

**From the holdings of the Department of Special Collections , Princeton University Library,
Princeton, NJ**

These documents are either in the public domain, or can only be used for educational and research purposes (Fair use) as per U.S. Copyright law (text below). By accessing this file, all users agree that their use falls in the public domain or within fair use as defined by the copyright law. We do not charge any permission or usage fees for the publication of images of material in our collections, including those provided by Princeton University Library, via our website, catalog, or directly from staff. For more information regarding copyright, credit, and citation please review the information on the Copyright, Credit and Citation Guidelines page of our website.

<https://library.princeton.edu/special-collections/policies/copyright-credit-and-citation-guidelines>. If you have any questions, including how to order higher-quality images from the collection, contact Special Collections Public Services via our Ask Us! Form. <https://library.princeton.edu/special-collections/ask-us>

Master's Theses, 2022-2026

Computer Science, 2023

2025

Effort, No Progress: Formulating Implicit Movements for Just-In-Time Instructions

Desta, Beza T

Princeton, NJ : Princeton University

<http://arks.princeton.edu/ark:/88435/dsp01hq37vs01z>

U.S. Copyright law: The copyright law of the United States (Title 17, §108, United States Code) governs the making of photocopies or other reproductions of copyrighted material. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship, or research. If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of fair use that user may be liable for copyright infringement. This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve a violation of copyright law. <https://www.copyright.gov/title17/title17.pdf>

EFFORT, NO PROGRESS: FORMULATING IMPLICIT MOVEMENTS FOR JUST-IN-TIME INSTRUCTIONS

BEZA DESTA

ADVISOR: PROFESSOR PARASTOO ABTAHI

A MASTER'S THESIS PRESENTED TO THE FACULTY OF PRINCETON UNIVERSITY
IN CANDIDACY FOR THE DEGREE OF MASTERS IN SCIENCE OF ENGINEERING

DEPARTMENT OF COMPUTER SCIENCE

PRINCETON UNIVERSITY

MAY 2025

I hereby declare that I am the sole author of this thesis.

I authorize Princeton University to lend this thesis to other institutions or individuals for the purpose of scholarly research.

Beza Desta

I further authorize Princeton University to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

Beza Desta

Abstract

When observing the execution of manual tasks, instructors provide follow-up instructions instead of waiting for mistakes to occur or for users to explicitly ask for help [37]. We manually coded 500 video clips—captured from an egocentric view of the instructee—to identify the implicit cues that prompt these interventions. Our annotations revealed a systematic classification of uncertainty-insinuating cues expressed through hand and head movements [4]. Based on this finding, we propose a just-in-time instruction system that tracks user and object movements to emulate two groups of implicit cues: goal-agnostic cues and goal-dependent cues. We implemented three methods to encode the most frequently observed cues: for the first group, we detect repeated movements of the head or of the objects currently being manipulated; for the second group, we leverage measures of task progress and error to infer uncertainty. By identifying when progress is suboptimal, our system can infer that the user is uncertain and deliver timely instructions.

Acknowledgements

I am eternally grateful to the army of friends, colleagues and mentors that formed in Princeton for keeping me going.

I am incredibly thankful to have had an advisor, Prof. Parastoo Abtahi, who was (and is) intentional in investing so much in my growth as a researcher.

I specifically want to thank Andrés Monroy-Hernández, Anja Kalaba, Kiyosu Maeda, Ching-Yi Tsai, Mo Kari and Lauren Wang for all the thoughtful feedback they have given on this work at different times.

Carrie Louer, Esther Hayter, Morgan Murdock, Hayat Shibeshi, Tigist Mebrate, Libby Feray, Chris Branner-Augmon, and Jensens, thanks for making my time at Princeton full of joy.

Last but not least, my family, whose prayers are the reason I reach the end.

To my Father in Heaven

Contents

Abstract	iii
Acknowledgements	iv
List of Tables	viii
List of Figures	ix
1 Introduction	1
2 Related Works	4
2.1 Mind in Motion	5
2.2 Just-In-Time Instruction Giving	6
2.2.1 Reactive	6
2.2.2 Proactive	7
2.3 Progress and Effort Approaches	8
3 Preliminary Exploration and Takeaways	10
3.1 Our approach	10
3.1.1 Dataset	10
3.1.2 Coding	11
3.2 Goal-Agnostic Cues and Goal-Dependent Cues	12
3.2.1 Goal-Agnostic Cues	12
3.2.2 Goal-Dependent Cues	12

4	System Implementation	17
4.1	Method	17
4.1.1	Apparatus	17
4.1.2	Heuristic Based System for Goal-Agnostic Cues	17
4.1.3	Progress and Effort Quantification for Goal-Dependent Cues .	19
5	Application	23
5.1	Other assembly tasks	23
5.2	Navigation	24
6	Limitations and Future Work	25

List of Tables

List of Figures

3.1	Initial Findings of General Cues	14
3.2	Broad groupings of the implicit cue codes	14
3.3	Goal-Agnostic Cues	15
3.4	Goal-Independent Cues (b) Example of object during a task.	15
3.5	The frequency of instances Goal-dependent cues were observed during our coding of 500 video clips	15
3.6	Goal-Independent Cues (b) Example of object during a task.	15
3.7	Goal-Dependent Cues	16
3.8	This structured breakdown visually and conceptually distinguishes be- tween the intended task flow, general signals of uncertainty, and specific action-based cues that can inform timely instructional support.	16
3.9	Goal-Dependent Cues Frequency	16
3.10	The frequency of instances Goal-dependent cues were observed during our coding of 500 video clips	16
4.1	(a).OptiTrack Camera Setup	18
4.2	(b).object Tracking	18
4.3	(a) The OptiTrack camera setup used for tracking. (b) Example of object during a task.	18

Chapter 1

Introduction

People often prefer not to read instructions [30]. Many consumers ignore instructions altogether or find themselves with extra parts after they think they are done [11]. For example, a market survey of public call boxes in the late 20th century showed that approximately 75 % of users did not refer to the instructions [32]. Several studies have explored why people tend to ignore instructions [38]. In a seminal work, Wright et al. identified unstructured manuals as a major issue: most instruction sets assume readers will start at the beginning and read through each step sequentially, memorizing them. In reality, people prefer to skip around and only consult a manual when they have a specific question. Moreover, human instructors intuitively determine when someone engaged in a task needs guidance. Although the need for a system that delivers the right instruction at the right time has long interested HCI researchers, most existing systems are reactive and require explicit input from the user. Such systems can be cumbersome—or even impossible—to use, depending on the task. Consequently, researchers have shifted focus toward proactive instruction-giving systems that rely on implicit inputs such as the environment, past interactions, and user physiological data [39, 26]. The literature has investigated various implicit methods—some of which we observed in our preliminary explorations—such as gaze direction [14], constant time

delays [15], user context [27], pupillary response [16], and other physiological signals [26]. However, many of these cues require multiple sensors and are impractical in real-world settings, limiting their generalizability.

Our work is inspired by psychological research showing that people use motion (e.g., hand gestures, body movements) to think through complex cognitive tasks [34]. These observations were echoed by Wang et al. in the HoloAssist dataset, where instructors frequently provided proactive guidance based on perceived implicit inputs [37]. Motivated by this, we conducted a manual preliminary exploration of 500 egocentric video clips from HoloAssist to identify the implicit cues that signal an instructee’s need for assistance. We found that uncertainty often manifests in hand and head movements: repeated reach-and-retract hand motions or side-to-side head movements implied that the user was unsure how to proceed, prompting instructors to offer additional or more detailed guidance.

We developed a taxonomy of implicit cues from the HoloAssist dataset, classified them into descriptive codes, and counted their frequency. Two major groups emerged—goal-dependent cues and goal-agnostic cues—which appeared most frequently. From these, we selected the most common behavioral signals of uncertainty (e.g., repeated reach-retract motions) to inform our system design.

- **Goal-agnostic cues:** We implemented a heuristic approach that triggers instructions based on users’ hand and head movements alone.
- **Goal-dependent cues:** We devised a principled decision mechanism that translates task progress and error signals into intervention decisions.

We frame instruction timing as an optimal decision problem: given the user’s current state—quantified by progress (task advancement) and effort (kinematic cost)—should the system intervene now or wait? To solve this, we propose a reinforcement-learning-inspired formulation where:

- Progress acts as reward (calculated via alignment between current and goal states),
- Effort acts as cost (measured by hand/head movement velocity).

The system intervenes when the ratio of progress to effort falls below a threshold (i.e., the user expends energy without advancement). This ensures instructions are triggered not only by observed cues (e.g., hesitation) but by their functional impact on task performance. Crucially, our approach requires only an egocentric view—no facial or eye tracking—and can generalize across diverse tasks in real-world settings.

We quantify progress by measuring the alignment difference between the current and goal states of task components over time. Effort is quantified by the velocity of hand and head movements. An upward-trending progress curve indicates clarity, while a downward-trending or fluctuating curve signals uncertainty and the need for guidance. By monitoring these values, our system determines the ideal timing for instruction delivery.

Our proactive approach—triggering guidance based on quantified progress—eliminates the need for users to make overt errors before receiving help. Additionally, by delivering instructions only when needed, our system may aim to reduce cognitive load and error rates while allowing users to access guidance in a need-based manner.

Our main research contributions are

- A codebook of implicit cues and themes of implicit cues of instructees prior to receiving instruction
- A formulation of implicit cues by quantifying progress of task and effort of user
- A just-in-time uncertainty detection system that can be used to deliver relevant instructions

Chapter 2

Related Works

Prior work has explored instruction design [1], instruction presentation [25], and instruction adaptation [39] for a plethora of users and user goals [36] and yet there seems to be a gap in the literature on a dedicated exploration on giving instructions at the right time.

Some works have addressed this gap by building around the idea of a context-aware system. However the context addressed pools around two major areas. One is that the context detected usually requires a number of information that requires high speed communication, well-performing visual models, and/or sensing components[21, 20, 26]. And the second is that it relies on the assessment that a complete and discrete step has been executed via detection or user self-reporting[20, 39, 18]. We posit that depending on what the system relies on to determine the timing of instruction we can classify previous works as *proactive* and *reactive*.

Our work builds on the vast literature that has explored ways of giving instructions in real-time. Tasks that physically involve a user in any capacity such as putting make-up on [33] or dancing [2] have been explored on providing instructions or guidance in real time using Augmented Reality(AR). But simply using any kind of AR to complete tasks does not result in decreased time of completion or error rate. Funk et al.,[10],

found that using Head-Mounted Displays as a form of AR to communicate instructions actually led to users performing worse than baseline paper-back instructions. When Baumister et al., further explored as to why this is the case, they found that what contributes to the helpful nature of other forms of AR instructions is their affordance to make instructions available with less extraneous effort exerted by the user.

As we observed in the dataset of 166 hours of human-human instructions of bi-manual tasks collected by HoloAssist [37], we know that humans provide instructions proactively before the individual who is performing the task needs to explicitly ask for one. Hence, we believe that it is important to look to other fields of psychology, anthropology, and neuroscience to understand how that comes to be.

2.1 Mind in Motion

Decision and action are interwoven: motor areas both receive accumulating evidence and feed back into the decision process, so that choice is realized through the dynamics of ongoing movements rather than being fully formed before action onset[31]. Moments of indecision don't just slow the mind—they literally interrupt the body's movement. Early researchers noticed that rats would occasionally pause at choice points in mazes and look back and forth as though confused about which way to proceed. They termed this behavior "vicarious trial and error" (VTE), speculating that the rats were mentally exploring potential future options rather than simply executing preprogrammed actions. More-recent experiments examining VTE in rats suggest that there are direct parallels to human processes of deliberative decision making, working memory and mental time travel. [28]. This is a concept used by multiple existing works when determining when to give instructions[15, 39, 18].

However uncertainty manifests itself in more ways than just pauses. Under target uncertainty, movement paths exhibit increased spatial variability as multiple potential

actions compete during execution. For example: Reaching movements show zigzag trajectories when deciding between ambiguous targets. [19]. This further portrays Tversky’s work in ”spraction.” Hand movements during speech aren’t just expressive- they actively organize ideas [34]. When multiple potential actions are spatially viable (e.g., reaching for two targets), the brain generates competing motor plans for each hand simultaneously. This creates a ”desirability-driven” neural competition where uncertainty emerges from unresolved plan conflicts. Behavioral signature: Movements exhibit spatial averaging (e.g., initial paths between targets) as the sensorimotor system delays commitment to a single plan until evidence thresholds are met. [7, 12].

2.2 Just-In-Time Instruction Giving

Just-in-time instruction refers to delivering guidance at the precise moment it is needed during task execution, rather than all at once or beforehand. Prior research has explored a variety of strategies for timing and triggering instructional interventions. We classify these approaches into two broad categories-reactive and proactive-each with further subtypes based on their triggering mechanisms.

2.2.1 Reactive

Reactive instruction is delivered in response to user actions or explicit requests, typically after an event has occurred[6]

- **Step-Completion or Error:** We see such assistance in a number of everyday tasks such as grammar and spelling checks, or error and warning flags or Copilot auto-completions in coding IDEs. Such systems highlight next steps post-task completion to align with pre-defined ”correct action” sequences, reducing user dead-ends.

- **User-initiated:** Early and foundational systems, such as UNIX man pages or DOS help, are classic examples where users explicitly request information about specific commands or features. These systems are effective for users who know what they need to look for, offering quick reference without interrupting the workflow[6]. Advanced systems maintain a user model to personalize the help experience. These systems can deliver information that is new and relevant to the user’s current goal or knowledge state, but still rely on explicit user requests to trigger help. [29]

2.2.2 Proactive

Proactive instruction systems anticipate the user’s needs and deliver guidance before errors or explicit requests occur. We have seen multiple ways of implementing this approach in HCI and psychology literature, which describes proactive systems as those that “do not necessarily await explicit human input but rather initiate and drive actions based on observed behavior, sensor readings, or contextual cues” [35]. These systems often monitor user attention, behavior, and environmental factors to predict intentions and intervene at optimal moments, thereby preventing mistakes and supporting seamless human-computer collaboration.

- **User-Based:** User-based proactive instruction relies on real-time monitoring of user behavior detecting repeated errors, lengthy pauses, prompting timely interventions [39, 15]. Advanced interfaces can monitor eye gaze, heart rate, or EEG to infer cognitive load or attention lapses, proactively offering help when overload is detected [26, 13]. Proactive systems can further adapt instructions based on user expertise, learning progress, or prior interactions, ensuring relevance and reducing unnecessary interruptions. [8, 3].
- **Environment-Based:** Environment-based proactive instruction is triggered

by changes or cues in the task environment or context rather than the user’s direct actions. Systems monitor the workspace for events such as the introduction of new tools, changes in task state, or the presence of hazards, and provide relevant instructions automatically [5, 9]. Instructions may also be delivered when users enter specific locations (e.g., AR overlays near machinery) or reach critical moments in a workflow (e.g., before a complex assembly step) [15, 13, 14]. Additionally, proactive prompts can be triggered when required resources become available or when environmental constraints change [21].

2.3 Progress and Effort Approaches

Progress tracking is a critical component in robotics, particularly for tasks involving assembly, manipulation, and interaction with dynamic environments. Researchers have developed various methods to model, predict, and optimize task progress, integrating instruction comprehension and real-time trajectory adjustments. For instance, robots can leverage visual progress representations to monitor task advancement and adaptively select subsequent subgoals, as demonstrated by Wang et al. [17]. This approach significantly enhances manipulation policy effectiveness in long-horizon tasks. Recent advancements also highlight the use of vision-language models to predict temporal progress from visual trajectories, facilitating learning and adaptation across diverse tasks and domains [22]. These methods demonstrate the potential of combining progress tracking with decision-making frameworks and integrating real-time feedback and adaptability.

While progress tracking in robotics has made significant strides, modeling human behavior, however, poses unique challenges. Unlike robots, humans are free agents with diverse and unpredictable behaviors. In robotics, algorithms such as maximum entropy inverse optimal control [40] have been employed to model behavior and

decision-making under uncertainty. These methods extract patterns from task execution, providing insights into intention and progression. Similarly, fuzzy-rule-based systems [41] have been used to predict human intention in collaborative settings, such as human-robot handovers, by combining wearable sensing data with decision models. However, these methods often rely on resource-intensive computations and lack the ability to generalize across diverse environments and users, leaving gaps in their applicability to real-world, dynamic, and user-specific scenarios.

While robotic systems benefit from precise sensors, deterministic algorithms, and predictable environments, human users are influenced by cognitive states, distractions, and emotional factors that complicate traditional progress tracking. Current systems often fail to see *where* to look in order to capture nuanced human uncertainty and are limited in their ability to adaptively provide support based on real-time, non-verbal cues.

Chapter 3

Preliminary Exploration and Takeaways

3.1 Our approach

We aimed to observe the implicit actions that users make between any completed steps that indicate uncertainty and a need for an intervention from an instructor. We hypothesize that these defined actions would be the indicators for an automated system that a user is uncertain or confused about what the current or following step should be in the assembly process and hence needs help.

3.1.1 Dataset

We constructed a CodeBook based on a subset of an egocentric dataset of two people collaborating to build one of four types of furniture. We resourced 1110 clips from 329 videos, that had according to the annotations of the dataset, a follow-up intervention. A *follow-up intervention* is any instance in the annotations that had a conversation purpose labeled: "an instructor follow-up instruction." Such conversations have this label when an instructor contributes an instruction without an explicit input from the

student. We clipped each instance of a follow-up intervention 5 seconds prior to when it starts and resulted in 1110 viable clips. We hypothesize that these instances would indicate what implicit cues, if any, triggered the instructors to proactively intervene.

3.1.2 Coding

In order to code the 500 clips, we first picked 100 video clips to formulate our initial labels for cues that possibly instigated the interventions. Out of the 100 video clips, the first 50 video clips were sequentially selected and the latter 50 video clips were randomly selected. We provide the labels for each implicit cue and their descriptions in the supplementary material.

Once we coded the 500 video clips for the observed cues, we were able to recognize 415 cues that were implicit that caused the instructor to intervene in order to give a verbal guidance and we categorized these cues in 6 broad categories as shown in Figure 3.2. The other 85 clips were coded as *unknown* or *explicit cues* after an agreement between our 2 main coders. (Figure 3.1)

The signals in the least observed categories of *interjection*, *gesture*, and *inaction* were, in fact, familiar as they were expressive and/or have been used in previous works[15, 39]. Mini-Errors were a category of cues that were associated with the state of the pieces involved in the task that would cause inconvenience or error in the long run. While these cues were somewhat prominent, we also believed that such cues were a deviation from the cues we initially set out to find. Finally, we were able to identify another group of cues that were actually more frequent and further proved what we have learned from the field of embodied cognition. And these cues fell into 2 categories: Goal-Independent Cues and Goal-Dependent Cues.

3.2 Goal-Agnostic Cues and Goal-Dependent Cues

Building on the embodied cognition framework introduced in 2.1 where decision-making emerges through sensorimotor engagement – our preliminary exploration revealed two distinct classes of uncertainty signals that drive proactive instruction timing. These categories align with Tversky’s ”spraction” principle, where spatial-action dynamics simultaneously reflect and shape cognitive states [34].

3.2.1 Goal-Agnostic Cues

Generalized behaviors that indicate uncertainty regardless of the specific assembly step. These include repeated rotation (A2-1), repeated shifting (A2-2), slow hesitant movements (A2-3), bringing a part closer for inspection (A2-4), and repeated shaking (A2-5). Such actions reflect underlying cognitive deliberation or confusion and are not tied to a particular assembly goal. Instead, they represent embodied manifestations of indecision, as users physically explore or reconsider their options before proceeding. These cues manifest as **generalized motor hesitations** observable across tasks, independent of specific assembly goals (Figure 3.3). We observed such Goal-Agnostic cues a total of 112 times in the 500 video clips we assessed. A break down of the frequency of each cue is show in 3.5.

3.2.2 Goal-Dependent Cues

Figure 3.7 illustrates goal-dependent cues-behaviors that signal uncertainty or error in direct relation to the current assembly objective. Blue-framed panels (B1-1 to B1-4) depict actions such as reaching toward an incorrect piece, retracting from correct or incorrect contact, and aborting an incorrect reach. Green-framed panels (B2-1 to B2-4) show picking up the wrong piece, moving toward an incorrect state, reaching with incorrect orientation, and progressing toward an inconvenient or suboptimal

state. These cues are tightly coupled to the task context and often precede instructor interventions, as they directly impact the likelihood of successful task completion. We observed such Goal-Agnostic cues a total of 169 times in the 500 video clips we assessed. A break down of the frequency of each cue is show in 3.9

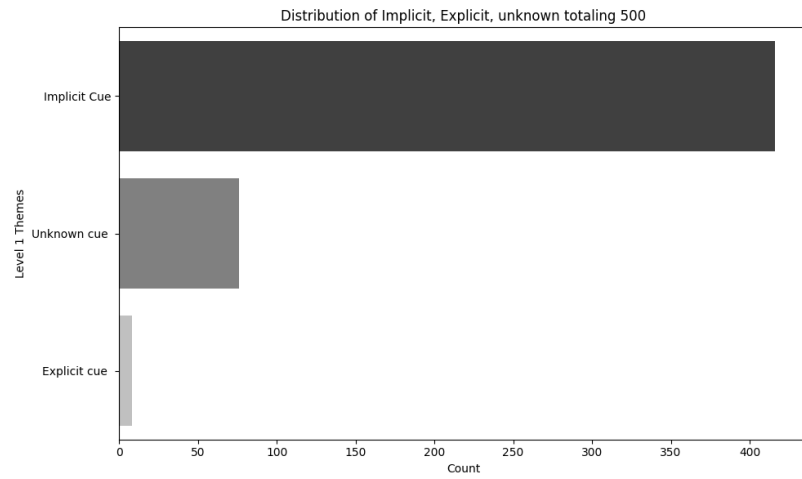


Figure 3.1: Initial Findings of General Cues

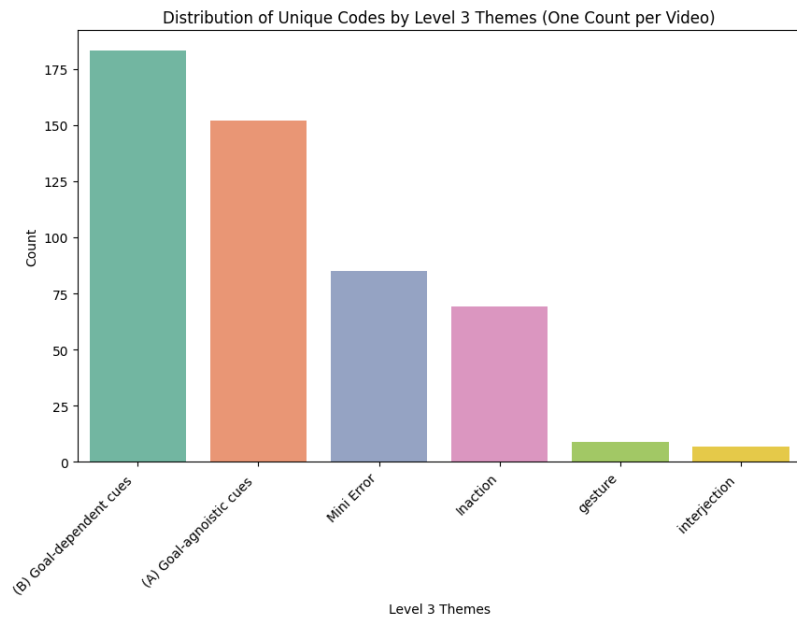


Figure 3.2: Broad groupings of the implicit cue codes

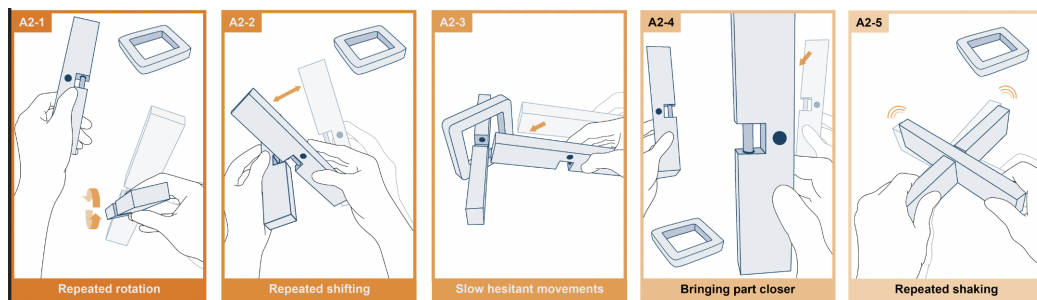


Figure 3.3: Goal-Agnostic Cues

Figure 3.4: Goal-Independent Cues (b) Example of object during a task.

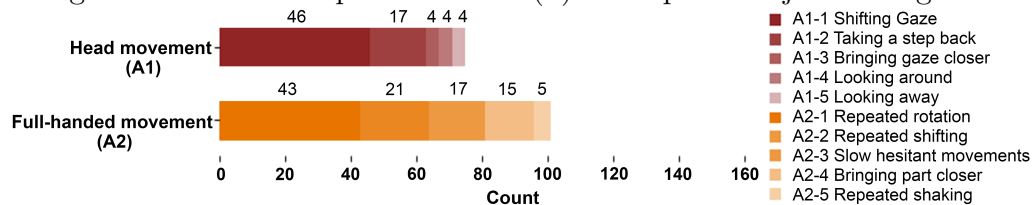


Figure 3.5: The frequency of instances Goal-dependent cues were observed during our coding of 500 video clips

Figure 3.6: Goal-Independent Cues (b) Example of object during a task.

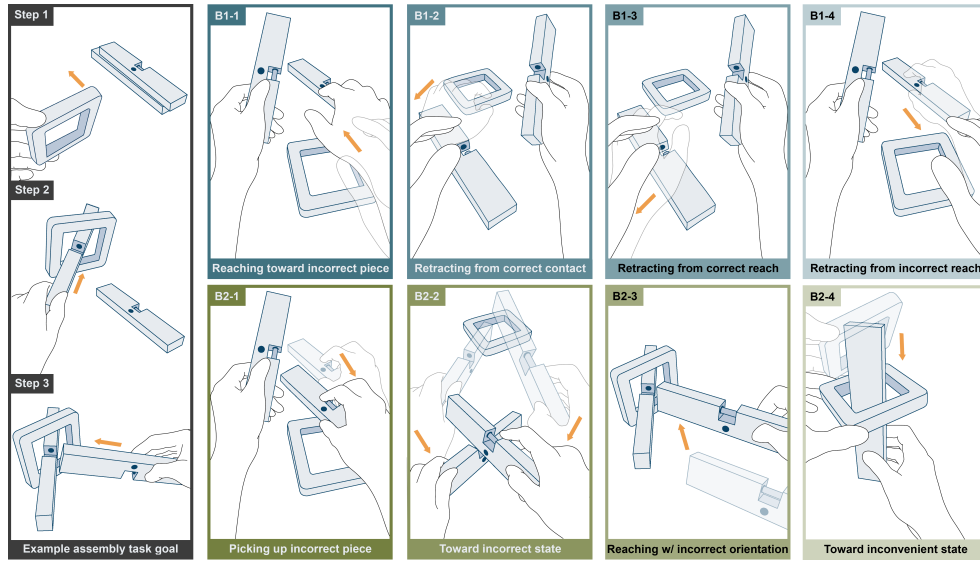


Figure 3.7: Goal-Dependent Cues

Figure 3.8: This structured breakdown visually and conceptually distinguishes between the intended task flow, general signals of uncertainty, and specific action-based cues that can inform timely instructional support.

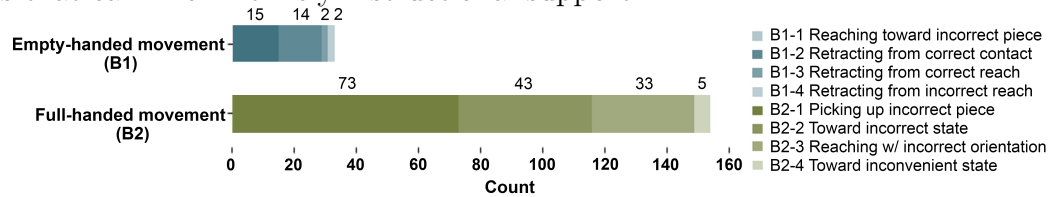


Figure 3.9: Goal-Dependent Cues Frequency

Figure 3.10: The frequency of instances Goal-dependent cues were observed during our coding of 500 video clips

Chapter 4

System Implementation

4.1 Method

Building on the insights from our preliminary exploration, we developed a movement-based system to detect user uncertainty and provide just-in-time instructions during manual tasks. Our method leverages implicit cues observed in user behavior, such as hand and head movements, combined with progress and effort metrics. This system aims to eliminate unnecessary interruptions while proactively delivering instructions at the ideal time—before the user makes an error or experiences undue hesitation.

4.1.1 Apparatus

We used OptiTrack motion tracking cameras to track the position and rotation of task components and users' hands. Reflective markers were attached to the hands and objects to capture precise movements in real time.

4.1.2 Heuristic Based System for Goal-Agnostic Cues

These movements serve as implicit cues of uncertainty, as identified in our preliminary analysis of 500 egocentric video clips. We implemented 2 of the most frequently

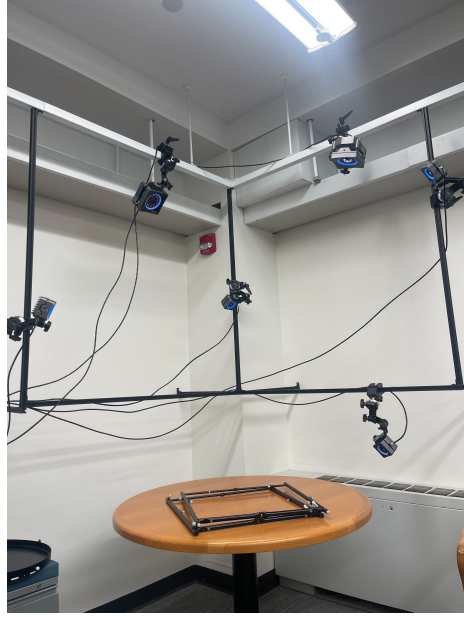


Figure 4.1: (a).OptiTrack Camera Setup

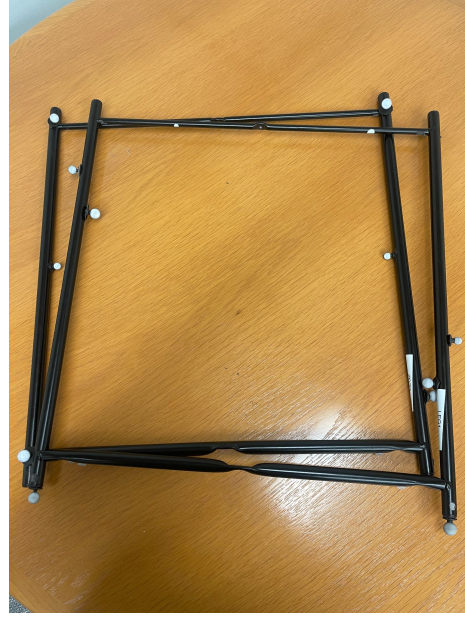


Figure 4.2: (b).object Tracking

Figure 4.3: (a) The OptiTrack camera setup used for tracking. (b) Example of object during a task.

observed cues:

- A1-1 Gaze Shift: Frequent angular movements are detected by looking at a rolling window of history of positional vectors and rotational quaternion of the head to detect if there are movements that happen without the accompanied movement of the hands or any of the pieces in the task.
- A2-1 and A2-2: Cyclical manipulations of components are detected through temporal pattern analysis within a rolling window of motion vectors. The system identifies uncertainty when objects return to previously held orientations (rotational oscillation) or positions (translational oscillation) multiple times within a short interval. When this cyclic pattern exceeds a threshold frequency (≥ 2 occurrences), the system interprets this as hesitation behavior indicative of decision uncertainty.

4.1.3 Progress and Effort Quantification for Goal-Dependent Cues

Our system quantifies *Progress* and *Effort* to evaluate task execution and determine when to provide instructions. The progress metric relies on the relative position and orientation of task components compared to their goal states, while the effort metric evaluates user motion efficiency.

Progress Calculation

Progress is measured as the alignment of task components relative to their fully assembled states. Specifically, for a pair of components i and j , progress is computed based on their position and orientation offsets using the following terms:

- $\mathbf{r}_{ij}$: The relative position vector between components i and j , calculated as $\mathbf{r}_{ij} = \mathbf{P}_j - \mathbf{P}_i$, where $\mathbf{P}_i$ and $\mathbf{P}_j$ are the absolute positions of components i and j .
- $\mathbf{r}_{ij}^{\text{goal}}$: The goal relative position vector representing the desired position of component j relative to i in the fully assembled state.
- $\mathbf{q}_{ij}$: The relative orientation quaternion between components i and j , calculated as $\mathbf{q}_{ij} = \mathbf{q}_j \otimes \mathbf{q}_i^{-1}$, where $\mathbf{q}_i$ and $\mathbf{q}_j$ are their absolute orientations.
- $\mathbf{q}_{ij}^{\text{goal}}$: The goal relative orientation quaternion representing the desired rotational relationship in the fully assembled state.
- ϵ_P : Position tolerance, below which position errors are considered negligible.
- ϵ_Q : Orientation tolerance, below which orientation errors are considered negligible.

Progress for a pair of components (i, j) is computed using the following formula:

$$\text{Progress} = 1 - \frac{1}{N} \sum_{(i,j) \in \text{Pairs}} \left(w_P \cdot \frac{\max\left(0, \|\mathbf{r}_{ij} - \mathbf{r}_{ij}^{\text{goal}}\| - \epsilon_P\right)}{\Delta P_{\max} - \epsilon_P} + w_Q \cdot \frac{\max\left(0, 1 - |\mathbf{q}_{ij} \cdot \mathbf{q}_{ij}^{\text{goal}}| - \epsilon_Q\right)}{1 - \epsilon_Q} \right), \quad (4.1)$$

where:

- w_P and w_Q are weights for the position and orientation errors, with $w_P + w_Q = 1$.
- $\Delta P_{\max}$: The maximum possible relative position error.
- N : The total number of component pairs, calculated as

$$N = \frac{N_{\text{pieces}}(N_{\text{pieces}} - 1)}{2},$$

where N_{pieces} is the total number of components.

Effort Calculation

Effort measures the user's physical actions during the task, focusing on the velocity of positional and rotational displacements of the hands. High effort with minimal progress indicates inefficiency, signaling the need for intervention. By monitoring these metrics in real-time, the system can identify when a user is struggling or uncertain, prompting timely instructions to guide them toward task completion.

Progress-Effort Patterns and Trigger Mechanisms

Our system dynamically evaluates task state transitions based on two metrics: **Progress** and **Effort**. These metrics, derived from position and orientation calculations, guide the system's intervention strategy. The trigger mechanism for each state is as follows:

1. **Ideal Progress (No Intervention Required):** When progress increases steadily with proportional effort (stable velocity in hand movements and stable rotations), the progress remains ideal. This corresponds to users effectively aligning task components and builders executing steps with confidence and precision.
2. **Fluctuating Progress (Hesitation Detected):** Progress oscillates between increases and decreases, often accompanied by moderate to high effort. This is identified by fluctuations in the position and orientation metrics:

$$\|\mathbf{r}_{ij} - \mathbf{r}_{ij}^{\text{goal}}\| \quad \text{and} \quad |\mathbf{q}_{ij} \cdot \mathbf{q}_{ij}^{\text{goal}}|.$$

This cue is further refined by our Heuristic-Based trigger of A2 cues. As fluctuations can be caused by more reasons other than hesitation or uncertainty. But specifically looking at those cues helps us identify a more reliable fluctuating progress.

3. **Neutral Progress (Stalled State):** Progress remains static, failing to improve over a time window, despite measurable effort. The system detects this state with our heuristic based A1 trigger. Not only the stalled state is important but also that the head is moving shows that effort is being exerted but no progress is being made.
4. **Negative Progress (Error State):** Progress decreases significantly, as indicated by increasing position error $\|\mathbf{r}_{ij} - \mathbf{r}_{ij}^{\text{goal}}\|$ or decreasing rotational alignment $|\mathbf{q}_{ij} \cdot \mathbf{q}_{ij}^{\text{goal}}|$. Simultaneously, effort remains high, reflecting misaligned components or incorrect task execution. The system intervenes to deliver corrective instructions, preventing compounding errors. This state aligns with study findings where builders deviated from the desired trajectory, requiring immediate

guidance.

By tying these states directly to progress-effort calculations, our system ensures that interventions are grounded in measurable task dynamics. These mechanisms address the implicit cues identified in our explorations, to enable timely and need-based instruction delivery.

Chapter 5

Application

Our approach can be generalized in a variety of tasks as long as a goal state is identified.

5.1 Other assembly tasks

- Legos: LEGO building presents is an obvious application domain for our instruction-timing approach. Like furniture assembly, it requires extensive bilateral coordination as users manipulate small pieces with both hands simultaneously. The spatial and temporal coordination demands create observable uncertainty markers and can be picked up as long as a goal state is known.
- Assembling a device: The challenge in such tasks can be framed the same way we have framed furniture assembly. Assembling a computer can be framed as a series of subtasks where one connects two subcomponents at a time. Progress and Effort can be detected

5.2 Navigation

The uncertainty detection principles we’ve developed can be extended to navigation tasks, where decision-making is similarly embodied in physical movements. Just as assembly tasks reveal ”the mind overflowing into the world” through physical actions, navigation involves continuous sensorimotor coupling with the environment.

Goal-independent navigation cues include head oscillations at decision points, slowing pace before intersections, or repeated rotational movements-behaviors analogous to the ”vicarious trial and error” patterns observed in our assembly studies. Goal-dependent cues manifest as trajectory deviations from optimal paths, hesitation when approaching incorrect routes, or reverting to previous positions after initial directional commitments.

These physical manifestations of uncertainty emerge in navigating physical environments (e.g., finding your way back out of a hospital exam room) making our detection framework further generalizable.

Chapter 6

Limitations and Future Work

While our system introduces a novel approach to providing just-in-time instructions, several technical and methodological limitations highlight areas for future refinement. On the technical side, the system relies on precise motion tracking, which can fail due to occlusions, calibration issues, or environmental constraints, leading to delays or inaccuracies in instruction delivery.

From a methodological perspective, our approach faces challenges in coming up with a strategy to adapt for the variability in user expertise. Novices may require more frequent interventions, while experienced users may find the system intrusive.

Finally, achieving proactive, human-like instruction timing remains a challenge, as the system largely reacts to measurable changes rather than inferring deeper intent such as user’s desire to explore. The level of proactivity desired is in fact still a big field to explore in various scenarios. People have varying levels of patience when it comes to a system that proactively intervenes. This variation is not only observed between people but also depending on the task or activity the system intervenes on [24, 23].

Bibliography

- [1] M. Agrawala, D. Phan, J. Heiser, J. Haymaker, J. Klingner, P. Hanrahan, and B. Tversky. Designing effective step-by-step assembly instructions. *ACM Trans. Graph.*, 22(3):828–837, July 2003.
- [2] F. Anderson, T. Grossman, J. Matejka, and G. Fitzmaurice. Youmove: enhancing movement training with an augmented reality mirror. In *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology*, UIST '13, page 311–320, New York, NY, USA, 2013. Association for Computing Machinery.
- [3] I. Arroyo, D. G. Cooper, W. Burleson, B. P. Woolf, K. Muldner, and R. Christopherson. Emotion sensors go to school. In *Proceedings of the 2009 Conference on Artificial Intelligence in Education: Building Learning Systems That Care: From Knowledge Representation to Affective Modelling*, page 17–24, NLD, 2009. IOS Press.
- [4] S. Castelo, J. Rulff, E. McGowan, B. Steers, G. Wu, S. Chen, I. Roman, R. Lopez, E. Brewer, C. Zhao, et al. Argus: Visualization of ai-assisted task guidance in ar. *IEEE Transactions on Visualization and Computer Graphics*, 2023.
- [5] A. K. Dey, G. D. Abowd, and D. S. and. A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications. *Human-Computer Interaction*, 16(2-4):97–166, 2001.

- [6] A. Dix. Human–computer interaction: A stable discipline, a nascent science, and the growth of the long tail. *Interacting with computers*, 22(1):13–27, 2010.
- [7] V. Enachescu, P. Schrater, S. Schaal, and V. Christopoulos. Action planning and control under uncertainty emerge through a desirability-driven competition between parallel encoding motor plans. *PLoS computational biology*, 17(10):e1009429, 2021.
- [8] G. Fischer. User modeling in human–computer interaction. *User modeling and user-adapted interaction*, 11:65–86, 2001.
- [9] G. Fischer. Context-aware systems: the ‘right’ information, at the ‘right’ time, in the ‘right’ place, in the ‘right’ way, to the ‘right’ person. In *Proceedings of the International Working Conference on Advanced Visual Interfaces, AVI ’12*, page 287–294, New York, NY, USA, 2012. Association for Computing Machinery.
- [10] M. Funk, T. Kosch, and A. Schmidt. Interactive worker assistance: comparing the effects of in-situ projection, head-mounted displays, tablet, and paper instructions. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp ’16*, page 934–939, New York, NY, USA, 2016. Association for Computing Machinery.
- [11] F. GANIER. Factors affecting the processing of procedural instructions: implications for document design. *IEEE Transactions on Professional Communication*, 47(1):15–26, 2004.
- [12] J. Gordon, A. Maselli, G. L. Lancia, T. Thiery, P. Cisek, and G. Pezzulo. The road towards understanding embodied decisions. *Neuroscience & Biobehavioral Reviews*, 131:722–736, 2021.

- [13] S. Henderson and S. Feiner. Exploring the benefits of augmented reality documentation for maintenance and repair. *IEEE Transactions on Visualization and Computer Graphics*, 17(10):1355–1368, 2011.
- [14] S. J. Henderson and S. K. Feiner. Augmented reality in the psychomotor phase of a procedural task. In *2011 10th IEEE International Symposium on Mixed and Augmented Reality*, pages 191–200, 2011.
- [15] G. Huang, X. Qian, T. Wang, F. Patel, M. Sreeram, Y. Cao, K. Ramani, and A. J. Quinn. Adaptutar: An adaptive tutoring system for machine tasks in augmented reality. In Y. Kitamura, A. Quigley, K. Isbister, T. Igarashi, P. Bjørn, and S. M. Drucker, editors, *CHI '21: CHI Conference on Human Factors in Computing Systems, Virtual Event / Yokohama, Japan, May 8-13, 2021*, pages 417:1–417:15, New York, NY, USA, 2021. ACM.
- [16] S. T. Iqbal, X. S. Zheng, and B. P. Bailey. Task-evoked pupillary response to mental workload in human-computer interaction. In *CHI '04 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '04, page 1477–1480, New York, NY, USA, 2004. Association for Computing Machinery.
- [17] X. Kang and Y.-L. Kuo. Incorporating task progress knowledge for subgoal generation in robotic manipulation through image edits, 2024.
- [18] M. Kraus, M. Schiller, G. Behnke, P. Bercher, M. Dorna, M. Dambier, B. Glimm, S. Biundo, and W. Minker. ”was that successful?” on integrating proactive meta-dialogue in a diy-assistant using multimodal cues. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, ICMI '20, page 585–594, New York, NY, USA, 2020. Association for Computing Machinery.
- [19] M. Krüger and J. Hermsdörfer. Target uncertainty during motor decision-making: The time course of movement variability reveals the effect of differ-

- ent sources of uncertainty on the control of reaching movements. *Frontiers in Psychology*, 10:41, 2019.
- [20] C. Li, G. Wu, G. Y.-Y. Chan, D. G. Turakhia, S. C. Quispe, D. Li, L. Welch, C. Silva, and J. Qian. Satori: Towards proactive ar assistant with belief-desire-intention user modeling, 2025.
- [21] Z. Liu, Z. Zhu, E. Jiang, F. Huang, A. M. Villanueva, X. Qian, T. Wang, and K. Ramani. Instrumentar: Auto-generation of augmented reality tutorials for operating digital instruments through recording embodied demonstration. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery.
- [22] Y. J. Ma, J. Hejna, A. Wahid, C. Fu, D. Shah, J. Liang, Z. Xu, S. Kirmani, P. Xu, D. Driess, T. Xiao, J. Tompson, O. Bastani, D. Jayaraman, W. Yu, T. Zhang, D. Sadigh, and F. Xia. Vision language models are in-context value learners, 2024.
- [23] C. Meurisch, C. A. Mihale-Wilson, A. Hawlitschek, F. Giger, F. Müller, O. Hinz, and M. Mühlhäuser. Exploring user expectations of proactive ai systems. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(4):1–22, 2020.
- [24] O. Miksik, I. Munasinghe, J. Asensio-Cubero, S. R. Bethi, S. Huang, S. Zylfo, X. Liu, T. Nica, A. Mitrocsak, S. Mezza, et al. Building proactive voice assistants: When and how (not) to interact. *arXiv preprint arXiv:2005.01322*, 2020.
- [25] J. B. Morrison and B. Tversky. The (in)effectiveness of animation in instruction. In *CHI '01 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '01, page 377–378, New York, NY, USA, 2001. Association for Computing Machinery.

- [26] A. A. Nargund, A. Aponte, A. Caetano, and M. Sra. Modband: Design of a modular headband for multimodal data collection and inference. In *Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23 Adjunct, New York, NY, USA, 2023. Association for Computing Machinery.
- [27] V. Pejovic and M. Musolesi. Interruptme: designing intelligent prompting mechanisms for pervasive applications. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '14, page 897–908, New York, NY, USA, 2014. Association for Computing Machinery.
- [28] A. D. Redish. Vicarious trial and error. *Nature Reviews Neuroscience*, 17(3):147–159, 2016.
- [29] B. R. Severitt, N. Castner, and S. Wahl. Bi-directional gaze-based communication: A review. *Multimodal Technologies and Interaction*, 8(12), 2024.
- [30] H. A. Simon and J. R. Hayes. Understanding complex task instructions. In D. Klahr, editor, *Cognition and Instruction*, pages 269–286. Lawrence Erlbaum Associates, 1976.
- [31] L. A. Suchman. *Plans and situated actions: The problem of human-machine communication*. Cambridge university press, 1987.
- [32] K. Szlichcinski. Telling people how things work. *Applied Ergonomics*, 10(1):2–8, 1979.
- [33] A. Truong, P. Chi, D. Salesin, I. Essa, and M. Agrawala. Automatic generation of two-level hierarchical tutorials from instructional makeup videos. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

- [34] B. Tversky. *Mind in Motion: How Action Shapes Thought*. Basic Books, 2019.
- [35] N. Van Berkel, M. B. Skov, and J. Kjeldskov. Human-ai interaction: intermittent, continuous, and proactive. *Interactions*, 28(6):67–71, 2021.
- [36] H. van der Meij. Goal-orientation, goal-setting, and goal-driven behavior in minimalist user instructions. *IEEE Transactions on Professional Communication*, 50(4):295–305, 2007.
- [37] X. Wang, T. Kwon, M. Rad, B. Pan, I. Chakraborty, S. Andrist, D. Bohus, A. Feniello, B. Tekin, F. V. Frujeri, N. Joshi, and M. Pollefeys. Holoassist: an egocentric human interaction dataset for interactive ai assistants in the real world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20270–20281, October 2023.
- [38] P. Wright. The instructions clearly state... can’t people read? *Applied Ergonomics*, 12(3):131–141, 1981.
- [39] J. J. Yang, L. Qiu, E. A. Corona-Moreno, L. Shi, H. Bui, M. S. Lam, and J. A. Landay. Amma: Adaptive multimodal assistants through automated state tracking and user model-directed guidance planning. In *2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pages 892–902, 2024.
- [40] B. D. Ziebart, A. L. Maas, J. A. Bagnell, and A. K. Dey. Human behavior modeling with maximum entropy inverse optimal control. In *AAAI spring symposium: human behavior modeling*, volume 92, 2009.
- [41] R. Zou, Y. Liu, Y. Li, G. Chu, J. Zhao, and H. Cai. A novel human intention prediction approach based on fuzzy rules through wearable sensing in human–robot handover. *Biomimetics*, 8(4):358, 2023.