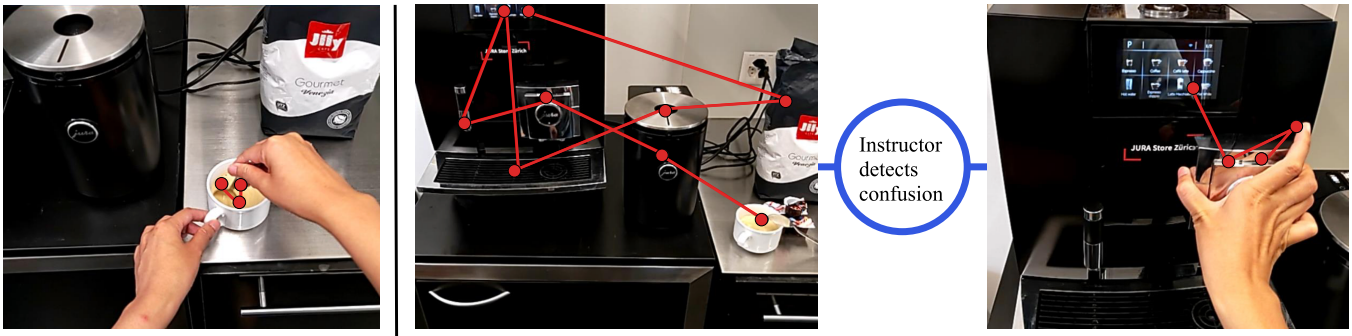
"Uh, *This* One?": Leveraging Behavioral Signals for Detecting Confusion during Physical Tasks

Maia Stiber
mstiber@jhu.edu
Johns Hopkins University
Baltimore, Maryland, USA

Dan Bohus
dbohus@microsoft.com
Microsoft Research
Redmond, Washington, USA

Sean Andrist
sandrist@microsoft.com
Microsoft Research
Redmond, Washington, USA



successful completion

Instructor: "Make a cup of coffee."

Instructor: "Great! Next, please empty the drip tray."

intervention for confusion mitigation

Instructor *follow-up*: "To remove the drip tray, take the thing where the coffee came out and put it a bit higher to slide it up."

Figure 1: First person view of someone performing a physical task (preparing coffee) according to an instructor's directions. Red dots and lines illustrate the performer's gaze trajectory. This work explores various modeling approaches that leverage behavioral signals from gaze, head, and hand movement for confusion detection in physical tasks.

Abstract

A longstanding goal in the AI and HCI research communities is building intelligent assistants to help people with physical tasks. To be effective in this, AI assistants must be aware of not only the physical environment, but also the human user and their cognitive states. In this paper, we specifically consider the detection of confusion, which we operationalize as the moments when a user is "stuck" and needs assistance. We explore how behavioral features such as gaze, head pose, and hand movements differ between periods of confusion vs no-confusion. We present various modeling approaches for detecting confusion that combine behavioral features, length of time, instructional text embeddings, and egocentric video. Although deep networks (*e.g.*, V-Jepa) trained on full video streams perform well in distinguishing confusion from non-confusion, simpler models leveraging lighter weight behavioral features exhibit similarly high performance, even when generalizing to unseen tasks.

ICMI '24, November 04–08, 2024, San Jose, Costa Rica

ACM ISBN 979-8-4007-0462-8/24/11
https://doi.org/10.1145/3678957.3685727

CCS Concepts

• **Human-centered computing** → **User models**.

Keywords

Confusion Detection; Human Behavioral Signals; Multimodal AI

ACM Reference Format:
Maia Stiber, Dan Bohus, and Sean Andrist. 2024. "Uh, *This* One?": Leveraging Behavioral Signals for Detecting Confusion during Physical Tasks. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '24), November 04–08, 2024, San Jose, Costa Rica.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3678957.3685727

1 Introduction

AI-based assistance systems for tasks in the physical world have the potential to augment human abilities by interpreting physical environments, interactive context, and user activities. However, the fluency and effectiveness of AI assistance relies on cooperation and coordination with human teammates [15, 51]. When humans collaborate on tasks, they naturally respond and adapt to each other's cognitive states—such as confusion, frustration, intention, and engagement—to provide assistance in ways that are timely, relevant, and helpful. For AI agent collaborators to fluidly provide assistance in physical tasks, they will similarly need to understand and react to users' cognitive states.

One effective way for an AI system to gain insight into user cognitive states is through human behavior modeling. Explicit behavioral

signals, such as speech, as well as implicit behavioral signals, such as gaze and facial expressions, can be used by AI systems to estimate the user's mental model of the task at hand, their intentions, and their cognitive states. Prior works in human-robot interaction have used behavioral signals such as facial expressions, gaze, and acoustic features to develop models for robot error detection and for recognizing user intentions [3, 23, 50]. In human-computer interaction more generally, behavioral signals such as EEG, heart rate, eye gaze patterns, and facial expressions have been used for assessing, *e.g.,* trust [1], cognitive load [45], and frustration [17].

In this paper, we focus on modeling implicit behavioral signals for detecting user confusion during AI-assisted procedural tasks in the physical world. Different tasks require different behavioral patterns for task completion [48] and systems need to be able to proactively detect confusion across tasks to be useful. Unmitigated user confusion can lead to frustration towards the task [2, 24] and disengagement [14], so systems should be aware and intervene upon detection [46]. Prior works researching confusion detection modeled implicit behavioral signals such as gaze-based features, facial expressions, speech, and screen-based features (*e.g.,* mouse behavior). These works were primarily situated in contexts involving screen-based tasks—such as learning [60] and dialogue tasks [30]—or where the user is seated, such as driving [18]. To the best of our knowledge, there has been little work in the area of confusion detection in fully 3D physical interactions, nor on assessing generalizability across a variety of procedural tasks.

In this paper, we operationalize confusion—a complex and nuanced cognitive state—in the context of procedural tasks in the physical world. Recognizing that behaviorally observable confusion is a narrower phenomenon than the internal subjective experience of confusion, we operationalize confusion as a period of time when a user is visibly "stuck" and needs assistance to make task progress. Taking inspiration from human-human interaction, we leveraged HoloAssist [56], a large-scale dataset of human-human task assistance, to identify sequences of observable confusion in task execution. Figure 1 shows an example of a human instructor responding to a task performer's nonverbal display of confusion and providing follow-up information to mitigate. We trained various models for detecting confusion from behavioral features combined with other modalities such as instruction text embeddings and egocentric video. Our work makes the following contributions:

- We define *confusion* in the context of procedural tasks such that it can be operationalized by an AI-powered physical task assistant in the physical world.
- We train models from behavioral features for confusion detection during physical tasks and compare with other modalities like video and instruction text embeddings.
- We explore the benefits of a multimodal approach to detecting confusion in physical tasks.

2 Background

2.1 AI Assistants for Physical Tasks

AI assistants for physical tasks can provide both instructive guidance and automatic feedback to improve a human's task performance and mitigate task breakdowns. AI assistants have been shown to be effective in complex procedural tasks by lowering user mental demand, reducing error rates during task execution, and decreasing the time it takes to repair errors [8, 16, 19, 52]. However, there are mixed results regarding their impact on task completion time, with some showing increases [52] and others, decreases [8, 16]. To be effective, intelligent interactive systems must perceive a user's actions and the environment, rationalize the implications of actions, and establish and maintain grounded, productive interactions with the user [10].

A potentially effective way for providing assistance is through Augmented Reality (AR) devices. In addition to conventional instructions, these devices can also overlay visual cues on the physical world [11], leading to improved spatial awareness [57]. As such, they have found applications in various fields, including training for surgical tasks [7], providing guidance during physical exercises [53], helping cognitively impaired workers [16], and assisting repair and maintenance tasks in manufacturing, military, and space [8, 19].

AI assistants have potential to enhance task performance and maintain efficient human-assistant interaction by being *proactive* [21]. Unlike purely reactive systems, proactive systems not only respond to user actions but also anticipate upcoming events, thereby building trust [25]. Example use cases for proactive interventions include providing alerts when a user makes (or ideally when the user is about to make) a mistake in an assembly task [13], and offering guidance on how to navigate and formulate queries for how-to videos related to physical tasks [31]. Creating proactive systems requires modeling whether and when an intervention is necessary, and what the intervention style should be [38]. These aspects are crucial as different individuals prefer varying types and amounts of proactive assistance [31, 33], and an excessively proactive approach may not be ideal [42]. Moreover, implementing these techniques requires perception and detection of explicit and implicit signals, so systems must be cautious to avoid excessive false positives, as these could degrade the interaction [12, 34].

2.2 Modeling Cognitive States from Behavior

Effective human-computer interactions necessitate a shared understanding between the user and the system [32]. The system needs to be aware of human capabilities, and must understand, predict, and adapt to their cognitive states [20, 41]. User cognitive states (*e.g.,* frustration, confusion, intention, and engagement) can significantly impact interaction dynamics and fluency [4]. Therefore, generating an accurate model of a user's cognitive state and intentions is necessary for fluent and effective interactions [15, 44, 51].

People produce explicit and implicit behavior signals not only in social interactions, but also when engaged in non-social, physical tasks [35, 50]. These signals can serve as observable reflections of users' intentions and internal states [9, 15, 51]. Estimates of cognitive states must capture both the inherent variability of human responses and underlying consistent behavioral patterns across tasks and humans. Behavioral signals have been used to detect errors in physical tasks (*i.e.,* hand and gaze movements [59]), user uncertainty (*i.e.,* facial expressions, gaze, head movement, and speech [47]), and user intention in human-robot interaction (*i.e.,* facial expressions, gaze, and verbalizations [3, 23]). Additionally, various behavioral signals like EEG, heart rate, and eye gaze have proven useful in gauging user trust in AI suggestions [1], while

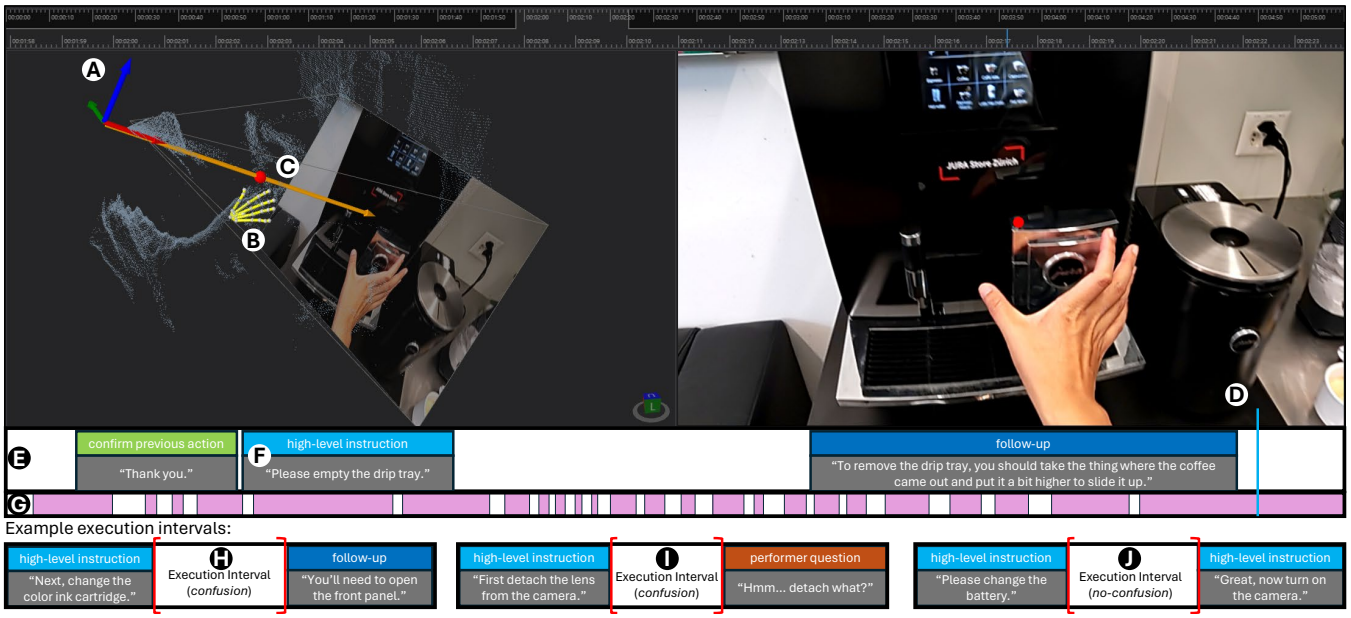


Figure 2: Data modalities from the HoloAssist dataset, and example execution intervals computed from utterance annotations. (A) Performer head pose. (B) Hand tracking. (C) 3D gaze tracking. (D) Performer's egocentric RGB view at every frame. (E) Instructor and performer utterance annotations from HoloAssist. (F) Text transcripts for instructions. (G) We compute temporal gaze fixations using the I-VT algorithm [39]. (H) Example of a *confusion* execution interval between the end of a high-level instruction and the start of an instructor follow-up. (I) Example of a *confusion* execution interval between the end of a high-level instruction and the start of a performer question. (J) Example of a *no-confusion* execution interval between the end of a high-level instruction and the start of the next high-level instruction.

other metrics—such as gaze patterns and facial expressions—have been used to quantify cognitive load and frustration [17, 45].

The cognitive state of *confusion* is important for intelligent systems to adapt and respond to. Unmitigated confusion not only prevents effective interaction, but is also a precursor to user frustration and disengagement [2, 14, 24]. The implications of not addressing confusion promptly and adequately can derail the user experience and deteriorate the effectiveness of systems, especially in important areas like surgery [22], learning [2], and driving [18].

Therefore, systems should be confusion-aware and intervene upon detection [46]. Gaze is a common indicator and was found to be statistically significantly different between states of confusion and confidence [18, 30, 55, 58]. Additional studies have integrated other behavioral, verbal, and physiological signals—such as emotions [30], head pose [22], verbal interactions [46], facial expressions [27], mouse movements [26], and EEG [36, 60]—demonstrating their relevance in detecting confusion across verbal interactions and 2D screen-based interactions. However, there appears to have been little work to understand and detect confusion in procedural tasks conducted in the 3D physical world.

3 Confusion during Physical Tasks

When developing an AI assistant for procedural tasks that is able to address when a user is confused and provide proactive help, the assistant needs a model to infer the user's internal cognitive states. One route is to learn from human-human interactions, as humans are adept at detecting when others are confused based on explicit or implicit signals. Examples of such signals are: deviation in task progress with respect to the observer's mental model of what should have happened, prior instruction complexity, and human behavioral signals. We propose the use of these explicit and implicit signals for detecting overt confusion. Our initial modeling approach simplifies the problem space into binary classification—intervals of time are labeled to correspond either to *confusion* or *no-confusion*.

3.1 Defining Confusion

In psychology, *confusion* is commonly defined as an epistemic emotion or affective state where there is a misalignment between a person's existing knowledge and new information [14, 54]. Starting from that definition and adapting it to the context of procedural tasks in the physical world (where only task completion, not learning, is the goal), we define confusion as an *internal cognitive state where there is a misalignment between a person's mental model of the task or world state and the actual state of the task or world.* In proactive AI assistant systems, taking action based on inferred confusion needs to be done cautiously, in order to reduce false positives and so that the assistant is not perceived as annoying [12, 34]. Therefore, we operationalize user "confusion" as a perceptible moment in time where assistance would be both relevant and helpful. In addition, from a practical standpoint, this form of confusion can be labelled from a third-party perspective, without having the participants self report when they *felt* confused. However, it is important to

emphasize that our operationalized definition of confusion for this work does not capture internal mental processes. Therefore, in the remaining parts of the paper, we move forward with this narrower operationalized definition of *behaviorally observable* confusion, acknowledging the broader and more subtle notion of confusion as a hidden *affective state* that it is related to.

3.2 Curating a Confusion Interval Dataset

Our overall goal is to uncover when and how confusion is exhibited and responded to during physical procedural tasks, and to train machine learning models to detect those moments of confusion. To this end, we curated a targeted dataset of "confusion intervals" from the publicly available HoloAssist dataset [56]. HoloAssist was not specifically collected to study confusion, but it does contain naturalistic data of human *performers* executing 22 real-world procedural tasks under the guidance of human *instructors* (222 total participants). Several modalities are captured, including the performer's egocentric video, audio, depth, and behavioral signal data (*i.e.,* gaze, hands, and head) from a HoloLens 2 device (Figure 2). Annotations with timestamped transcripts of all instructor and performer utterances are also provided, each labeled with a "conversation purpose." For this work, we focus on utterances with the following labels:

- *High-level instruction*: a task step provided by the instructor. Without these instructions, the performer would not know what to do. For example, "Next, insert the micro SD card."
- *Follow-up*: an instructor utterance that is intended to clarify, elaborate, or provide additional information with respect to the current high-level instruction. For example, "Flip it around, it's the slot on the bottom."
- *Question*: a performer utterance that is asking a question about the current task step. For example, "Uh, *this* one?"

We use 14 of the 22 HoloAssist tasks in our curated dataset, excluding all furniture assembly and disassembly tasks, as these typically involved only a single *high-level instruction* at the start (*e.g.,* "Now disassemble the table"), and then no further instructions or follow-ups were needed. The remaining 14 tasks exhibit many more step-by-step high-level instructions, follow-ups, and performer questions throughout the interactions, and involve objects of varying size and complexity, such as printers, coffee machines, cameras, and circuit breakers (full list in Figure 3). HoloAssist contains 1114 total sessions involving these tasks. The *high-level instructions* in our dataset subset contained 8.69 words on average.

We extracted *execution intervals* from the HoloAssist data, and assigned ground truth labels of *confusion* and *no-confusion*. We define an *execution interval* as the period of time between the end of one *high-level instruction* and the start of the first following *intervention*. An intervention can either be the next high-level instruction, an instructor follow-up, or a performer question (see above). We do not consider intervals in which the performer makes a mistake and the instructor corrects their task error. Otherwise, if the intervention was a proactive follow-up from the instructor or a question asked by the performer, we take it to mean that the performer needed help in order to proceed, and therefore assign to that interval a label of *confusion* (according to our operationalized definition above). If on the other hand the next intervention is a new high-level instruction, we infer that the performer successfully completed the current step without any need for help, and assign a label of *no-confusion*. Concrete examples are visualized in Figure 2.

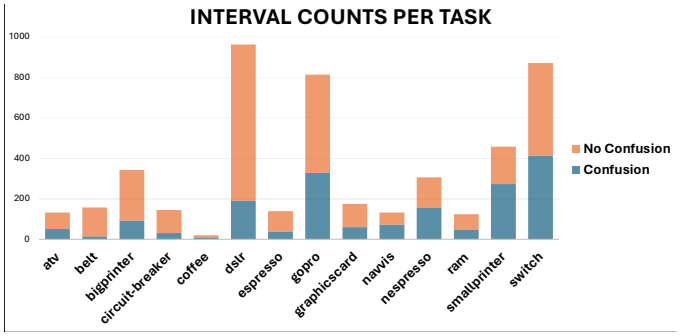


Figure 3: Graph illustrating the count of *confusion* and *no-confusion* execution intervals computed from the HoloAssist dataset, broken down by task type.

When extracting and labeling execution intervals from the dataset, we exclude intervals that are extremely short or long (less than two seconds and greater than 60 seconds), or if there are any other utterances occurring within the last three seconds of the interval before the start of the intervention (*e.g.,* performer's self-talk or the instructor asking the performer to adjust their HoloLens camera view). See Figure 3 for a breakdown of intervals across task.

3.2.1 Train and Test Sets. We created a task-stratified train-test split containing 1,442 *confusion* intervals (average length = 9.40s) and 2,325 *no-confusion* intervals (average length = 14.12s) for a total of 3,767 intervals. We also started with 1,044 intervals in the test set. Although the above heuristics work reasonably well for curating a set of *confusion* and *no-confusion* execution intervals, the existing HoloAssist transcriptions and intervention annotations are somewhat noisy, which can lead to incorrectly assigned ground truth on some intervals. For example, if an intervention was incorrectly labeled as a *follow-up* when it was actually a new *high-level instruction*, then the corresponding interval label should change from *confusion* to *no-confusion*, and vice versa. If the intervention was actually the instructor correcting a mistake in the performer's action, then the interval should not have been included in our dataset at all. In order to ensure a clean test set for evaluation, the authors manually inspected all 1,044 automatically extracted intervals (this time-intensive process was not feasible for the train set) and determined that 940 were labeled correctly (90.0%), 31 needed to be removed (3.0%), and 73 needed to switch labels (7.0%), resulting in 1,013 corrected intervals in the test set (333 *confusion*, 680 *no-confusion*).

4 Exploring Behavioral Features

Implicit behavioral signals are a powerful modality for modeling user internal cognitive states, effective in both capturing the variability of people's behaviors and responses, and having the potential to generalize across environments, embodiments, and tasks [29, 50]. Prior work has shown that gaze is a strong predictor for confusion in 2D interaction (*e.g.,* [40]). The HoloAssist dataset provides not

only gaze signals, but also video, depth, audio, head, and hand signals during procedural task execution, allowing us to explore other behavior signal modalities. A goal of this work is to determine how these different implicit behavioral signals vary based on confusion, and if they are good input features for confusion detection.

Anecdotally, we observed from the data that when performers were confused, they often exhibited a few different categories of behaviors: fast eye movement in conjunction with lack of hand movement, lack of eye movement in conjunction with fast hand movement, getting closer to objects (either through bringing the object closer to the head or moving the head closer to the workspace), and taking a step back from the workspace.

Based on prior work and our observations, we chose to focus on gaze, hand, and head features. We handcrafted several low-level features pertaining to each of these modalities to filter out some of the noise that often comes with behavior signals and distill them into a smaller number of features. This approach was designed to prevent overfitting to the dataset, allowing for potential generalizability across different tasks. The features were chosen so they would not explicitly represent task-specific and environment-specific information, such as absolute positions. Additionally, since behavioral signals might fluctuate over the course of an entire execution interval, we sampled each low-level feature over different temporal segments in the interval to capture changes over time.

We defined 48 behavioral features across gaze (7 features), hands (36 features), and head (5 features) modalities, and computed all features for the full interval, first two seconds of the interval, and last two seconds of the interval, resulting in 144 total features. For gaze features within the interval, we considered angular eye speed (average and standard deviation), number of gaze fixations per second, fixation duration (average, standard deviation, and total percentage of the interval), and percentage of valid gaze tracking values. For hand features, we considered linear joint speeds (average, standard deviation, and percentage of valid values) and distance from the head (average, standard deviation, and percentage of valid values) for the wrist, index tip, and thumb tip joints on the left and right hands. Finally, for the head, we compute features for linear and angular head speed (average and standard deviation) as well as percentage of valid head tracking values in the interval.

4.1 Implicit Behavioral Signals Dataset Analysis

We first analyzed our curated dataset and computed features to understand the differences in behavioral signals exhibited by performers during intervals of *confusion* vs *no-confusion*. The following analyses were computed using Mann-Whitney U Tests as the data was not normally distributed. Additionally, since our analysis includes multiple t-tests, we used the Benjamini-Hochberg correction to adjust our p-values and control for Type I errors.

All gaze features exhibited statistically significant differences between *confusion* and *no-confusion* intervals—with confusion intervals exhibiting faster and more variable gaze with shorter and more frequent fixations. This result matches our anecdotal observations that periods of confusion seem to often be marked by fast scanning behaviors as the performer attempts to find an object or part that is relevant to completing the current task step. The head features of average linear and angular speed were also significantly different across interval types, with faster and more variable speeds during *confusion*. For the hands, left-hand joints (wrist, index tip, and thumb tip) were statistically significant indicators, with lower variance in joint speeds and larger average joint speeds in *confusion* intervals than *no-confusion* ones. The right-hand joints' average speeds were statistically significant with lower speed in *confusion* intervals than *no-confusion* ones. This result can be intuitively explained by the hand moving less when the performer is confused and not sure what to do next, but moving faster when the performer is confidently executing the physical task step. See Appendix A for detailed statistical test for each feature.

The statistically significant differences in these features across *confusion* vs *no-confusion* intervals suggests the potential power that these features might have in *detecting* user confusion.

5 Learning to Detect Confusion

Informed by prior work and the data analysis above, we observed that behavioral signals have the potential to be discriminitive, and could contain useful information for confusion modeling and detection. We explore whether behavioral signals can be used to detect confusion and whether these models can be augmented with other features without compromising generalizability. Our approach simplifies the problem space to a binary classification between *confusion* and *no-confusion*, taking into account the entire interval of time leading up to a potential intervention (*e.g.,* following up with a clarification or moving on to the next step). This approach in essence learns *how* to intervene based on a confusion classification, but determining *when* to intervene is left for future work.

5.1 Method

We first look to leverage lightweight classification models for future integration into interactive agent systems for real-time confusion detection. The inputs into these models are the behavioral features, described above, calculated across the three temporal segments (full interval, first two seconds, and last two seconds). Models are trained to predict *confusion* as the positive class.

In addition, to gain a better understanding of this problem space, we explored whether and how other sources of knowledge, such as the content of the instructions, video-based information, and temporal information can aid with our confusion detection task. We constructed models that leverage these knowledge sources and we also explored combining them with the behavioral features in a multimodal approach. Specifically, the instruction transcripts, video, and temporal information were modeled as follows:

- *Instructions*: To leverage the content of the instructions, we used the CLIP [43] embeddings of the text of the *high-level instructions* for each interval. The intuition behind including this information is that different instructions may have different prior likelihoods of causing confusion.
- *Video*: To leverage the ego-centric video data, we used V-Jepa [5], a pretrained state-of-the-art video encoding model that can be used in video prediction tasks. Following the evaluation protocol described in [5], we kept the V-Jepa pretrained backbone encoder frozen (used the ViT-L model) and trained an additional cross-attention module with a learnable

query token, followed by a linear classifier, to predict confusion. Our intention was to explore how a heavier-weight, state-of-the-art video model fares against and complements a simpler, behavioral feature-based approach.
- *Temporal*: Finally, we also explored the use of the length of each interval (in seconds) as a source of information.

As mentioned above, the instruction, video and temporal features were also combined with the behavioral signals. For the instruction and temporal features, we simply concatenated the CLIP embeddings or interval duration with other features as input for a confusion detection model. When using video data, we employed two methods of integration. In a first approach, dubbed *concatenate*, we simply concatenated the additional features (e.g., behavioral, instruction embedding, time) to the output vector from the cross-attention probe operating over the V-Jepa encoder, and fed the larger (concatenated) vector into a final linear layer. In the second approach, dubbed *late fusion*, we used the prediction probabilities generated by trained video model described earlier as a single additional feature, which was then combined with the other features for processing by another light-weight confusion detection model.

For non-video models, we explored the use of Random Forest, SVM, XGBoost, and Neural Nets, with Explainable Boosting Machines (EBM) [37] being our ultimate choice. We construct an EBM model with 100 bins and two max leaves.

We conducted two tests: (1) The **within-task test** sought to determine different features' abilities to detect confusion, irrespective of the physical task. The training dataset consisted of data from all of the tasks. (2) The **between-task test** aimed to identify the potential that different features have to generalize across different tasks. Training was conducted in a "leave-one-task-out" regime.

5.2 Performance Measures

We used the following metrics to assess model performance with various sets of features in detecting confusion for both the within-task and between-task tests.

- *Accuracy (ACC)*: The percentage of intervals classified correctly. 67.1% of the test set intervals are labeled *no-confusion*.
- *AUC*: The area under the Receiver Operating Characteristic (ROC) curve, representing the true positive rate against the false positive rate.
- *Precision (PR)*: Number of correctly identified confusion intervals (true positive) divided by the total number of intervals classified as confusion (true positive + false positive).
- *Recall Given 90% Precision (R-90PR)*: In the context of an AI assistant, achieving high precision is crucial because mistakenly intervening when a user did not actually need help can significantly disrupt and frustrate the user [12, 34]. Therefore, an understanding of expected recall when a system is tuned for high precision is informative.

5.3 Results

5.3.1 Within-Task Tests. To evaluate the performance of models leveraging various knowledge sources—behavioral, instructions, video, temporal—we employed five-fold cross-validation in our within-task tests. We created five different models, each trained on four of the folds. Then, we evaluated each of these models on the test set. We compared the results across inputs using a Mann-Whitney U Test, adjusted for multiple comparisons with the Benjamini-Hochberg correction. We specifically examined how low-dimensional inputs (time, text, behavior) compared to performance using only behavioral signals, to understand their contributions to confusion detection. This analysis also aimed to determine how contextualizing behavioral signals affects performance. Additionally,

Table 1: Performance of models combining features from various sources of information (within-task). Metrics are accuracy (ACC), area under the curve (AUC), precision (PR), and recall given 90% precision (R-90PR). Non-video models are EBMs. Video models use the V-Jepa architecture. Values represent an average across each training fold computed on the test set. Statistical significance (p-values adjusted using Benjamini-Hochberg correction) is reported where * is p = .05, ** p = .01, and * is p < .001. A human annotator baseline (on 100 test examples) is also reported.**

	Within-Task			
	ACC	AUC	PR	R-90PR
Human ($\kappa = 0.837, p < .001$)	0.91	N/A	0.925	N/A
Mann-Whitney U Test (Comparing to *Behavior*)				
Behavior	0.738	0.767	0.626	0.053
Time	0.671**	0.626**	0**	0*
Instruction Embedding	0.732	0.765	0.612	0.071
Time + Instruction Embedding	0.806**	0.838**	0.739**	0.391**
Behavior + Time	0.744	0.780	0.630	0.059
Behavior + Time + Instruction	0.811**	0.858**	0.737**	0.440**
Mann-Whitney U Test (Comparing to *Behavior + Time + Instruction*)				
Video (V-Jepa)	0.806	0.865	0.716	0.378
ALL (Late Fusion)	0.811	0.858	0.738	0.432
ALL (Concatenate)	0.811	0.858	0.715	0.282

Table 2: Performance of EBM models trained on various subsets of behavioral features and temporal segments (within-task). Each row indicates which features were included in the model. The four metrics are accuracy (ACC), area under the curve (AUC), precision (PR), and recall given 90% precision (R-90PR). In bold are the highest values for each metric.

		Within-Task			
		ACC	AUC	PR	R-90PR
	All Behavior	0.738	**0.767**	0.626	0.053
Behavior Used	Gaze Only	0.692	0.685	0.542	0.002
	Head Only	0.654	0.605	0.443	0.001
	Hands Only	0.728	0.731	**0.637**	0.040
	Gaze + Head	0.697	0.703	0.549	0.005
	Gaze + Hands	**0.739**	0.765	0.633	**0.070**
	Head + Hands	0.725	0.727	0.616	0.023
Temporal Used	Full Interval	0.696	0.692	0.559	0.280
	First Two Seconds	0.678	0.600	0.560	0.002
	Last Two Seconds	0.710	0.696	0.595	0.004
	Full + First	0.705	0.703	0.572	0.023
	Full + Last	0.723	0.757	0.594	0.026
	First + Last	0.719	0.732	0.601	0.011

we assessed heavier weight computationally intensive video-based models (V-Jepa), combining with other modalities, against the best performing set of non-video lightweight models. Finally, to establish an upper bound of performance for this problem space, we asked two naive coders to annotate a subset of 100 test intervals. They annotated each interval by watching the video (no audio), along with the text transcript of the corresponding high-level instruction.

Table 1 illustrates each model's average performance across training folds. Human annotators achieved 91% accuracy, with an intercoder reliability of Cohen's $\kappa = 0.84, p < .001$. All models performed better than the majority class baseline (67.1%), with the multimodal approach of *Behavior + Time + Instructions* performing the best, statistically significantly outperforming the *Behavior*-only model. The *Time + Instructions* model is not far behind. The *Time*-only model performed the worst across all metrics. When exploring the heavyweight video-based models, the late fusion of *Video + Behavior + Time + Instructions* performed best. However, when comparing it to that of the lighter-weight EBM model of *Behavior + Time + Instructions*, there was no statistical significance across all metrics. This lack of significant difference was consistent across all video input-related models when compared to *Behavior + Time + Instructions*. See Appendix B Table 2 for the p-values.

To understand how different behavioral features influence model performance, we conducted several ablation tests (Table 2). For the behavior ablation, rather than removing individual features, we systematically removed sources of behavior: gaze, head, and hands. During this ablation, the features still spanned across all temporal segments. The gaze and hands modalities appeared to perform best. The head modality by itself performed the worst. For the temporal ablations, we removed features corresponding to each temporal segment, but retained all behavioral features. The first two-second segment alone had the lowest performance. The final two-second segment showed the highest performance among the individual segments, which is intuitive because confusion behaviors are likely most pronounced immediately before the intervention. The most effective temporal inputs combined features from all segments.

5.3.2 Generalizability. We also evaluated how well the models might perform on unseen tasks. To do this, we conducted a leave-one-task-out validation, where we tested each model's ability to generalize across various tasks. Table 3 shows the model's performance for different features (specifically the features that had the best accuracy and precision in the within-task test) across tasks. We found that the best models (lightweight and heavyweight) generally had similar across-task performance. However, the performances of each feature with respect to the tasks was highly variable. It is important to note that the amount of intervals per task vary greatly, so the metrics reported for the tasks with few intervals—such as coffee or RAM—might be less reliable and more susceptible to noise. The *Behavior + Time + Instructions* model appeared to perform the best on the "big printer", "espresso", "nespresso", and "Nintendo Switch" tasks. The *Video* model performs the best on "DSLR", "graphics card", "Navvis", and "RAM" tasks. The *Video + Behavior + Time + Instructions* model performs the best on "belt", "circuit breaker", and "GoPro" tasks. See Appendix C Table 3 for additional results for the *Behavior* and *Time + Instructions* inputs. In summary, *Behavior* appears to generalize better to new tasks than *Time + Instructions*.

6 Discussion

In this paper, we explored how intervals of confusion might be detected in physical tasks, with the ultimate aim of modeling this cognitive state for potential real-time detection during task execution. Detecting and modeling confusion is a difficult problem with many factors, such as confusion type, task context, and how a person implicitly expresses confusion. Our goal is to create a model that captures these subtle signs and can be used across different tasks. We highlight the important role that observing human behavior plays in understanding confusion and discuss the advantages of using a multimodal approach that considers both behavioral signals and other contextual features.

6.1 Confusion is Complex

Understanding and modeling confusion is a difficult problem. Results from the within-task test (training and testing on all tasks together) showed that the best performing model achieved about 81.1% accuracy (AUC = 0.858, precision = 0.737). Even with a state-of-the-art video-based model (V-Jepa) over high-dimensional features (egocentric video clips), we achieved an accuracy of only 80.6% (AUC = 0.865, precision = 0.716). The performance of both the lighter-weight models with lower dimensional features and the heavier-weight model with higher dimensional features for the simplified problem space underpins the intrinsic complexity of modeling confusion during physical tasks.

This complexity might partially stem from the fact that a human can experience different types of confusion for the same task and instruction, where behavioral patterns can be completely different. In addition, the complexity could also stem from the variability in how confusion can manifest across disparate tasks. As evidenced by our between-task tests, there is considerable performance variability across tasks irrespective of the feature type.

6.2 Using Behavioral Features

Regardless of task, humans seem to consistently signal their confusion through their behavior. Gaze, hand, and head features were all significantly different between *confusion* and *no-confusion* intervals. For example, we observed trends of increased gaze and left-hand activity, with faster angular head movements in confusion intervals. These observations were also reflected when it came to training confusion classification models from these signals (see Table 1). Within-task tests revealed that we are able to detect confusion using behavioral signals. In fact, when contextualized (*e.g., Behavior + Time + Instructions*), they deliver performance comparable to that of video-based, computationally intensive methods across all metrics.

Generalizability: Evaluating these models across task is also important when considering that AI assistants will need to be able to generalize their abilities to tasks unseen during training time. The results from the between-task test show that the best performing models, both lightweight behavior-based and heavyweight video-based, have the potential to generalize to unseen physical tasks (see Appendix C and Table 3).

6.3 Benefits of Multimodal Approach

A multimodal approach, combining behavioral signals with other contextual features, has the potential to improve the estimation of

Table 3: Performance of the three best features from the within-task test in generalizing across task types. Metrics are accuracy (ACC), area under the curve (AUC), precision (PR), and recall given 90% precision (R-90PR). Results were computed using leave-one-task-out. The bold numbers were the highest numbers for each metric and task.

	Lightweight				*Heavyweight*				*Combined*			
	Behavior + Time + Instructions				**Video**				**Video + Behavior + Time + Instructions**			
	ACC	AUC	PR	R-90PR	ACC	AUC	PR	R-90PR	ACC	AUC	PR	R-90PR
ATV	0.400	0.364	0.188	0	**0.633**	0.631	0.364	0	0.600	**0.670**	**0.389**	0
Belt	**0.727**	0.713	0.231	**0.200**	0.568	0.651	0.182	0	0.705	**0.769**	**0.250**	**0.200**
Big Printer	**0.613**	**0.899**	**0.426**	**0.565**	0.525	0.661	0.358	0.130	0.538	0.712	0.370	0.391
Circuit Breaker	0.675	0.500	0.182	0	0.675	0.598	0.182	0.167	**0.800**	**0.721**	**0.500**	**0.500**
Coffee	0.778	**0.778**	1	0.333	0.778	0.667	1	0.333	0.778	0.556	1	0.333
DSLR	**0.621**	0.644	0.237	0.094	0.526	**0.727**	**0.241**	0.031	0.389	0.709	0.200	**0.188**
Espresso	**0.889**	**0.986**	0.750	**0.857**	0.963	0.907	**1**	**0.857**	0.852	0.914	0.667	0.286
Gopro	0.747	0.795	0.722	0.232	0.785	0.862	0.775	0.089	**0.823**	**0.900**	**0.868**	**0.392**
Graphics Card	0.595	0.650	0.278	**0.111**	**0.738**	**0.731**	**0.429**	0	0.690	0.721	0.357	0
Navvis	0.680	0.507	**1**	0.111	**0.760**	**0.833**	**1**	**0.556**	0.640	0.802	0.500	0.111
Nespresso	**0.686**	0.679	**0.640**	0	0.549	0.583	0.500	0.087	0.627	**0.691**	0.600	**0.130**
RAM	0.621	0.461	0.250	0	**0.793**	**0.877**	**0.556**	**0.571**	0.655	0.643	0.385	0
Small Printer	0.636	0.775	0.865	0.318	0.607	0.781	**0.962**	**0.606**	**0.664**	**0.815**	0.941	0.591
Switch	0.746	**0.840**	**0.739**	0.468	0.713	0.791	0.652	0.278	**0.773**	0.837	0.711	0.405

human cognitive states. Our results demonstrate that combining behavioral signals with time and instruction embeddings yields the best confusion detection performance, irrespective of the task.

Part of the reason why adopting a multimodal approach is important for detecting confusion in physical tasks is that confusion seems to have a strong prior. The *Time + Instructions* model was successful for confusion detection—better than behavioral signals alone. This performance shows us that some instructions are, by default, inherently more prone to confusion than others. While behavioral signals vary based on the person, task, and human cognitive state, the text feature captures a consistent underlying tendency towards confusion. Therefore, combining *Instructions* and behavioral signals facilitates improved confusion detection, as the instruction embeddings provide a stable prior, while behavioral signals introduce variability as well as enabling streaming. It is important to note that *Time + Instructions* alone, while having good confusion detection performance, would be insufficient for deployment because that input can only be used to detect the "potential" for an instruction to be confusing. This alone is not useful for just-in-time intervention. However, behavioral signals can help models detect the onset of confusion in its early stages, enabling preemptive interventions and support before the user experience is significantly affected [49].

6.4 Future Work and Limitations

As our results have shown, detecting and modeling confusion is a complex problem, which highlights the fact that several different factors need to be considered when modeling human cognitive states during task execution. Future work should explore why certain tasks were not successfully generalized to, such as analyzing if the complexity, physicality, or cognitive demand of the task correlates with generalization capability.

Future work should also address transitioning our approach to a real-time streaming paradigm, allowing for just-in-time feedback during physical task execution. One potential strategy using our model in real-time could involve classifying segments up to the current frame to determine if the participant is confused at that point and developing policies to decide when and how the AI should intervene. Then the streaming model could be integrated into an AI assistant displayed via augmented reality (*e.g.,* [6]), which has been shown to be a promising technology for interactively delivering instructions for complex physical tasks, by lowering user mental demand and reducing error rate during task execution [8, 52].

One limitation for this work is that all of these tasks are dependent on instructions provided by an instructor. However, not all physical tasks are structured or dependent on instructions in that way. Confusion detection should be explored outside of highly structured tasks and consider the performance of different features.

Another limitation arises from how the confusion ground truth labels for the HoloAssist dataset were derived: they were indirectly inferred based on the overt behaviors of the participants. Future investigations might explore other approaches for accurately labeling confusion in datasets, ensuring robustness and reliability.

Work in the area of confusion and physical tasks should expand beyond just detection to other areas of confusion management: classification, severity assessment, and mitigation. Future work should classify confusion *types* to select appropriate mitigation strategies. For example, if an AI assistant system interacts with a user through an augmented reality headset and the user is confused about an object's location, the system could generate holograms to point to different object locations. Additionally, we should consider how to implement escalating mitigation interventions if the confusion persists [28]. This adaptive strategy would enhance an AI assistance's ability to navigate the nuances of confusion during task execution and enhance user experience and performance in real-time.

Acknowledgments

We would like to thank Rich Caruana, Neel Joshi, Vibhav Vineet, and Xin Wang for their help and insights during this project.

References

[1] Kumar Akash, Wan-Lin Hu, Neera Jain, and Tahira Reid. 2018. A classification model for sensing human trust in machines using EEG and GSR. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 8, 4 (2018), 1–20.

[2] Amaël Arguel, Lori Lockyer, Ottmar V Lipp, Jason M Lodge, and Gregor Kennedy. 2017. Inside out: detecting learners' confusion to improve interactive digital learning environments. *Journal of Educational Computing Research* 55, 4 (2017), 526–551.

[3] Reuben M. Aronson and Henny Admoni. 2018. Gaze for error detection during human-robot shared manipulation.

[4] Brian P Bailey and Joseph A Konstan. 2006. On the need for attention-aware systems: Measuring effects of interruption on task performance, error rate, and affective state. *Computers in human behavior* 22, 4 (2006), 685–708.

[5] Adrien Bardes, Quentin Garrido, Jean Ponce, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. 2024. Revisiting Feature Prediction for Learning Visual Representations from Video. *arXiv:2404.08471* (2024).

[6] Dan Bohus, Sean Andrist, Nick Saw, Ann Paradiso, Ishani Chakraborty, and Mahdi Rad. 2024. SIGMA: An Open-Source Interactive System for Mixed-Reality Task Assistance Research - Extended Abstract. In *2024 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE.

[7] Sanne MBI Botden, Sonja N Buzink, Marlies P Schijven, and Jack J Jakimowicz. 2008. ProMIS augmented reality training of laparoscopic procedures face validity. *Simulation in healthcare* 3, 2 (2008), 97–102.

[8] Adam M Braly, Benjamin Nuernberger, and So Young Kim. 2019. Augmented reality improves procedural work on an international space station science instrument. *Human factors* 61, 6 (2019), 866–878.

[9] Alexandra Bremers, Alexandria Pabst, Maria Teresa Parreira, and Wendy Ju. 2023. Using Social Cues to Recognize Task Failures for HRI: A Review of Current Research and Future Directions. *arXiv preprint arXiv:2301.11972* (2023).

[10] Sonia Castelo, Joao Rulff, Erin McGowan, Bea Steers, Guande Wu, Shaoyu Chen, Iran Roman, Roque Lopez, Ethan Brewer, Chen Zhao, et al. 2023. ARGUS: Visualization of AI-Assisted Task Guidance in AR. *arXiv preprint arXiv:2308.06246* (2023).

[11] Subramanian Chidambaram, Hank Huang, Fengming He, Xun Qian, Ana M Villanueva, Thomas S Redick, Wolfgang Stuerzlinger, and Karthik Ramani. 2021. Processar: An augmented reality-based tool to create in-situ procedural 2d/3d ar instructions. In *Designing Interactive Systems Conference 2021*. 234–249.

[12] Andrea Cuadra, Shuran Li, Hansol Lee, Jason Cho, and Wendy Ju. 2021. My bad! repairing intelligent voice assistant errors improves interaction. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–24.

[13] Guodong Ding, Fadime Sener, Shugao Ma, and Angela Yao. 2023. Every Mistake Counts in Assembly. *arXiv preprint arXiv:2307.16453* (2023).

[14] Sidney D'Mello and Art Graesser. 2014. Confusion and its dynamics during device comprehension with breakdown scenarios. *Acta psychologica* 151 (2014), 106–116.

[15] Zoran Duric, Wayne D Gray, Ric Heishman, Fayin Li, Azriel Rosenfeld, Michael J Schoelles, Christian Schunn, and Harry Wechsler. 2002. Integrating perceptual and cognitive modeling for adaptive and intelligent human-computer interaction. *Proc. IEEE* 90, 7 (2002), 1272–1289.

[16] Markus Funk, Sven Mayer, and Albrecht Schmidt. 2015. Using in-situ projection to support cognitively impaired workers at the workplace. In *Proceedings of the 17th international ACM SIGACCESS conference on Computers & accessibility*. 185–192.

[17] Joseph Grafsgaard, Joseph B Wiggins, Kristy Elizabeth Boyer, Eric N Wiebe, and James Lester. 2013. Automatically recognizing facial expression: Predicting engagement and frustration. In *Educational data mining 2013*.

[18] Shabnam Haghzare, Jennifer L Campos, and Alex Mihailidis. 2022. Classifying older drivers' gaze behaviour during automated versus non-automated driving: A preliminary step towards detecting mode confusion. *International Journal of Human–Computer Interaction* (2022), 1–14.

[19] Steven Henderson and Steven Feiner. 2010. Exploring the benefits of augmented reality documentation for maintenance and repair. *IEEE transactions on visualization and computer graphics* 17, 10 (2010), 1355–1368.

[20] Guy Hoffman and Cynthia Breazeal. 2007. Cost-based anticipatory action selection for human–robot fluency. *IEEE transactions on robotics* 23, 5 (2007), 952–961.

[21] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 159–166.

[22] Benedikt Hosp, Myat Su Yin, Peter Haddawy, Ratthaphum Watcharopas, Paphon Sa-Ngasoongsong, and Enkelejda Kasneci. 2021. States of Confusion: Eye and Head Tracking Reveal Surgeons' Confusion during Arthroscopic Surgery. In *Proceedings of the 2021 International Conference on Multimodal Interaction*. 753–757.

[23] Dimosthenis Kontogiorgos, Minh Tran, Joakim Gustafson, and Mohammad Soleymani. 2021. A systematic cross-corpus analysis of human reactions to robot conversational failures. In *Proceedings of the 2021 International Conference on Multimodal Interaction*. 112–120.

[24] Barry Kort, Rob Reilly, and Rosalind W Picard. 2001. An affective model of interplay between emotions and learning: Reengineering educational pedagogy-building a learning companion. In *Proceedings IEEE international conference on advanced learning technologies*. IEEE, 43–46.

[25] Matthias Kraus, Marvin Schiller, Gregor Behnke, Pascal Bercher, Michael Dorna, Michael Dambier, Birte Glimm, Susanne Biundo, and Wolfgang Minker. 2020. "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*. 585–594.

[26] Sébastien Lallé, Cristina Conati, and Giuseppe Carenini. 2016. Predicting Confusion in Information Visualization from Eye Tracking and Interaction Data.. In *IJCAI*. 2529–2535.

[27] Na Li, John D Kelleher, and Robert Ross. 2022. Detecting Interlocutor Confusion in Situated Human-Avatar Dialogue: A Pilot Study. *arXiv preprint arXiv:2206.02436* (2022).

[28] Na Li and Robert Ross. 2022. Dialogue Policies for Confusion Mitigation in Situated HRI. *arXiv preprint arXiv:2208.09367* (2022).

[29] Na Li and Robert Ross. 2022. Transferring Studies Across Embodiments: A Case Study in Confusion Detection. *arXiv preprint arXiv:2206.01493* (2022).

[30] Na Li and Robert Ross. 2023. Hmm, You Seem Confused! Tracking Interlocutor Confusion for Situated Task-Oriented HRI. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. 142–151.

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

[32] Joseph B Lyons and Paul R Havig. 2014. Transparency in a human-machine context: Approaches for fostering shared awareness/intent. In *Virtual, Augmented and Mixed Reality. Designing and Developing Virtual and Augmented Environments: 6th International Conference, VAMR 2014, Held as Part of HCI International 2014, Heraklion, Crete, Greece, June 22-27, 2014, Proceedings, Part I 6*. Springer, 181–190.

[33] Christian Meurisch, Cristina A Mihale-Wilson, Adrian Hawlitschek, Florian Giger, Florian Müller, Oliver Hinz, and Max Mühlhäuser. 2020. Exploring user expectations of proactive AI systems. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (2020), 1–22.

[34] Ondrej Miksik, I Munasinghe, J Asensio-Cubero, S Reddy Bethi, ST Huang, S Zylfo, X Liu, T Nica, A Mitrocsak, S Mezza, et al. 2020. Building proactive voice assistants: When and how (not) to interact. *arXiv preprint arXiv:2005.01322* (2020).

[35] Nicole Mirnig, Gerald Stollnberger, Markus Miksch, Susanne Stadler, Manuel Giuliani, and Manfred Tscheligi. 2017. To err is robot: How humans assess and act toward an erroneous social robot. *Frontiers in Robotics and AI* 4 (2017), 21.

[36] Zhaoheng Ni, Ahmet Cem Yuksel, Xiuyan Ni, Michael I Mandel, and Lei Xie. 2017. Confused or not confused? Disentangling brain activity from EEG data using bidirectional LSTM recurrent neural networks. In *Proceedings of the 8th acm international conference on bioinformatics, computational biology, and health informatics*. 241–246.

[37] Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. 2019. Interpretml: A unified framework for machine learning interpretability. *arXiv preprint arXiv:1909.09223* (2019).

[38] Florian Nothdurft, Stefan Ultes, and Wolfgang Minker. 2014. Finding appropriate interaction strategies for proactive dialogue systems—an open quest. In *Proceedings of the 2nd European and the 5th Nordic Symposium on Multimodal Communication*, Vol. 110. Citeseer, 73–80.

[39] Anneli Olsen. 2012. The Tobii I-VT fixation filter. *Tobii Technology* 21 (2012), 4–19.

[40] Mariya Pachman, Amaël Arguel, Lori Lockyer, Gregor Kennedy, and Jason Lodge. 2016. Eye tracking and early detection of confusion in digital learning environments: Proof of concept. *Australasian Journal of Educational Technology* 32, 6 (2016).

[41] Amit Kumar Pandey, Lavindra de Silva, and Rachid Alami. 2016. A human-robot competition: Towards evaluating robots' reasoning abilities for hri. In *Social Robotics: 8th International Conference, ICSR 2016, Kansas City, MO, USA, November 1-3, 2016 Proceedings 8*. Springer, 138–147.

[42] Zhenhui Peng, Yunhwan Kwon, Jiaan Lu, Ziming Wu, and Xiaojuan Ma. 2019. Design and evaluation of service robot's proactivity in decision-making support process. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[44] Ismo Rakkolainen, Ahmed Farooq, Jari Kangas, Jaakko Hakulinen, Jussi Rantala, Markku Turunen, and Roope Raisamo. 2021. Technologies for multimodal interaction in extended reality—a scoping review. *Multimodal Technologies and Interaction* 5, 12 (2021), 81.

[45] P Ramakrishnan, B Balasingam, and F Biondi. 2021. Cognitive load estimation for adaptive human–machine system automation. In *Learning control*. Elsevier, 35–58.

[46] Mao Saeki, Kotoka Miyagi, Shinya Fujie, Shungo Suzuki, Tetsuji Ogawa, Tetsunori Kobayashi, and Yoichi Matsuyama. 2022. Confusion detection for adaptive conversational strategies of an oral proficiency assessment interview agent. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, Vol. 2022. 3988–3992.

[47] Lisa Scherf, Lisa Alina Gasche, Eya Chemangui, and Dorothea Koert. 2024. Are You Sure?-Multi-Modal Human Decision Uncertainty Detection in Human-Robot Interaction. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. 621–629.

[48] Yuki Shiga, Takumi Toyama, Yuzuko Utsumi, Koichi Kise, and Andreas Dengel. 2014. Daily activity recognition combining gaze motion and visual features. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*. 1103–1111.

[49] Maia Stiber, Russell Taylor, and Chien-Ming Huang. 2022. Modeling human response to robot errors for timely error detection. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 676–683.

[50] Maia Stiber, Russell H Taylor, and Chien-Ming Huang. 2023. On using social signals to enable flexible error-aware HRI. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. 222–230.

[51] Aaquib Tabrez, Matthew B Luebbers, and Bradley Hayes. 2020. A survey of mental modeling techniques in human–robot teaming. *Current Robotics Reports* 1 (2020), 259–267.

[52] Arthur Tang, Charles Owen, Frank Biocca, and Weimin Mou. 2003. Comparative effectiveness of augmented reality in object assembly. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 73–80.

[53] Richard Tang, Xing-Dong Yang, Scott Bateman, Joaquim Jorge, and Anthony Tang. 2015. Physio@ Home: Exploring visual guidance and feedback techniques for physiotherapy exercises. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 4123–4132.

[54] Elisabeth Vogl, Reinhard Pekrun, and Kristina Loderer. 2021. Epistemic emotions and metacognitive feelings. *Trends and prospects in metacognition research across the life span: A tribute to anastasia efklides* (2021), 41–58.

[55] Lennart Wachowiak, Peter Tisnikar, Gerard Canal, Andrew Coles, Matteo Leonetti, and Oya Celiktutan. 2022. Analysing eye gaze patterns during confusion and errors in human–agent collaborations. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 224–229.

[56] Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, Neel Joshi, and Marc Pollefeys. 2023. 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)*.

[57] Zhuo Wang, Xiaoliang Bai, Shusheng Zhang, Mark Billinghurst, Weiping He, Yang Wang, Dong Han, Gong Chen, and Jianghong Li. 2021. The role of user-centered AR instruction in improving novice spatial cognition in a high-precision procedural task. *Advanced Engineering Informatics* 47 (2021), 101250.

[58] Andreas Winklbauer, Barbara Stiglbauer, Michael Lankes, and Maurice Sporn. 2023. Telling Eyes: Linking Eye-Tracking Indicators to Affective Variables. In *Proceedings of the 18th International Conference on the Foundations of Digital Games*. 1–11.

[59] Julian Wolf, Quentin Lohmeyer, Christian Holz, and Mirko Meboldt. 2021. Gaze comes in handy: Predicting and preventing erroneous hand actions in ar-supported manual tasks. In *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 166–175.

[60] Tao Xu, Jiabao Wang, Gaotian Zhang, Ling Zhang, and Yun Zhou. 2023. Confused or not: decoding brain activity and recognizing confusion in reasoning learning using EEG. *Journal of Neural Engineering* 20, 2 (2023), 026018.

"Uh, *This* One?": Leveraging Behavioral Signals for Detecting Confusion during Physical Tasks: Supplementary Materials

Maia Stiber
mstiber@jhu.edu
Johns Hopkins University
Baltimore, Maryland, USA

Dan Bohus
dbohus@microsoft.com
Microsoft Research
Redmond, Washington, USA

Sean Andrist
sandrist@microsoft.com
Microsoft Research
Redmond, Washington, USA

A STATISTICAL ANALYSIS OF BEHAVIORAL SIGNAL FEATURES

In order to determine the effect that confusion has on implicit behavioral signals exhibited by participants, we evaluated the dataset using Mann-Whitney U tests and controlling for false discovery rate through Benjamini-Hochberg test. Table 1 displays the numerical values of the Mann-Whitney U tests ran for all of the handcrafted low-level behavioral features extracted, as described in Section 4.

B WITHIN-TASK TEST STATISTICAL TESTS

Table 2 shows the Mann-Whitney U test values calculated for comparing model performance in the within-task test.

C ADDITIONAL RESULTS FOR THE BETWEEN-TASK TEST

Table 3 contains the results of the between-task test for the *Behavior* and *Time + Instructions* inputs.

Table 1: Mann-Whitney U test values (accounting for false discovery rate using Benjamini-Hochberg correction) used to evaluate the significance of behavioral signal features with respect to *confusion* and *no-confusion* intervals. The tests were conducted on the full execution intervals. The bolded features exhibited statistically significant differences across interval type.

		Confusion Interval Mean	No-confusion Interval Mean	Mann-Whitney U Test
Gaze	**Fixations Per Second**	1.22 ± 0.61	1.09 ± 0.55	U = 2495054, p < .001
	Average Completed Fixation Duration	0.76 ± 0.88	0.92 ± 0.96	U = 3276030, p < .001
	SD Completed Fixation Duration	0.60 ± 0.74	0.85 ± 1.00	U = 3359105, p < .001
	Percent Interval Fixation Overlap	0.80 ± 0.16	0.82 ± 0.14	U = 2993328, p = .0038
	Average Gaze Speed	0.56 ± 0.51	0.49 ± 0.38	U = 2363822, p < .001
	SD Gaze Speed	1.12 ± 1.05	1.10 ± 1.08	U = 2626381, p < .001
Head	**Average Head Linear Speed**	0.084 ± 0.092	0.067 ± 0.077	U = 2499418, p < .001
	SD Head Linear Speed	0.069 ± 0.066	0.06 ± 0.061	U = 2603831, p < .001
	Average Head Angular Speed	0.22 ± 0.17	0.19 ± 0.15	U = 2566847, p < .001
	SD Head Angular Speed	0.17 ± 0.14	0.17 ± 0.306	U = 2705490, p = .011
Left Hand	**Average Left Index Speed**	1.9 ± 20.4	1.57 ± 13.8	U = 2597784, p < .001
	SD Left Index Speed	5.65 ± 34.6	5.90 ± 29.7	U = 2708021, p = .013
	Average Left Index DistHead	0.48 ± 0.15	0.47 ± 0.17	U = 2882610, p = .51
	SD Left Index DistHead	0.06 ± 0.17	0.075 ± 0.97	U = 2717302, p = .021
	Average Left Thumb Speed	1.86 ± 20.4	1.53 ± 13.8	U = 2577204, p < .001
	SD Left Thumb Speed	5.59 ± 34.6	5.84 ± 29.7	U = 2688697, p = .0042
	Average Left Thumb DistHead	0.46 ± 0.14	0.45 ± 0.16	U = 2844762, p = .94
	SD Left Thumb DistHead	0.056 ± 0.17	0.071 ± 0.97	U = 2696553, p = .0067
	Average Left Wrist Speed	1.79 ± 20.4	1.48 ± 13.8	U = 2588481, p < .001
	SD Left Wrist Speed	5.49 ± 34.5	5.76 ± 29.7	U = 2683223, p = .0033
	Average Left Wrist DistHead	0.44 ± 0.13	0.44 ± 0.15	U = 2870304, p = .69
	SD Left Wrist DistHead	0.044 ± 0.17	0.061 ± 0.97	U = 2746217, p = .079
Right Hand	**Average Right Index Speed**	1.29 ± 7.14	1.33 ± 6.70	U = 2722990, p = .026
	SD Right Index Speed	4.55 ± 16.2	6.02 ± 24.8	U = 2820876, p = .81
	Average Right Index DistHead	0.50 ± 0.12	0.50 ± 0.141	U = 2892744, p = .39
	SD Right Index DistHead	0.061 ± 0.053	0.065 ± 0.21	U = 2854810, p = .83
	Average Right Thumb Speed	1.25 ± 7.12	1.29 ± 6.69	U = 2729467, p = .036
	SD Right Thumb Speed	4.50 ± 16.2	5.97 ± 24.8	U = 2829903, p = .89
	Average Right Thumb DistHead	0.48 ± 0.12	0.48 ± 0.14	U = 2866402, p = .73
	SD Right Thumb DistHead	0.058 ± 0.053	0.062 ± 0.21	U = 2843412, p = .94
	Average Right Wrist Speed	1.18 ± 7.05	1.23 ± 6.68	U = 2721788, p = .026
	SD Right Wrist Speed	4.39 ± 16.1	5.87 ± 24.8	U = 2817821, p = .79
	Average Right Wrist DistHead	0.46 ± 0.10	0.46 ± 0.13	U = 2857312, p = .81
	SD Right Wrist DistHead	0.044 ± 0.048	0.050 ± 0.21	U = 2932962, p = .079

Table 2: Mann-Whitney U test values (accounting for false discovery rate using Benjamini-Hochberg correction) used to evaluate the difference in model performances for the within-task tests.

	Mann-Whitney U Test			
	ACC	AUC	PR	R-90PR
Mann-Whitney U Test (Comparing to *Behavior*)				
Time	U = 1, p = .014	U = 1, p = .014	U = 1, p = .013	U = 2.5, p = .034
Time + Instruction	U = 25, p = .017	U = 25, p = .014	U = 25, p = .014	U = 25, p = .014
Behavior + Time	U = 17, p = .39	U = 21, p = .12	U = 15, p = .69	U = 15, p = 69
Behavior + Time + Instruction	U = 25, p = .017	U = 25, p = .014	U = 25, p = .014	U = 25, p = .014
Mann-Whitney U Test (Comparing to *Behavior + Time + Instruction*)				
Video (V-Jepa)	U = 6.5, p = .50	U = 20, p = .49	U = 3, p = .33	U = 6, p = .50
ALL (Late Fusion)	U = 11, p = .92	U = 12, p = .98	U = 13, p = .98	U = 10, p = .97
ALL (Concatenate)	U = 11, p = .92	U = 11, p = .84	U = 6, p = .50	U = 1, p = .33

Table 3: Performance of the *Behavior* and *Time + Instructions* in generalizing across physical task (between-task test). The four metrics were accuracy (ACC), area under the curve (AUC), precision (PR), and recall given 90% precision (R-90PR). Results were computed using leave-one-out combination for each task. The bold numbers were the highest numbers for each metric and task.

	Behavior				Time + Instructions			
	ACC	AUC	PR	R-90PR	ACC	AUC	PR	R-90PR
ATV	**0.667**	**0.574**	**0.333**	**0.250**	0.433	0.341	0.200	0
Belt	0.682	**0.692**	0.154	0	**0.727**	0.646	**0.230**	0
Big Printer	**0.775**	0.783	**0.6**	0.174	0.575	**0.839**	0.403	**0.261**
Circuit Breaker	**0.600**	**0.475**	0.083	**0.167**	0.425	0.412	**0.130**	0
Coffee	0.667	**0.889**	0	**0.667**	**0.778**	0.861	**1**	**0.667**
DSLR	**0.584**	**0.688**	**0.247**	0.031	0.579	0.548	0.200	**0.094**
Espresso	0.852	0.793	0.714	0	**0.889**	**0.871**	**0.833**	**0.714**
Gopro	**0.753**	**0.853**	**0.870**	**0.268**	0.690	0.737	0.595	0.232
Graphics Card	**0.571**	**0.556**	0.200	0	0.535	**0.556**	**0.25**	**0.111**
Navvis	0.480	0.549	0.167	0	**0.680**	**0.667**	**1**	**0.111**
Nespresso	0.607	0.611	**0.667**	**0.043**	**0.647**	**0.689**	0.619	0
RAM	**0.655**	**0.552**	**0.364**	**0.043**	0.621	0.532	0.167	0
Small Printer	**0.570**	**0.739**	**0.828**	0	0.561	0.718	0.778	**0.303**
Switch	0.635	**0.850**	0.667	0	**0.785**	0.723	**0.803**	**0.519**