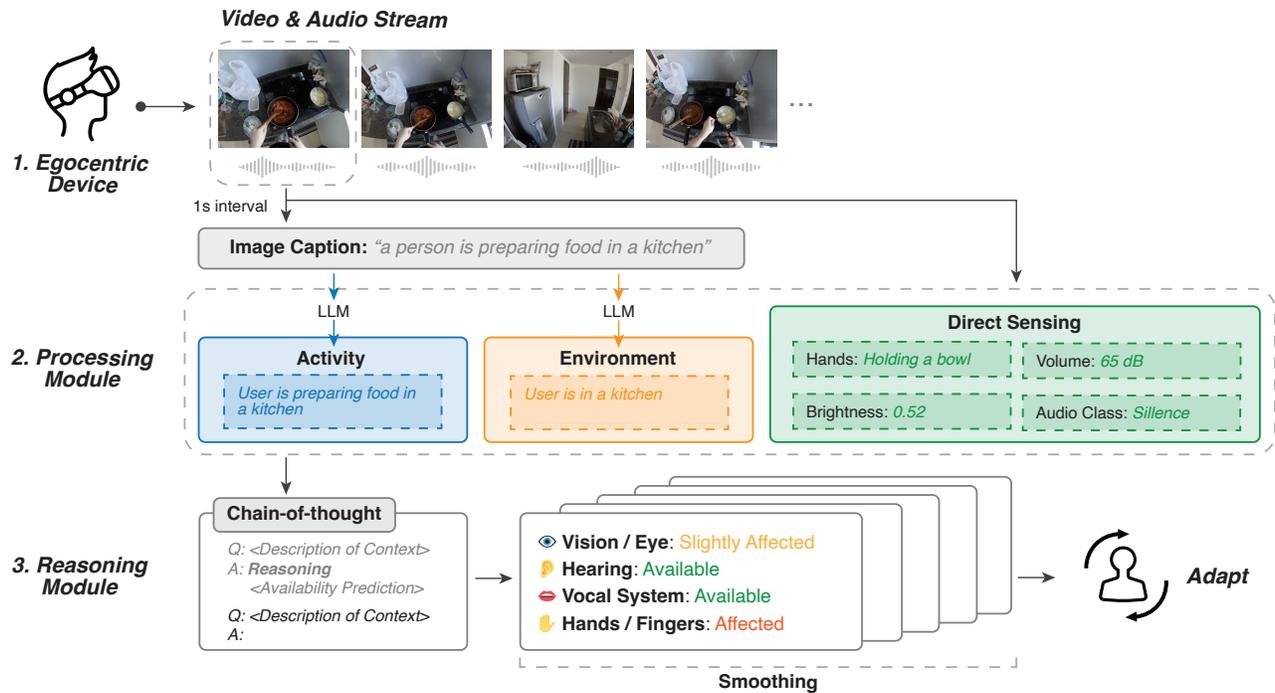


Figure 4: The Human I/O pipeline comprises three components: (1) an camera and microphone capturing the user’s egocentric video and audio stream; (2) video and audio data processing using computer vision, NLP, and audio analysis to obtain contextual information, including user’s activity, environment, and direct sensing ; and (3) sending contextual information to a large language model with chain-of-thought prompting techniques, predicting channel availability, and incorporating a smoothing algorithm for enhanced system stability.

	Activity	Raw BLIP-2 Output → Refined (GPT-3 text-curie-001)
(1)		A tennis court with lights on at night. User is playing tennis at night on a lit court.
(2)		A view of the inside of a passenger bus. User is riding on a bus.
(3)	Environment 	A person washing dishes in a sink. User is in a kitchen
(4)		A person playing a video game on a computer. User is in a room, likely indoor

Figure 5: Examples of using GPT-3 (text-curie-001) to refine raw image caption results from BLIP-2 to get more accurate descriptions of the current activity and environment.

computer”. Following the integration with GPT-3, the description becomes: “User is in a room, likely indoors”.

4.3.3 Direct Sensing. Since activity and environment detection may miss information that is not be adequately represented by their high-level descriptions, we also implement direct sensing techniques to gather a more comprehensive set of data. Specifically,

we consider hand detection, volume level, audio classification, and environmental brightness.

First, our hand detection algorithm consists of three stages. We first use the MediaPipe Hands model [55] to obtain keypoint localizations of 21 3D hand-knuckle coordinates for both hands. If no hand is detected at this stage, the process is halted, and outputs “No hand is detected.”

If a hand is detected, we proceed by utilizing the MediaPipe object detection model (efficientdet_lite0) [14] to detect if either hand is holding any object and what object is it holding (details in Appendix B). If a hand is holding an object, the system outputs “Hand is holding <Object>.”

Finally, if a hand is detected but not holding an object, we use the BLIP-2 Visual Question Answering (VQA) model to ask, “What are the hands doing?”. This unconstrained approach is particularly useful when hand landmarks or objects are not accurately recognized (e.g., object detection model cannot identify drumsticks), or in complex scenarios such as typing, washing hands, etc. In this case, we use the result of BLIP-2 VQA as output.

In addition, Human I/O directly senses the availability of the hearing and vision channels through volume level, audio event classification, and brightness measurements. Volume level is determined using the Web Audio API, and measurements are smoothed and converted to decibels. Audio event classification leverages the pre-trained YAMNet model [42], capable of detecting 521 distinct

audio events. Brightness is assessed using relative luminance, following the WCAG accessibility guideline and Rec. 709 coefficients². Temporal smoothing is applied to volume level and brightness measurements. For further details on the implementations, please refer to [Appendix B](#).

4.4 Reasoning Module

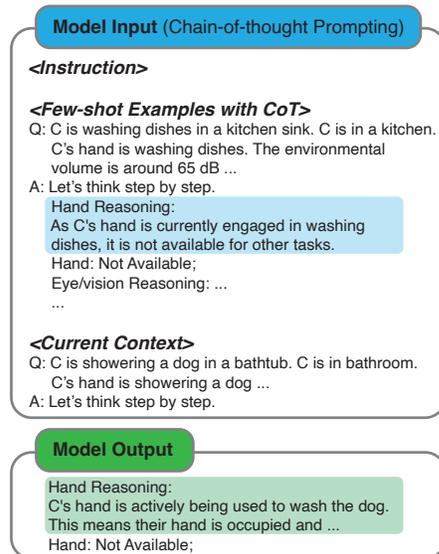


Figure 6: An illustration of our prompt structure leveraging chain-of-thought (CoT, highlighted) to enable LLMs to predict channel availability from the context.

We integrate all intermediate results and employ chain-of-thought (CoT) prompting [52] with GPT-4 to predict the availability of human input/output channels. CoT involves providing the model with a prompt composed of triples: `<input, CoT, output>`, where the chain-of-thought contains a series of intermediate natural language reasoning steps leading to the final output. This method allows the model to accumulate and maintain context and consistency throughout the prediction process.

While CoT is prevalent in LLM research, our method converts multimodal data into structured text, merging audio-visual and language models. Combining textual descriptions with quantitative sensory data, our approach offers richer context, enhancing system accuracy and nuance in predictions. For Human I/O, the prompt consists of three components: *Instruction*, *Few-shot Examples*, and *Current Context*. Please refer to [Figure 6](#) for the prompt structure and [Appendix C](#) for the complete prompt used in our system.

4.4.1 Instruction. A prefix that clarifies the definitions of our four-level scale availability, as outlined in [subsection 3.3](#), and the task. The task is described as follows:

“Given the current activity and environment as described below, determine the availability of the user’s vision/eye, hearing, vision, and hands/fingers channels.”

²Relative luminance: https://www.w3.org/WAI/GL/wiki/Relative_luminance

4.4.2 Few-shot Examples. Three few-shot examples are provided, each comprising input, chain-of-thought, and output. We selected three examples representing different availability in various channels, different activities and environments: washing dishes in a kitchen, playing an acoustic guitar in a room, and working on a laptop in a library. For each few-shot example, the input contains activity and environment descriptions, along with direct sensing outputs for hands, volume level, audio classification, and brightness.

We construct the chain-of-thought by defining the intermediate reasoning steps that the model should follow to derive the availability of a channel based on the context. For example, in the case of the hearing channel, the chain of thought may involve considering the volume level, the presence of noise or other sound events, and the user’s current activity (e.g. playing guitar) and environment (e.g. in a library). The model is guided through these reasoning steps and prompted to provide a final output predicting the availability of the hearing channel. A similar process is followed for the other I/O channels.

The output is a four-level availability score, as described in [subsection 3.3](#).

4.4.3 Current Context. We combine the formatted outputs from the processing module. For example, for the situation in [Figure 4](#), the combined context is:

Q: User is preparing food in a kitchen. User is in a kitchen. User’s hand is holding a bowl. The environmental volume is around 65dB. No audio event is detected in the environment. The luminance value of the current environment is 0.52, in the range of 0 to 1.

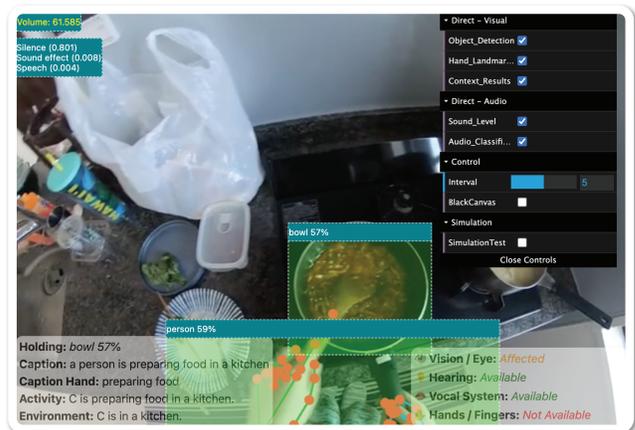


Figure 7: The web-based visualization interface of our prototype Human I/O system. The prediction results are displayed in the bottom-right. Intermediate auditory information is shown on the top-left corner, and visual information is shown on the bottom-left. The settings is shown as a drop-down menu on the top-right. Detected objects and hands are also highlighted in the video in real-time.

4.4.4 Availability Prediction. We combine instruction, few-shot examples, and the current context as the final input to GPT-4. We

also added a suffix “A: *Let’s think step by step*” [24]. The model output contains the reasoning and availability prediction of each channel.

4.4.5 Temporal Smoothing. To account for potential fluctuations in the prediction output, we apply a smoothing window of size 5 to the model’s availability predictions. Human I/O runs every second, and with the smoothing window in place, it only outputs an availability prediction if more than three of the past five predictions are the same. This majority prediction is then output as the final availability determination. If there is no majority prediction within the past five predictions, the system outputs “Unsure”. This smoothing technique reduces the potential impact of brief, sporadic changes in the user’s context or sensing outputs.

4.4.6 Lite Version. Although chain-of-thought provides robust reasoning to ensure prediction accuracy, generating the reasoning texts is the bottleneck of the system speed. To enable more real-time prediction, we created a lite version by using GPT-3.5-turbo and removing all intermediate reasoning steps in the few-shot examples in the prompt and “*Let’s think step by step*” (Appendix C). This substantially decreases inference speed to under 1 second, as the system only needs to generate availability predictions for four channels (around 10 tokens), with a lighter model. However, we observe a decrease in performance. We provide a quantitative analysis comparing the two approaches in section 5.

4.5 Human I/O Visualization Interface

We developed a Human I/O Visualization Interface (Figure 7), primarily as a tool for developers and researchers, to provide real-time monitoring of the processing outputs (along with hand landmarks and object detection) and predictions of availability for all four communication channels. The visualization interface has been designed with the flexibility to analyze both live and pre-recorded video feeds. We can conveniently conduct technical evaluations by loading pre-recorded videos and analyzing the data. In addition, the interface offers a logging system that records all intermediate results throughout the processing pipeline. The interface is deployed live at <https://github.com/google/humanio>.

5 TECHNICAL EVALUATION WITH IN-THE-WILD VIDEOS

We evaluated Human I/O on 300 clips from a set of 60 in-the-wild egocentric video recordings under 32 different scenarios. We report the accuracy (mean absolute error and classification accuracy) and the consistency (intra-video variance) of our system on channel availability predictions.

5.1 Materials

Our sample was sourced from Ego4D v1 [15], an extensive egocentric dataset with over 3,670 hours of daily-life activity videos, recorded in-the-wild. We applied a filter to select videos shorter than two minutes with audio, and then randomly selected 60 videos, ensuring a diverse range of scenarios. Each video features a single, coherent activity. For each video, we randomly selected 5 non-overlapping clips aligned with the main activity, totaling 300 tested clips. Our final sample comprises 32 distinct scenarios (Figure 8),

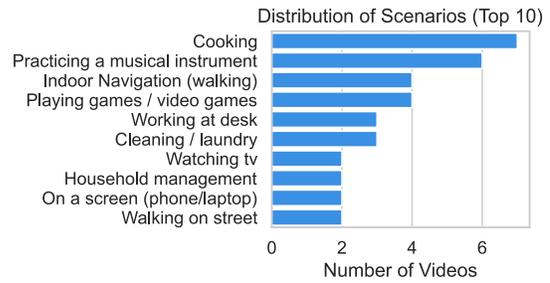


Figure 8: Distribution of the top 10 most common scenarios in our evaluation dataset. Full distribution of the scenarios is shown in Appendix subsection D.1.

such as cooking, practicing a musical instrument, working, playing with pets, cleaning, doing laundry, walking, sitting on a sofa etc.

We executed Human I/O on each video, logging various data points, including image caption, activity description, environment description, direct sensing results (hand presence, audio volume, audio class, and brightness), LLM reasoning output, and channel availability predictions. We recorded the smoothed outputs over the video windows. To compare performance, we also ran the lite version of Human I/O.

5.2 Data Annotation

Two researchers first independently watched the videos, and annotated the level of availability for vision/eyes, hearing, vocal system and hands/fingers channels based on the four-level scale we developed (subsection 3.3) on a spreadsheet. Similar to our formative study results, we reached a high level of inter-rater agreement (Cohen’s Kappa $\kappa = 0.862$). We then discussed the mismatches and resolved all disagreements. In the final sample, 10.0% of the channels are labeled as “Unavailable”, 18.33% “Affected”, 30.42% “Slightly Affected”, and 41.25% “Available”.

5.3 Metrics

We assess Human I/O using three quantitative metrics: (1) *Mean Absolute Error (MAE)*: This metric captures the average discrepancy between predicted availability levels and the actual ground truth for each of the four channels within every video clip. Given the nature of our data, MAE offers a precise error depiction. This is because misclassifying a level as “Slightly Affected” instead of “Affected” is less erroneous than confusing “Unavailable” with “Available”. Availability levels are numerically translated to values ranging from 1 to 4. (2) *Averaged Classification Accuracy (ACC)*: In addition to MAE, we provide an accuracy measurement for a more intuitive understanding of classification performance. (3) *Intra-video Variance (VAR)*: Assessing the model’s consistency throughout a continuous audio-visual stream is crucial to prevent flickering predictions. The variance score is determined by first calculating it based on video ID and channel then averaging the results.

Channels	Human I/O			Human I/O Lite		
	MAE	ACC	VAR	MAE	ACC	VAR
Vision/Eyes	0.25	76.0%	0.17	0.45	62.0%	0.16
Hearing	0.23	86.7%	0.05	0.37	63.3%	0.11
Vocal System	0.08	92.0%	0.01	0.23	87.3%	0.20
Hands/Fingers	0.33	72.7%	0.18	0.70	46.7%	0.30
Total	0.22	81.8%	0.10	0.44	64.8%	0.19

Table 1: Technical evaluation of Human I/O and Human I/O Lite. We report the mean absolute error (MAE), average classification accuracy (ACC), and average intra-video variance (VAR) for four channels and the overall results. Our system estimates availability levels with small margins of error and variance. In Human I/O, 96.0% of predictions are within a discrepancy of 1 from the actual value.

5.4 Results

We analyzed the performance of Human I/O in predicting the availability of each input/output channel. Table 1 summarizes our results. The system achieves a mean absolute error (MAE) of 0.22, and an average accuracy of 81.8% across all channels. All MAEs of the four channels are under 0.33. These low MAE values indicate that the system’s predictions closely align with the actual availability, with deviations being less than a third of the actual level on average. Breaking down the performance by individual channels, we observed accuracies of 76.0% for eyes, 86.7% for hearing, 92.0% for vocal, and 72.7% for hands, with corresponding MAEs of 0.25, 0.23, 0.08, and 0.33. Notably, 96% of the system’s predictions differed from the actual values by 1 or less, and no predictions had a difference greater than 2, demonstrating a reliable performance across different channels. Additionally, the overall low variance in predictions (0.1), with channel-specific variances of 0.17 (eyes), 0.05 (hearing), 0.01 (vocal), and 0.18 (hands), reflects the system’s consistency. This consistency is important for practical applications, as it ensures that the system reliably predicts the level of impairments across different instances in the recording.

For Human I/O Lite, we observe overall slightly inferior performance compared to the full model. However, the MAE for Human I/O Lite is still at a very low level around 0.44, showing promising ability to predict SIIDs even with reduced computational resources.

5.5 Latency

We evaluated our computation pipeline’s latency: Human I/O with a full chain-of-thought prompt had an average latency of 19.95 seconds with GPT-4, and 7.33 seconds with GPT-3.5-turbo. The Human I/O Lite (without chain-of-thought reasoning) exhibited a significant improvement, with an average latency of 2.07 seconds.

From our observations in the dataset, user’s channel availability typically spans more than one minute. This suggests that even a delayed prediction could still be accurate when using it in real-time. In addition, while latency might matter more for shorter activities that last only a few seconds, our formative study suggests that those

short activities likely do not necessitate adaptations. However, we acknowledge that certain scenarios may still demand low latency. For instance, when users require rapid adaptations in response to changing situations (e.g., entering a noisy subway while still on an important phone call).

Future system can incorporate temporal segmentation techniques to detect changes in scenes or activities, thus eliminating the need for the system to run continuously at one-second intervals. Alternatively, as the current system’s main bottleneck is the inference speed of LLMs, the use of lighter weight models, potentially fine-tuned on an extensive dataset, could be employed to enable faster computation and improve system performance. Such models may also provide a more concise representation of SIIDs predictions.

In addition, future work could explore identifying pre-impairment scenarios. That is, to anticipate situational impairments before they occur, thereby preparing the system for a timely response. For instance, the system might start making inferences as soon as a faucet is opened, predicting the imminent unavailability of hands due to washing. While acknowledging the potential for false positives, this method may enhance responsiveness of the system particularly in dynamic environments.

5.6 Failure Cases Analysis

Our results indicate that the system have similarly effective performance at predicting the availability of the vision, hearing and vocal channels, while the performance in predicting hand availability is less satisfactory. After reviewing the failure cases, we speculate that this lower performance for the hands channel may be attributed to several factors: (1) The complexity of hand-related tasks and the wide range of possible hand impairments make it challenging to accurately capture all nuances of hand-related SIIDs. (2) Occlusion that might affect the quality of the data captured by the egocentric vision system. (3) The few-shot examples may not be sufficiently diverse to represent hand-related SIIDs, thus affecting the model’s performance in this specific channel.

In addition, we observed that many incorrect predictions are related to inaccurate activity recognition. The system tends to fail in situations with unclear activities, such as walking around in the bathroom. However, it performs well when the activities are more explicit, such as washing hands, cooking food, or playing drums. Future versions of the system can explore balancing the weights for the more confident direct sensing results, and reduce the over-reliance on high-level descriptions of the activity and environment.

The system also struggles to discern subtle differences between similar activities. For instance, consider two kitchen-related videos where one person is washing their hands and another is stirring food in a pan. The system detects both activities as “preparing food in the kitchen”, and classifies hand availability as “unavailable” in both cases. However, hands can be briefly used for other tasks while stirring food, so they should be considered as “affected” rather than “unavailable”.

Anecdotally, even when the system misrecognizes the activity completely (e.g., scrubbing wood with sandpaper as climbing a wooden ladder), it sometimes still provides relatively accurate channel predictions due to similarities in hand occupancy, environmental volume level, and other factors.

6 EXPERIENCING HUMAN I/O IN REAL-TIME

While our focus remains on the detection of SIIDs, we conducted an additional study to understand users' integrated experiences when Human I/O is employed to enable common interface adaptation strategies.

6.1 Procedure

We conducted a user study with Human I/O Lite to understand its potential and challenges for real users in assisting them in managing incoming tasks during various SIIDs, when combined with common adaptation strategies. The study setup involved an egocentric camera (Logitech C930e), AR glasses (Nreal Light) for adaptive displays, and a custom website displayed on a touch screen tablet to simulate incoming tasks. The website was connected to the Human I/O system via a web socket.

We recruited 10 participants from Google (age range: 22–36, avg=28.0, std=3.95) with diverse backgrounds, including students, software engineers, research scientists, designers, product managers, and marketing analysts. Half of the participants reported familiarity with AR wearables (rating > 3 on a scale of 1 to 5).

Participants first familiarized themselves with Human I/O. They then simulated four scenarios in a lab space: working, washing hands, hair drying, and eating. Scenarios were selected to represent impairments in each of the vision, hands, hearing, and vocal channels. During each scenario, participants received a notification or a task on the tablet, including phone calls, text messages, and video watching. We designed corresponding adaptations strategies on Human I/O to provide adaptations when impairments were detected. For example, if the hands channel is detected to be affected, Human I/O will automatically display a phone call on AR glasses and prompt users to answer/reject the call by saying “yes” or “no”; if the hearing channel is affected, it will automatically turn on captions for a video. Details of all user study scenarios are shown in Table 2. Each scenario was conducted with and without Human I/O assistance in a within-subject design with counter-balanced order. All participants experienced four scenarios with and without Human I/O. Participants followed the adaptation strategies to overcome the situational impairments when Human I/O is activated. When Human I/O is not active, they were asked to address the incoming task and resume back to their previous task (e.g., pause hand washing and answer the phone call).

After each scenario, participants completed the NASA Task Load Index (TLX) questionnaire, assessing mental demand, physical demand, temporal demand, overall performance, effort, and frustration level on a 7-point scale (from 1–Lowest to 7–Highest). We also conducted a semi-structured interview to gather qualitative feedback on the participants' experiences.

6.2 Findings

Human I/O Enhances User Experience. Participants unanimously preferred the Human I/O experience. They found the system to be robust (P8, P9), accurate (P1, P2, P3, P6, P10), and helpful (P7, P8, P9) in various aspects. Participants reported that Human I/O significantly reduced their mental, physical, and temporal demand, as well as their effort and frustration level, and significantly improved their performance to manage incoming tasks during SIIDs

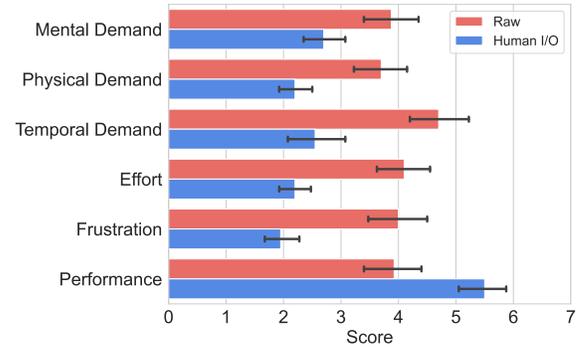


Figure 9: Participants' ratings to Task Load Index questions (on a scale of 1-low to 7-high) for their experience with SIIDs with and without Human I/O in user study. All rating differences are statistically significant with $p < 0.001$ via Wilcoxon signed-rank tests.

(Figure 9, all $p < 0.001$ via Wilcoxon signed-rank tests). Participants mentioned that when not using the system, they often had to “*delay my physical activities or my intent to interact*”. Participants highlighted the system's capability to support their workflow without interruption, as it allowed them to continue their current activities without delays. P4 noted that Human I/O enabled them to “*maintain [their] existing workflow and focus on the task at hand.*” This sentiment was echoed by P6, who appreciated the “*time-saving*” aspect of the system. P7 also emphasized the convenience and efficiency that Human I/O provided:

“It was much more convenient. Especially when my hand is captured by something else, then I'm still able to respond promptly to the request. It's essentially allowing me to do the thing that I want to do but easier and way faster.” – P7

Need for Personalization. Interestingly, we observed that for some tasks the perceived usefulness of Human I/O varied among participants. In particular, while some users found Human I/O less helpful for tasks that involved quickly interacting with phone notifications and then returning to work, others reported that the notifications severely interrupted their focus and thought processes (P2). The impairments in these cases are not severe enough to warrant the adaptations.

As another example, 6 participants mentioned that the hand-washing scenario was the most helpful because “*my hands are all wet and there's practically no way to touch the screen.*” (P4). However, P5 did not share this sentiment, stating that they were comfortable interacting with their device: “*If my I just wash my hands with water I think it's acceptable for me to click the screen.*”

In light of these varied responses, we see a need for a more personalized availability scale into Human I/O to better address individual preferences. Researchers and developers should note that it might be wrong to assume that users always want the UI to adapt based on their situations. Recognizing this difference in user preference underscores the importance of integrating different levels and options within Human I/O.

#	Scenario	Incoming Task	Impaired Channel	Adaptation
1	Working (Typing)	Receive a text message	Vision/Eyes	Display message on glasses
2	Hand Washing	Receive a phone call	Hands/Fingers	Switch to voice commands
3	Hair Drying	Watch a video	Hearing	Turn on captions
4	Eating Chips	Receive a phone call	Vocal	Suggest auto-reply

Table 2: Scenarios, incoming tasks, and corresponding adaptations participants experimented with in our simulation user study.

Participants also expressed desire in customizing the adaptation procedures, emphasizing that familiarity with these adaptations would make the system much more comfortable and effective to use. As one participant noted:

“When one input method is unavailable, I might have multiple alternatives, such as nodding or using voice commands. It would be even better if it could be customized to match my preferences.” – P10

Raising Awareness of SIIDs. One interesting finding from our study is that before experimenting Human I/O, many participants had not recognized the extent of the situational impairments affecting their daily lives. As a result, they would often deem many tasks unfeasible, give up on them, or seek alternative methods. P4’s comment illustrates this realization:

“Before it’s more like, there’s no way. I have to finish my hand washing and drying it out before I’m able to respond. It’s kinda impossible without the system. And the system gives me some new capabilities.” – P4

For the participants, Human I/O not only facilitated interaction with their devices but also served as an awareness tool, spotlighting the situational impairments they routinely encountered. This newfound awareness allowed users to extend their capabilities beyond their initial expectations, opening up new possibilities for interactions that they had not previously considered.

The insights from our user study involving 10 participants demonstrated the practical application and potential of Human I/O in enhancing user experience in the presence of SIIDs. Future research should explore an extensive deployment study that will involve a larger and more diverse group of participants, which can provide a more comprehensive understanding of the system’s utility across different user demographics and contexts in the long-term.

7 DISCUSSION AND FUTURE WORK

Privacy and Ethics. Maintaining privacy and upholding ethical standards are crucial in the design and deployment of SIIDs systems with active cameras and microphones.

We identify three main concerns: (1) Invasion of Privacy: Egocentric devices can unintentionally infringe on personal privacy by capturing and recording visuals and sounds without explicit consent. (2) Data Security and Storage: Using cloud servers for real-time machine learning model inferences requires rigorous measures like data anonymization, encryption, or on-device federated learning to forestall potential data breaches. (3) Inclusion, Bias, and Discrimination: The deployment of cameras, microphones, and LLMs may unintentionally exclude certain demographics or make inferences

based on race, gender, or other protected attributes, risking bias and discrimination.

Acknowledging these issues, we conducted user studies of Human I/O exclusively in controlled lab settings. It is imperative for subsequent researchers to prioritize privacy, ethical considerations, and the judicious use of technology.

While our system is exploratory, advancements in the field hint at promising solutions in future. The recent compact LLMs, such as LLAMA 7B and ALPACA, and fine-tuned models could facilitate on-device computations, ensuring data security. Recent developments by technology companies, exemplified by Apple’s Vision Pro, indicate a trend towards encryption and data anonymization protocols in everyday wearable devices. Furthermore, as HMD/AR devices gain prevalence, incorporating learnings from existing egocentric vision research is important. Such studies have already pioneered a variety of privacy-preserving techniques that we could incorporate.

Dimensions of availability. Our study introduces a four-level classification of channel availability, which has shown high accuracy in diverse scenarios. However, this classification could benefit from expansion beyond a single dimension. Future research should aim to develop a more nuanced understanding of channel availability, considering multiple dimensions such as the duration of the task, type of impairment, effort required in freeing up the channel, effort required in resuming the task, and the ability to multitask, etc.

For example, the dimension of duration highlights the variability in the length of time a channel is unavailable. Short-term unavailability, like a quick glance at a notification, may only need temporary adaptations. In contrast, longer durations of unavailability, such as during meetings, might require more substantial changes in user interface design. Another example is the ability to multitask. This dimension helps recognize situations where multitasking is feasible versus those where it may be disruptive. This understanding can inform the design of systems that better align with users’ capabilities and preferences, reducing the cognitive load and enhancing user experience.

Incorporating these multiple dimensions into the measurement of channel availability can provide users with finer controls over how their devices adapt to their changing needs and situations, and also offer developers with a more comprehensive range of options to consider and design for.

Incorporating More Sensing Techniques. The current implementation of Human I/O primarily relies on an egocentric view camera and microphone sensors, which may limit the system’s ability to accurately detect certain aspects of user interactions. For instance, it may not be able to determine if a user is brushing their

teeth, wearing headphones, or if their fingertips are wet. To enhance the capability and detection accuracy of Human I/O, future systems should incorporate additional sensing techniques.

One promising approach involves utilizing gaze tracking, available in many contemporary XR devices, to measure users' attention and gather more contextual information. Moreover, pupil diameter changes can be measured to estimate cognitive load [32]. While AR glasses or egocentric cameras provide a rich data source, their continuous usage throughout the day may not be desirable or feasible for users. Developing alternative, lower-resolution sensing methods that leverage mobile devices could be a more practical solution. For example, an approach similar to Google's Activity Recognition API [13] could be adapted to infer user availability based on patterns of device usage and motion data.

Dataset. Our prototype has demonstrated the potential for predicting channel availability using a few-shot chain-of-thought reasoning prompt. However, to develop a more robust, faster, and accurate model, and to establish a formal benchmark for evaluating situationally-aware systems, we recognize the need for a large-scale, extensive dataset. This dataset should encompass comprehensive features, similar to those found in the Ego4D dataset [15], while also incorporating lower-resolution features that can be obtained from everyday devices (e.g. smartphone IMU).

Toolkit. In our Human I/O prototype, we have implemented an extensible reasoning framework to infer SIIDs from multimodal sensors. As a proof-of-concept, we used egocentric cameras and microphones as input sources. By open-sourcing Human I/O, we hope to provide a toolkit that empowers XR and sensing researchers and developers to create more accessible systems. For example, by incorporating input from thermal cameras, eye-tracking cameras, depth sensing [8], Inertial Measurement Unit (IMU) signals, Ultra Wide Band (UWB) signals, street views for remote tourism [7], and the state-of-the-art sensing algorithms, Human I/O can be further expanded to create a holistic SIID detection framework for developing wearable applications.

Adaptation Strategies. While our paper primarily focuses on the detection of SIIDs, a complete self-adaptive system addressing SIIDs would also require the "adaptation policies", or ways to adjust the system's behavior given the detected impairments. However, adaptation strategies would require the efforts of a separate, comprehensive research project, which goes beyond the scope of this work.

An important question to investigate is whether a universal design can be achieved, and if we can develop algorithms to suggest appropriate adjustments based on user needs and contextual factors. Establishing robust evaluation frameworks, such as new metrics and user studies, will enable researchers and practitioners to assess the effectiveness of different adaptation approaches and develop systems that better accommodate user needs in various contexts.

We hope that the detection results from our system might serve as a starting point for app developers, aiding them in considering adaptation policies suited to their specific contexts.

A Situationally Aware Network. Another interesting direction for future research is to investigate the development of a situationally aware network that connects multiple devices. For example, a person is on a phone call while their spouse tries to dry their hair,

and a message appears on the spouse's device. The network, recognizing the ongoing phone call, could send reminders to prevent noise interference. This illustrates the potential of a collaborative system across multiple devices and multiple people that responds intelligently to a network of context.

8 CONCLUSION

In this paper, we presented a unified approach to detecting SIIDs based on the availability of human input/output channels. We shared insights from a formative study that guided the design of our system, emphasizing the importance of integrating contextual cues and proposing a four-level scale for measuring channel availability. Furthermore, we introduced Human I/O, a system that combines egocentric device, multimodal sensing, and large language models to predict channel availability. Our technical evaluation and user study demonstrated the effectiveness of Human I/O, and its potential in reducing user effort and improving performance in the presence of SIIDs. By abstracting SIIDs into channel availability, our work offers a step towards comprehensive detection of situational impairments; and based on this first step, we see an exciting future to build a general-purpose toolkit for SIIDs detection and enable developers to address a range of situational impairments in our daily lives.

ACKNOWLEDGMENTS

We would like to extend our thanks to Siyou Pei, Xiuxiu Yuan, Alex Olwal, Eric Turner, and Federico Tombari for providing assistance and/or reviews for the manuscript. We would also like to thank our reviewers for their insightful feedback.

REFERENCES

- [1] Djamila Romaiissa Beddiar, Brahim Nini, Mohammad Sabokrou, and Abdenour Hadid. 2020. Vision-Based Human Activity Recognition: A Survey. *Multimedia Tools and Applications* 79 (2020), 30509–30555. <https://doi.org/10.1007/s11042>
- [2] 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. <https://doi.org/10.5555/3495724.3495883>
- [3] Stuart K Card. 2018. *The Psychology of Human-Computer Interaction*. Crc Press.
- [4] Alexander Curtiss, Blaine Rothrock, Abu Bakar, Nivedita Arora, Jason Huang, Zachary Englhardt, Aaron-Patrick Empedrado, Chixiang Wang, Saad Ahmed, Yang Zhang, et al. 2021. FaceBit: Smart Face Masks Platform. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 4 (2021), 1–44. <https://doi.org/10.1145/3494991>
- [5] Jiangpeng Dai, Jin Teng, Xiaole Bai, Zhaohui Shen, and Dong Xuan. 2010. Mobile Phone Based Drunk Driving Detection. In *2010 4th International Conference on Pervasive Computing Technologies for Healthcare*. IEEE, IEEE, 1–8. <https://doi.org/10.4108/ICST.PERVASIVEHEALTH2010.8901>
- [6] Alan Dix, Janet Finlay, Gregory D Abowd, and Russell Beale. 2003. *Human-Computer Interaction*. Pearson Education.
- [7] Ruofei Du, David Li, and Amitabh Varshney. 2019. Geollery: A Mixed Reality Social Media Platform. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI, 685)*. ACM, 13. <https://doi.org/10.1145/3290605.3300915>
- [8] Ruofei Du, Eric Turner, Maksym Dzitsiuk, Luca Prasso, Ivo Duarte, Jason Dourgarian, Joao Afonso, Jose Pascoal, Josh Gladstone, Nuno Cruces, Shahram Izadi, Adarsh Kowdle, Konstantine Tsotsos, and David Kim. 2020. DepthLab: Real-time 3D Interaction With Depth Maps for Mobile Augmented Reality. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology (UIST)*. ACM, 829–843. <https://doi.org/10.1145/3379337.3415881>
- [9] Simon Frieder, Luca Pinchetti, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Christian Petersen, Alexis Chevalier, and Julius Berner. 2023. Mathematical Capabilities of ChatGPT. *ArXiv Preprint ArXiv:2301.13867* (2023). <https://doi.org/10.48550/arXiv.2301.13867>

- [10] Jon E Froehlich, Eric Larson, Tim Campbell, Conor Haggerty, James Fogarty, and Shwetak N Patel. 2009. HydroSense: Infrastructure-Mediated Single-Point Sensing of Whole-Home Water Activity. In *Proceedings of the 11th International Conference on Ubiquitous Computing*. 235–244. <https://doi.org/10.1145/1620545.1620581>
- [11] Mayank Goel, Leah Findlater, and Jacob Wobbrock. 2012. WalkType: Using Accelerometer Data to Accomodate Situational Impairments in Mobile Touch Screen Text Entry. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2687–2696. <https://doi.org/10.1145/2207676.2208662>
- [12] Mayank Goel, Alex Jansen, Travis Mandel, Shwetak N Patel, and Jacob O Wobbrock. 2013. ContextType: Using Hand Posture Information to Improve Mobile Touch Screen Text Entry. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2795–2798. <https://doi.org/10.1145/2470654.2481386>
- [13] Google. 2023. Google Activity Recognition API. <https://developers.google.com/location-context/activity-recognition>
- [14] Google. 2023. Object Detection Task Guide. https://developers.google.com/mediapipe/solutions/vision/object_detector
- [15] 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. <https://doi.org/10.1109/CVPR52688.2022.01842>
- [16] Bilge Günlül, A Murat Tekalp, and Peter J van Beek. 1998. Content-Based Access to Video Objects: Temporal Segmentation, Visual Summarization, and Feature Extraction. *Signal Processing* 66, 2 (1998), 261–280. <https://doi.org/10.1016/S0165>
- [17] Steve Hodges, Lyndsay Williams, Emma Berry, Shahram Izadi, James Srinivasan, Alex Butler, Gavin Smyth, Narinder Kapur, and Ken Wood. 2006. SenseCam: A Retrospective Memory Aid. In *UbiComp 2006: Ubiquitous Computing: 8th International Conference, UbiComp 2006 Orange County, CA, USA, September 17-21, 2006 Proceedings*. Springer, Springer, 177–193. https://doi.org/10.1007/1185356_11
- [18] Ellen Jiang, Edwin Toh, Alejandra Molina, Aaron Donsbach, Carrie J Cai, and Michael Terry. 2021. Genline and Genform: Two Tools for Interacting With Generative Language Models in a Code Editor. In *Adjunct Proceedings of the 34th Annual ACM Symposium on User Interface Software and Technology*. 145–147. <https://doi.org/10.1145/3474349.3480209>
- [19] Takeo Kanade and Martial Hebert. 2012. First-Person Vision. *Proc. IEEE* 100, 8 (2012), 2442–2453. <https://doi.org/10.1109/JPROC.2012.2205356>
- [20] Shaun K Kane, Jacob O Wobbrock, and Ian E Smith. 2008. Getting Off the Treadmill: Evaluating Walking User Interfaces for Mobile Devices in Public Spaces. In *Proceedings of the 10th International Conference on Human Computer Interaction With Mobile Devices and Services*. 109–118. <https://doi.org/10.1145/1409240.1409253>
- [21] Shian-Ru Ke, Hoang Le Uyen Thuc, Yong-Jin Lee, Jenq-Neng Hwang, Jang-Hee Yoo, and Kyoung-Ho Choi. 2013. A Review on Video-Based Human Activity Recognition. *Computers* 2, 2 (2013), 88–131. <https://doi.org/10.1109/CONFUENCE.2016.7508177>
- [22] Adil Mehmood Khan, Ali Tufail, Asad Masood Khattak, and Teemu H Laine. 2014. Activity Recognition on Smartphones via Sensor-Fusion and KDA-Based SVMs. *International Journal of Distributed Sensor Networks* 10, 5 (2014), 503291. <https://doi.org/10.1155/2014/503291>
- [23] Kris M Kitani, Brian D Ziebart, James Andrew Bagnell, and Martial Hebert. 2012. Activity Forecasting. In *Computer Vision—ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7–13, 2012, Proceedings, Part IV*. Springer, Springer, 201–214. <https://doi.org/10.1007/978-3-642-33765-15>
- [24] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large Language Models Are Zero-Shot Reasoners. *ArXiv Preprint ArXiv:2205.11916* (2022). <https://doi.org/10.48550/arXiv.2205.11916>
- [25] Gierad Laput, Karan Ahuja, Mayank Goel, and Chris Harrison. 2018. Ubicoustics: Plug-and-Play Acoustic Activity Recognition. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*. 213–224. <https://doi.org/10.1145/3242587.3242609>
- [26] Gierad Laput and Chris Harrison. 2019. Sensing Fine-Grained Hand Activity With Smartwatches. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13. <https://doi.org/10.1145/3290605.3300568>
- [27] Gierad Laput, Robert Xiao, and Chris Harrison. 2016. Viband: High-Fidelity Bio-Acoustic Sensing Using Commodity Smartwatch Accelerometers. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. 321–333. <https://doi.org/10.1145/2984511.2984582>
- [28] Gierad Laput, Yang Zhang, and Chris Harrison. 2017. Synthetic Sensors: Towards General-Purpose Sensing. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 3986–3999. <https://doi.org/10.1145/3025453.3025773>
- [29] Oscar D Lara and Miguel A Labrador. 2012. A Survey on Human Activity Recognition Using Wearable Sensors. *IEEE Communications Surveys & Tutorials* 15, 3 (2012), 1192–1209. https://doi.org/10.1007/978-3-031-24352-5_5
- [30] Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. 2022. Solving Quantitative Reasoning Problems With Language Models. *ArXiv Preprint ArXiv:2206.14858* (2022). <https://doi.org/10.48550/arXiv.2206.14858>
- [31] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping Language-Image Pre-Training With Frozen Image Encoders and Large Language Models. *ArXiv Preprint ArXiv:2301.12597* (2023). <https://doi.org/10.48550/arXiv.2301.12597>
- [32] David Lindlbauer, Anna Maria Feit, and Otmar Hilliges. 2019. Context-Aware Online Adaptation of Mixed Reality Interfaces. In *UIST '19: Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*. ACM. <https://doi.org/10.1145/3332165.3347945>
- [33] Xingyu Liu, Jun Zhang, Leonardo Ferrer, Susan Xu, Vikas Bahirwani, Boris Smus, Alex Olwal, and Ruofei Du. 2023. Modeling and Improving Text Stability in Live Captions. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA, 2023)*. ACM, 1–9. <https://doi.org/10.1145/3544549.3585609>
- [34] Xingyu "Bruce" Liu, Vladimir Kirilyuk, Xiuxiu Yuan, Peggy Chi, Xiang Chen, Alex Olwal, 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 (CHI)*. ACM. <https://doi.org/10.1145/3544548.3581566>
- [35] Xingyu "Bruce" Liu, Ruolin Wang, Dingzeyu Li, Xiang Anthony Chen, and Amy Pavel. 2022. CrossA11y: Identifying Video Accessibility Issues via Cross-Modal Grounding. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology (Bend, OR, USA) (UIST '22)*. Association for Computing Machinery, New York, NY, USA, Article 43, 14 pages. <https://doi.org/10.1145/3526113.3545703>
- [36] Steve Mann, James Fung, Chris Aimone, Anurag Sehgal, and Daniel Chen. 2005. Designing EyeTap Digital Eyeglasses for Continuous Lifelong Capture and Sharing of Personal Experiences. *Alt. Chi, Proc. CHI 2005* (2005). <https://doi.org/10.1007/978-3-319-07788-27>
- [37] Alexander Mariakakis, Mayank Goel, Md Tanvir Islam Aumi, Shwetak N Patel, and Jacob O Wobbrock. 2015. SwitchBack: Using Focus and Saccade Tracking to Guide Users' Attention for Mobile Task Resumption. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 2953–2962. <https://doi.org/10.1145/2702123.2702539>
- [38] Alex Mariakakis, Sayna Parsi, Shwetak N Patel, and Jacob O Wobbrock. 2018. Drunk User Interfaces: Determining Blood Alcohol Level Through Everyday Smartphone Tasks. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13. <https://doi.org/10.1145/3173574.3173808>
- [39] Lingfei Mo, Fan Li, Yanjia Zhu, and Anjie Huang. 2016. Human Physical Activity Recognition Based on Computer Vision With Deep Learning Model. In *2016 IEEE International Instrumentation and Measurement Technology Conference Proceedings*. IEEE, IEEE, 1–6. <https://doi.org/10.1109/I2MTC.2016.7520541>
- [40] Alex Olwal, Kevin Balke, Dmitrii Votintsev, Thad Starner, Paula Conn, Bonnie Chinh, and Benoit Corda. 2020. Wearable Subtitles: Augmenting Spoken Communication With Lightweight Eyewear for All-Day Captioning. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 1108–1120. <https://doi.org/10.1145/3379337.3415817>
- [41] 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. <https://doi.org/10.1145/3526113.3545616>
- [42] Manoj Plakal and Dan. Ellis. 2020. YAMNet. <https://github.com/tensorflow/models/tree/master/research/audioset/yamnet>
- [43] Danila Potapov, Matthijs Douze, Zaid Harchaoui, and Cordelia Schmid. 2014. Category-Specific Video Summarization. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part VI 13*. Springer, Springer, 540–555. <https://doi.org/10.1007/978-3-319-10599-35>
- [44] Tauhidur Rahman, Alexander Travis Adams, Mi Zhang, Erin Cherry, Bobby Zhou, Huaishu Peng, and Tanzeem Choudhury. 2014. BodyBeat: A Mobile System for Sensing Non-Speech Body Sounds. In *MobiSys, Vol. 14*. 2–594. <https://doi.org/10.1145/2594368.2594386>
- [45] Isidoros Rodomagoulakis, Nikolaos Kardaris, Vassilis Pitsikalis, E Mavroudi, Athanasios Katsamanis, Antigoni Tsiami, and Petros Maragos. 2016. Multimodal Human Action Recognition in Assistive Human-Robot Interaction. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, IEEE, 2702–2706. <https://doi.org/10.1109/ICASSP.2016.7472168>
- [46] Andrew Sears, Mark Young, and Jinjuan Feng. 2009. Physical Disabilities and Computing Technologies: An Analysis of Impairments. In *The Human-Computer Interaction Handbook*. CRC Press, 87–110. <https://doi.org/10.1201/9781410615862>
- [47] Hyun Soo Park and Jianbo Shi. 2015. Social Saliency Prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4777–4785. <https://doi.org/10.1109/CVPR.2015.7299110>
- [48] Krishna Srinivasan, Karthik Raman, Anupam Samanta, Lingrui Liao, Luca Bertelli, and Mike Bendersky. 2022. QUILL: Query Intent With Large Language Models Using Retrieval Augmentation and Multi-Stage Distillation. *ArXiv Preprint ArXiv:2210.15718* (2022). <https://doi.org/10.48550/arXiv.2210.15718>

- [49] Zehua Sun, Qihong Ke, Hossein Rahmani, Mohammed Benamoun, Gang Wang, and Jun Liu. 2022. Human Action Recognition From Various Data Modalities: A Review. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022). <https://doi.org/10.1109/TPAMI.2022.3183112>
- [50] Ying-Chao Tung, Mayank Goel, Isaac Zinda, and Jacob O Wobbrock. 2018. RainCheck: Overcoming Capacitive Interference Caused by Rainwater on Smartphones. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. 464–471. <https://doi.org/10.1145/3242969.3243028>
- [51] Bryan Wang, Gang Li, and Yang Li. 2022. Enabling Conversational Interaction With Mobile UI Using Large Language Models. *ArXiv Preprint ArXiv:2209.08655* (2022). <https://arxiv.org/pdf/2209.08655>
- [52] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of Thought Prompting Elicits Reasoning in Large Language Models. *ArXiv Preprint ArXiv:2201.11903* (2022). <https://doi.org/10.48550/arXiv.2201.11903>
- [53] Jacob O Wobbrock. 2019. Situationally Aware Mobile Devices for Overcoming Situational Impairments. In *Proceedings of the ACM SIGCHI Symposium on Engineering Interactive Computing Systems*. 1–18. <https://doi.org/10.1145/3319499.3330292>
- [54] Andy Zeng, Adrian Wong, Stefan Welker, Krzysztof Choromanski, Federico Tombari, Aavek 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). <https://doi.org/10.48550/arXiv.2204.00598>
- [55] Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang, and Matthias Grundmann. 2020. Mediapipe Hands: On-Device Real-Time Hand Tracking. *ArXiv Preprint ArXiv:2006.10214* (2020). <https://arxiv.org/pdf/2006.10214>
- [56] Zhongyi Zhou, Jing Jin, Vrushank Phadnis, Xiuxiu Yuan, Jun Jiang, Xun Qian, Jingtao Zhou, Yiyi Huang, Zheng Xu, Yinda Zhang, Kristen Wright, Jason Mayes, Mark Sherwood, Johnny Lee, Alex Olwal, David Kim, Ram Iyengar, Na Li, and Ruofei Du. 2023. InstructPipe: Building Visual Programming Pipelines With Human Instructions. <https://doi.org/10.48550/arXiv.2312.09672>

A A DECISION TREE FOR DETERMINING THE LEVEL OF CHANNEL AVAILABILITY

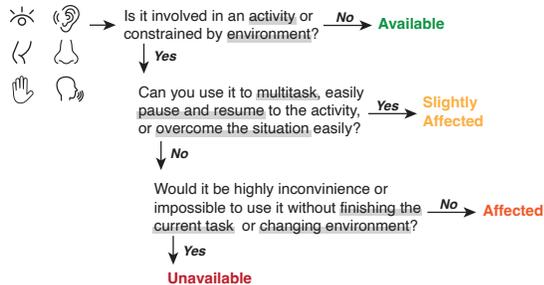


Figure 10: A decision tree based on our proposed four-level scale of channel availability.

We sketched a decision tree for determining the level of channel availability based on our proposed scale (Figure 10). We envision our approach may inspire future research to create more holistic and interactive labelling systems for SIIDs.

B HUMAN I/O IMPLEMENTATION DETAILS

We present implementation details of sensing volume, sound event, and brightness in the Human I/O system.

B.1 Direct Sensing: Object Detection

To determine if a hand H is holding an object O , we apply a rule-based method that considers the following criteria:

- (1) The confidence score of O is greater than 0.70.

- (2) The nearest distance between landmarks of H and the bounding box of O is less than 20 pixels.
- (3) The average distance between the thumb and the index, middle, ring, and pinky fingers is less than 0.25.
- (4) O 's predicted label is not "person".

These thresholds were determined empirically during system development and have demonstrated effectiveness on various objects. Our method prioritizes the minimization of false positives, resulting in a low recall but higher precision.

B.2 Direct Sensing: Volume Level

To sense the direct impact on the availability of the hearing channel, Human I/O measures the volume level of the environment. Specifically, we use the Web Audio API to access the user's microphone and the incoming audio stream. We maintain a buffer of the 20 most recent volume measurements to provide a continuous and smooth volume estimate. The computed volume levels are then converted to decibels:

$$\text{dB} = 20 \cdot \log_{10}(\text{volume}) + 100 \quad (1)$$

This allows for a more intuitive representation of the acoustic environment. We format the output as: "The environmental volume level is around <value> decibels."

B.3 Direct Sensing: Audio Classification

To provide a more comprehensive understanding of the audio environment, our system not only measures volume levels but also identifies specific audio events. We leverage YAMNet[42], a pre-trained deep neural network designed for audio event classification. YAMNet is capable of detecting 521 distinct audio event classes, including barking, laughter, siren, and silence, etc. During audio processing, we extracted the input buffer's channel data and fed it to the audio classifier. We stored the top three category names in an array, and output the top-1 audio class if the confidence score is greater than 0.70. We format the output as: "The environmental sound may contain <audio class>."

B.4 Direct Sensing: Brightness

To sense the direct impact on the availability of the vision/eye channel, Human I/O measures the brightness of the environment. We utilize relative luminance, a metric accounting for human perception of brightness. We follow the WCAG³ accessibility guideline to calculate luminance. Specifically, we first convert each frame to sRGB color space and calculate luminance using Rec. 709 coefficients:

$$Y = \frac{1}{255} (0.2126 * R + 0.7152 * G + 0.0722 * B) \quad (2)$$

We obtain a luminance value (Y) for each pixel, and compute the average luminance across all pixels, resulting in a single value representing perceived brightness in the range of 0 to 1. A value close to 0 indicates a very dark environment, while a value close to 1 indicates a very bright environment. A value around 0.5 suggests a medium brightness level. Similar to volume level, we applied temporal smoothing by averaging 20 luminance values in the past

³WCAG: https://www.w3.org/WAI/GL/wiki/Relative_luminance

second. We format the output as: “*The luminance value of the current environment is <value>, in the range of 0 to 1.*”

C PROMPTS FOR LARGE LANGUAGE MODELS

We elaborate on the detailed prompts for reasoning on user activity, their environment, and predicting human input/output channel availabilities.

C.1 Activity

“An egocentric view of User is showing” +
<BLIP-2 output> +

“Describe what User is doing briefly and objectively, as concisely as possible, without guesses or assumptions. Answer in the format of ‘User is...’. If it seems that User is not doing anything, answer ‘User is not doing anything’. If it cannot be inferred, answer ‘Unsure’.”

C.2 Environment

“An egocentric view of User is showing” +
<BLIP-2 output> +

“What location or environment is User likely to be in? Answer in the format of ‘User is in...’. If it cannot be inferred, answer ‘Unsure’.”

C.3 Channel Availability Prediction

C.3.1 Full Version.

[Instruction]

Available: The channel is currently not involved in any activity, or constrained by any environmental factors. It takes low to zero effort to use the channel to do a new task. Example: A user is sitting at their desk with their hands free, eyes not engaged in any task, and no background noise interfering with their hearing or speech.

Slightly Affected: The channel is engaged in an activity or constrained by an environmental factor. Given a new task that requires the channel, users can multitask, easily pause and resume to the current activity, or easily overcome the situation. Example: A user is holding a remote control, which can be quickly put down to free up their hand for another task.

Affected: The channel is involved in an activity or constrained by an environmental factor. Given a new task, the user may experience inconvenience or require effort to use the channel. Example: A user is carrying grocery bags in both hands, making it challenging to use their hands for other tasks without putting the bags down first.

Unavailable: The channel is completely unavailable due to an activity or environmental factor, and the user cannot use it for a new task without substantial adaptation or changing the environment. Example: A user is attending a loud concert, making it impossible for them to hear incoming notifications or carry on a conversation without stepping outside.

Given the current activity and environment as described below, what are the availability of C’s vision/eye, hearing, vision, and hands/fingers channels? For each channel provide reasons first and then answer using the scale defined above: available, slightly affected, affected or unavailable.

Separate each channel prediction with a semicolon (;).

<Few-shot Example 1>

Q: C is washing dishes in a kitchen sink. C is in a kitchen. C’s hand is washing dishes. The environmental volume is around 40 dB.

A: Let’s think step by step.

Eye Reasoning: While C is washing dishes, their eyes are partially occupied with the task at hand, which involves looking at the dishes to ensure they are clean. However, they can still glance away or multitask to some extent, so their vision is slightly affected but severely impacted.

Eye: Slightly Affected;

Hearing Reasoning: Washing dishes in the kitchen sink does not affect C’s hearing. They can still hear other things happening around them.

Hearing: Available;

Vocal Reasoning: C’s speech and voice are not impaired by washing dishes. They can still talk or communicate with others while performing the task.

Vocal: Available;

Hand Reasoning: As C’s hand is currently engaged in washing dishes, it is not available for other tasks. If C needs to use their hand for something else, they would need to stop washing dishes and wipe their hands. Hand: Not Available;

[ANSWER COMPLETED]

<Few-shot Example 2>

Q: C is playing an acoustic guitar in a room. C is in a small, cozy room with minimal furniture and decorations. C’s hand is playing an acoustic guitar. The environmental volume is around 58 dB.

A: Let’s think step by step.

Eye Reasoning: C is playing an acoustic guitar, which requires some attention to the placement of fingers on the frets and possibly looking at the sheet music or chords. However, their eyes are not entirely preoccupied with the task and can still be used for other tasks with some level of efficiency, by pausing playing guitar and potential put it away.

Eye: Affected;

Hearing Reasoning: As C is playing an acoustic guitar in a small room, the sound from the guitar is likely to be more noticeable. However, the environmental volume is low, which means that although their hearing may be somewhat affected by the sound of the guitar, they should still be able to hear other things, especially if they are loud or distinct.

Hearing: Affected;

Vocal Reasoning: Playing the guitar does not directly involve using one’s voice or speech, so C should be able to use their voice for other tasks while playing the guitar. However, their focus may be divided between playing the guitar and speaking, which could affect their ability to fully concentrate on either task.

Vocal: Available;

Hand Reasoning: C’s hand is actively engaged in playing the acoustic guitar. Using their hand for other tasks while playing the guitar would require some effort, as it would require them to stop playing the guitar.

Hand: Affected;

[ANSWER COMPLETED]

<Few-shot Example 3>

Q: C is working at a desk with a laptop. C is in a library. C's hand is typing on a computer. The environmental volume level is around 42 dB.

A: Let's think step by step.

Eye Reasoning:

C is currently using their eyes to focus on the laptop screen in front of them. While their attention is primarily on the laptop, they still have the ability to momentarily glance at other visual stimuli in their environment. However, their ability to focus on other tasks requiring visual attention may be somewhat affected.

Eye: Affected;

Hearing Reasoning:

The environmental volume level is low, which means that C is not experiencing any significant auditory impairment. They should be able to hear other things happening around them without much difficulty.

Hearing: Available;

Vocal Reasoning:

C is in a library, which typically has rules about maintaining a quiet environment. While their voice is physically available, using it for other tasks may be considered inappropriate or disruptive in this setting. Therefore, their ability to use their speech or voice for other tasks is situationally affected.

Vocal: Affected;

Hand Reasoning:

C is currently using their hands to interact with the laptop, such as typing or using the touchpad. They may be able to briefly use their hands for other tasks, but their ability to focus on other hand-related tasks might be affected while they are engaged with the laptop.

Hand: Affected;

[ANSWER COMPLETED]

[Current Context (example)]

Q: C is working on a piece of wood. C is in a workshop or a carpentry studio. C's hand is cutting a piece of wood. The environmental volume level is loud.

A: Let's think step by step.

C.3.2 Lite Version. In the lite version, we remove all reasoning steps in the few shot examples, as well as "let's think step by step.":

[Few-shot Example 1]

Q: C is washing dishes in a kitchen sink. C is in a kitchen. C hand is washing dishes. The environmental volume is around 40 dB.

A:

Eye: Slightly Affected;

Hearing: Available;

Vocal: Available;

Hand: Not Available;

[ANSWER COMPLETED]

[Few-shot Example 2]

Q: C is playing an acoustic guitar in a room. C is in a small, cozy room with minimal furniture and decorations. C's hand is playing an acoustic guitar. The environmental volume is around 58 dB.

A: Eye: Affected;

Hearing: Affected;

Vocal: Available;

Hand: Affected;

[ANSWER COMPLETED]

[Few-shot Example 3]

Q: C is working at a desk with a laptop. C is in a workspace or office environment. C's hand is typing on a keyboard. The environmental volume level is around 42 dB.

A: Eye: Affected;

Hearing: Available;

Vocal: Available;

Hand: Affected;

[ANSWER COMPLETED]

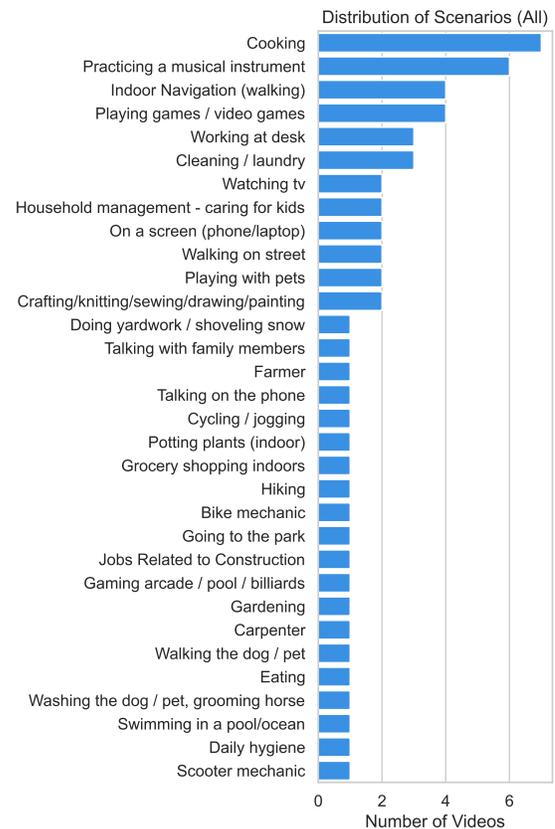


Figure 11: Distribution of the scenarios in our evaluation dataset that contains 60 egocentric video recordings from Ego4D dataset [15].

D EVALUATIONS

D.1 Technical Evaluation

D.1.1 Distribution of Scenarios. Figure 11 the full distribution of the scenarios of the video recordings used in our technical evaluation.

D.2 User Study

D.2.1 Task Load Index Questions. We list the set of task load index questions used in Section 6. After each task, we ask (from 1-very low, to 7-very high):

- (1) Mental Demand: How much mental and perceptual activity was required? Was the task easy or demanding, simple or complex?
- (2) Physical Demand: How much physical activity was required? Was the task easy or demanding, slack or strenuous?
- (3) Temporal Demand: How much time pressure did you feel due to the pace at which the tasks or task elements occurred? Was the pace slow or rapid?
- (4) Effort: How hard did you have to work (mentally and physically) to accomplish your level of performance?
- (5) Frustration Level: How irritated, stressed, and annoyed versus content, relaxed, and complacent did you feel during the task?
- (6) Overall Performance: How successful were you in performing the task? How satisfied were you with your performance?