Attune: A Self-Annotation Tool for Understanding Robot Operator Attention Profiles

Puqi Zhou
Department of Computer Science
George Mason University
Fairfax, VA, USA
pzhou@gmu.edu

Sungsoo Ray Hong
Department of Information Sciences and Technology
George Mason University
Fairfax, VA, USA
shong31@gmu.edu

David Porfirio
Department of Computer Science
George Mason University
Fairfax, VA, USA
dporfiri@gmu.edu

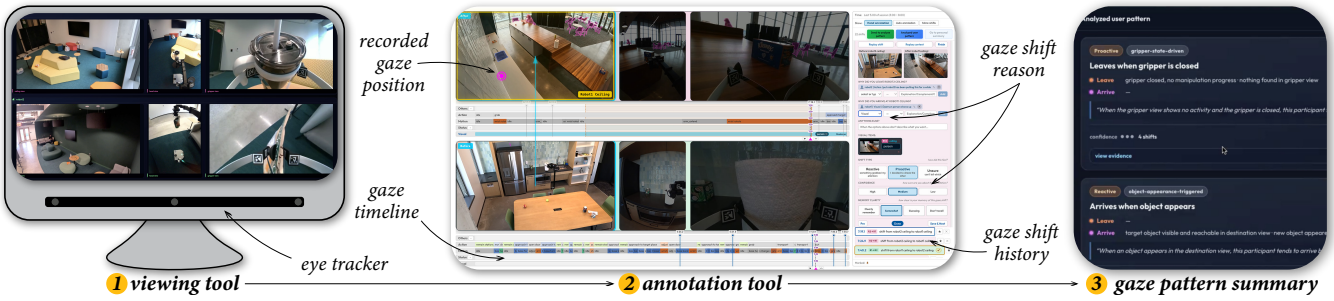


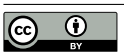
Figure 1: *Attune* enables robot operators to self-annotate attentional shifts and functions as follows: 1 an operator watches robots while *Attune* records eye gaze; 2 the operator plays back their gaze shifts and retroactively attributes causes to these shifts; and 3 *Attune* co-annotates additional shifts and extracts patterns to be presented to the operator for confirmation.

Abstract

Deploying robot fleets in complex, real-world environments requires human operators to supervise multiple robots simultaneously. Managing operator attention is a fundamental challenge of designing multi-robot supervision interfaces, encompassing both feed layout and feed content (i.e., robot *behavior* design). Thus far, designers lack empirical guidance on the latter—how to change a robot's behavior to capture, sustain, or relinquish operator attention during multi-robot supervision. In our vision of the future, designers should be able to use this guidance to calibrate robot behavior to different operator attention profiles. Treating operator eye gaze as a robot behavior design clue, we created a pre-deployment elicitation tool called *Attune*. *Attune* automatically identifies *when* meaningful gaze shifts occur, provides AI assistance for annotating *why* shifts occurred, and outputs a summary of operator gaze patterns for operator review. We evaluated *Attune* through a user study in which participants annotated the visual triggers that drew their attention. Our findings unveil variation in observed gaze patterns and reveal how *Attune* helps characterize operator attention.

Keywords

human-robot interaction, eye gaze, attention, annotation



UIST '26, Detroit, MI, USA

ACM ISBN 979-8-4007-2856-3/2026/11
https://doi.org/10.1145/3830398.3830514

ACM Reference Format:
Puqi Zhou, Sungsoo Ray Hong, and David Porfirio. 2026. *Attune*: A Self-Annotation Tool for Understanding Robot Operator Attention Profiles. In *The 39th Annual ACM Symposium on User Interface Software and Technology (UIST '26), November 02–05, 2026, Detroit, MI, USA.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3830398.3830514

1 Introduction

Multi-robot fleets are increasingly common on college campuses [29] and hospitals [15] for making deliveries, in warehouses for stocking and sorting items [73], cities and rural areas for searching for missing people [65], in addition to many other environments. While each robot in the fleet is mostly autonomous in practice, robot *operators* are still needed to supervise and manage each robot and to intervene or troubleshoot when necessary. Often, a single operator is tasked with monitoring multiple robots and must maintain attention to key events and intervene in a robot's operation. Many interfaces that facilitate this monitoring consist of multiple adjacent robot camera feeds streamed to a single screen [25, 74].

The design of such interfaces is crucial to maintaining operator performance, which prior work indicates is affected by operator stress, workload, and attention [57, 61]. Attention, in particular, has led to efforts to improve operator interface design [12], although this is difficult because operators possess individual differences due to personal psychology, task-related attentional biases, or being influenced by the setting within which the operator resides [66, 69]. Multi-robot viewing interfaces should therefore be customized for individual operator attention needs.

The first step towards customizing a multi-robot viewing interface for a specific operator is to understand the operator's attention

profile. For this paper, we refer to an *attention profile* as an operator-reviewed collection of recurring gaze-shift patterns, grounded in the operator's retrospective rationales and the robot and environmental context surrounding those shifts. In order to assess the operator's individual attention profile, including what interface features and robot actions they are most prone to looking at or being distracted by, we envision a *calibration step* that occurs when multi-robot fleets are deployed in a particular setting. Using this information, a designer can create or customize an interface tailored specifically to the operator or tweak the robots' behaviors in order to appropriately allocate operator attention and mitigate distractions. In this paper, we take a first step towards the latter, characterizing how robot behavior coincides with gaze shifts and how calibration can structure these shifts into attention profiles.

We therefore constructed *Attune*, a pre-deployment elicitation tool that facilitates this calibration step by helping characterize cross-robot gaze shifts on minimal, two-robot configurations, leaving larger configurations to future work. Shown in Figure 1, *Attune* first uses eye tracking to record operator gaze during a two-robot video observation step, selects *valid* gaze shifts, and then presents the operators with an *annotation tool* that guides them in retroactively attribute possible causes to a subset of these shifts. *Attune* then uses large language models (LLMs) to further annotate valid shifts and produce a summary of the operator's attention profile.

Our evaluation of *Attune* uses the two-robot setup as a foundational first step for validating the general feasibility of our proposed workflow, including cross-robot gaze shift production and subsequent annotation. Specifically, we conducted a user study that asks (**RQ1**) how should annotation tools be designed to elicit retroactive causal attributions for gaze shifts and (**RQ2**) what are the characteristics of gaze shifts during multi-robot supervision. Our results highlight the benefits of structured replay and contextual scaffolding and uncover both shared and variable gaze patterns between participants and scenarios. Our contributions are:

- *System—Attune*, a tool that elicits gaze-shift attributions.
- *Empirical*—an evaluation of *Attune*'s ability to capture gaze shift attributions, operator gaze patterns, and usability.
- *Design*—design implications for improving self-annotation tools for achieving our vision of operator gaze calibration.

2 Related Work

This work draws from prior research on human attention, interfaces for robot operators, and interfaces for data annotation.

2.1 Human Attention

The *Premotor Theory of Attention* ties the preparation of eye motor movement to attentional changes, thus linking gaze shifts (*saccades*) to attention [53]. In the context of human-robot supervisory control, there are several factors that predict attentional shifts towards different *areas of interest (AOIs)*—the *salience* of the AOI (e.g., its visual distinctiveness), the *expectancy* of observing events at a particular AOI, the *value* of looking at the AOI, and the *effort* of the shift [66–68]. Crucially, however, eye tracking alone cannot determine the cause of a shift. Furthermore, attention shifts can be a "top-down" (e.g., knowledge-driven) process [66], and can thus differ from person to person depending on an individual's own goals and experiences [69]. These limitations provide a motivation for using *Attune* to rationalize, or attribute possible causes to, gaze shifts. There is a substantial amount of prior work on both predicting and leveraging attention to manage human-centered computing [70]. Recently, Pei et al. [44] created *AttentionAR*, an augmented-reality interface that monitors user attention and strategically overlays hazard alerts in the user's headset. Managing human attention is also highly relevant in human-robot interaction, where robot behavioral design is critical to managing operator attention [42]. Prior work shows how to measure [32], interpret [58], guide [40, 51], and respond to [71] human attention shifts. As further justification for our own work, Saad et al. [55] show that robot behaviors can be chosen to appropriately manage human attention.

2.2 Interfaces for Robot Operators

Robot operators encompass a broad range of skilled professionals whose responsibilities depend on contextual expectations and the autonomous capabilities of the robot being operated, ranging from direct control (i.e., teleoperation) [19, 24, 46, 64] to monitoring semi-autonomous systems with only minimal intervention needed [5]. Interventions are often physical, such as clearing obstacles [47] and preventing bystander harm [6]. This paper instead focuses on *remote* monitoring [18, 31], which often occurs during the deployment of semi-autonomous "sidewalk" delivery robots [26, 29, 62], search-and-rescue robots [1], and robots that traverse areas that are hazardous to humans [11]. In mature deployments, operators are immersed in camera feeds for extended windows under either continuous monitoring [10, 21, 74] or manage-by-exception alert response paradigms [9, 25, 33]. *Attune* does not assume a fixed monitoring duration after calibration, intending to apply to both sustained monitoring and brief (e.g., alert-triggered) windows.

Interfaces must be carefully designed for these operators. In human-robot interaction, operator performance is uniquely affected by how the robot presents its autonomy, the reliability of its sensors, and the duration of its behaviors, among other aspects [16, 17]. Roy et al. [54] find that how the robot presents its autonomy to a remote operator affects situation awareness. Poorly designed interfaces that fail to manage these factors can lead to high operator workload, poor memory retention, misallocated attention, and, in general, reduced operator effectiveness [49]. This has led to the careful development of remote viewing interfaces such as Zhou et al. [74]'s multi-video ground robot sensemaking interface.

2.3 Behavioral Annotation Interfaces

Behavioral annotation draws from a long lineage of interfaces for qualitative analysis, much of which focuses on analyzing text content [2, 14, 34]. Our work instead draws from multimodal (e.g., video and audio) human *behavioral* annotation. *BORIS* [23] is one example, which enables annotators to assign qualitative codes to video and audio content. *DOTE* [37–39] and *ELAN* [36, 59] serve a similar purpose, and are popular in the human-centered computing community for conducting conversational (and more generally *interactional*) analyses between humans or between humans and robots [43, 45, 48]. Beyond *ROSAnnotator*, few such tools exist specifically for human-robot interaction [72]. None discussed thus far are intended for annotating human eye gaze behavior.

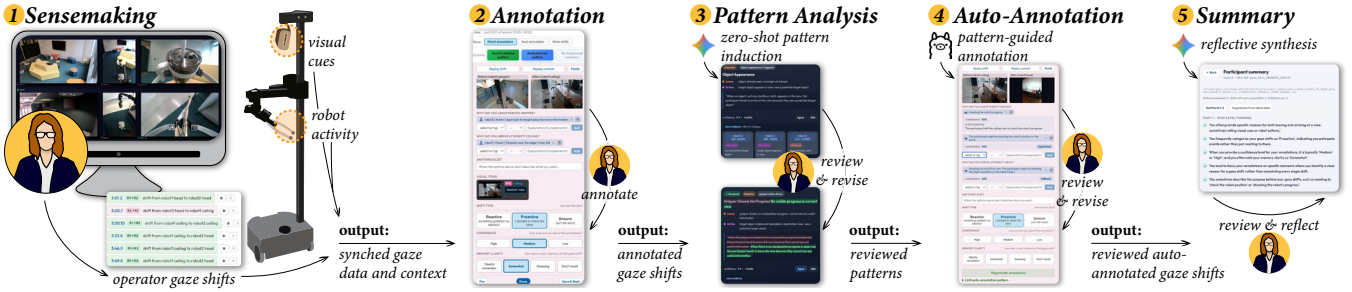


Figure 2: The *Attune* pipeline, with the output being a synthesized attention profile summarized for the operator to review.

Instead, we require an interface that (1) tracks operator eye gaze behavior; (2) automatically identifies significant gaze shifts (or *saccades*) versus stationary gaze fixations; (3) assists the operator in attributing possible causes to a subset of these saccades; and (4) automatically attributes causes to previously unseen saccades. Prior work satisfies subsets of these requirements. Existing research already addresses the binary classification between fixations and saccades [28, 63]. Similarly, classifying *what* a viewer fixates on, though laborious without assistance from eye gaze trackers [60], is usually an automatic process [13]. The *eyeNotate* tool, for example, semi-automatically classifies *areas of interest* from a viewer's gaze *fixation events* [4]. In addition to mapping fixation events to real-world targets, *gazeMapper* [41] facilitates analysis of gaze behavior in multi-party settings. With all of these aforementioned tools, however, the causes of saccades are unknown. We specifically seek a solution that enables viewers themselves (the *operators*) to rationalize both top-down and salience-driven saccades.

3 System Design

We designed *Attune* as a pre-deployment operator calibration facilitator for multi-robot monitoring, focusing on a minimal (two-robot) variant. The system combines multi-video monitoring, eye-tracking-based gaze recording, and both manual and LLM-assisted annotation in order to capture where operators look, and crucially, help them explain *why* they shifted their attention. Through the five phases shown in Figure 2, we aim to surface individual operator attention profiles, or their self-confirmed accounts of recurring gaze shift rationales. These phases form a one-time, human-in-the-loop, pre-deployment calibration step: (1) video sensemaking to capture operator gaze; (2) gaze-shift annotation; (3) cross-shift pattern analysis; (4) pattern-guided auto-annotation generation; and (5) synthesis and summarization of the operator's attention profile.

Scenario. We use a scenario to assist in our description of *Attune*. Consider a company that is deploying a fleet of robots at a hospital. The robots make deliveries between staff and patients, while a *robot operator* observes the fleet remotely and intervenes when necessary. It is crucial that the multi-robot supervision interface used by the operator effectively manages operator attention. For example, if one robot in the fleet is performing a high-risk task, the others should not draw the operator's attention. A designer who works for the company is responsible for ensuring that the behaviors of each robot in the fleet are appropriately calibrated to the operator's attention profile and uses *Attune* as a calibration facilitator.

3.1 Phase 1: Multi-Robot Video Sensemaking

Figure 3 depicts our sensemaking interface, which supports viewing two robots concurrently executing independent tasks. Each robot is accompanied by three synchronized camera views from the *ceiling*, the robot's *head* point-of-view (POV), and its *gripper* POV, ordered from left to right. In total, this yields six synchronized video streams. This design is intended to facilitate the characterization of the operator's gaze shifts both within the same robot (*cross-view* shifts) and across different robots (*cross-robot* shifts).

While the operator monitors the robot feeds, *Attune* records their eye gaze. Gaze is saved both as *pixel-level* gaze records (e.g., the exact position of eye gaze on the monitor) and *discretized* gaze records that specify which video region is being viewed at each moment. Gaze is synchronized with video playback to enable direct correspondence between gaze behavior and video events.

In addition to video time, gaze is aligned with discrete *robot activity*, inspired by the hierarchical abstractions that are common in the task planning literature [22, 27]. Shown in Table 1, the different levels of robot activity include (a) the high-level goal currently being pursued by the robot; (b) the subgoal that the robot is currently executing in pursuit of the larger goal; (c) the current action being executed; (d) the robot's low-level motion; and (e) the robot's joint state. In practice, this information can come from the robot itself.

In addition, *Attune* uses YOLO [52] to extract and segment visual cues from each video, which are stored as object labels and maskable

Table 1: Our hierarchical definition of robot activity.

Rob. Activity	Labels
Goal	Domain specific. Examples include *tidy kitchen.*
Subgoal	Domain specific. Examples include *wash dishes.*
Action	*Open, close, grab, carry, drop, navigate, approach, transport, wait, notify, activate, deactivate, pull, push, adjust [limb], change view, remain stationary.*
Motion	*Gripper open/close, base movement, arm retract/extend, arm lift/down, head move, wrist rotate, idle.*
Gripper State	Discrete based on openness: *closed, partially open, open.*
Wrist State	Pitch, yaw, and roll discretized into combined labels: *neutral, up, down, left, forward, right, (strong) rolled, stowed.*
Head State	Combinations of pan and tilt: *(far) left, forward, (far) right, back, (strongly) down, level, (strongly) up.*

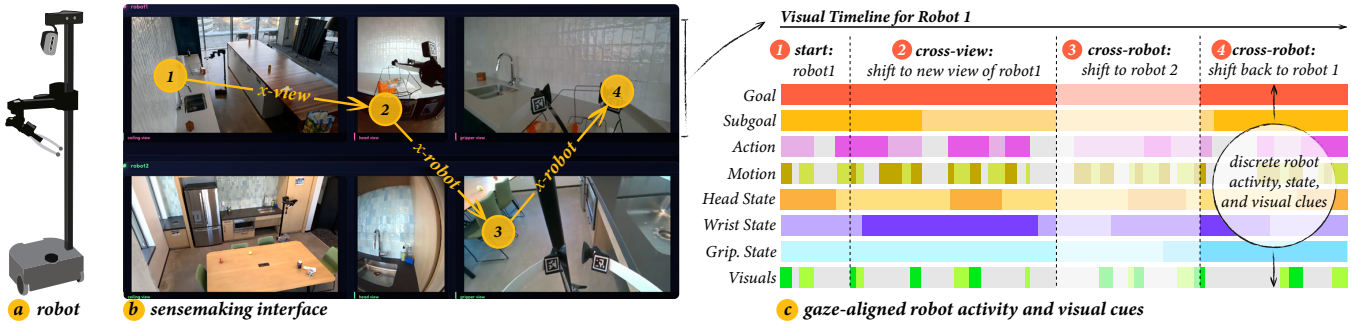


Figure 3: The sensemaking interface shows two a robots conducting tasks simultaneously, in which both robots stream b three video feeds to the interface. The interface tracks operator gaze, including cross-robot (x-robot) and cross-view (x-view) gaze shifts, which is aligned with c robot activity and visual cues.

Algorithm 1 Selecting Gaze Shifts for Review

1: Split video V into temporally equal bins $v \in \mathcal{V}$
2: $\mathcal{P}, \mathcal{M}, \mathcal{A} \leftarrow [\,]$ ▷ Curated, manual, and auto pools
3: **for all** $v \in \mathcal{V}$ **do**
4: **skip** v if no valid shift exists for either type
5: **repeat**
6: $g_r, g_v \leftarrow \textsc{SampleShifts}(v)$ ▷ cross-robot/view
7: **until** $\textsc{Curate}(g_r) \wedge \textsc{Curate}(g_v)$ ▷ See Table 2
8: $\textsc{Append}(\mathcal{P}, (v, g_r, g_v))$
9: **end for**
10: $\mathcal{P} \leftarrow \textsc{Shuffle}(\mathcal{P})$
11: Put m shifts of each type in $\mathcal{M}$ ▷ Manual pool
12: Put n shifts of each type in $\mathcal{A}$ ▷ Auto pool

visual regions. As a result, each gaze shift can be interpreted with reference not only to video time, but also to ongoing, discrete robot behaviors and visible task-relevant objects at that time.

Sensemaking in Action: In our scenario, the operator sits in front of a computer monitor and watches both robots conduct different tasks. Shown in Figure 3, *Attune* records and maps their gaze to robot activity and visual cues.

3.2 Phase 2: Data Annotation Interface

After the sensemaking session, the operator enters the *annotation tool* shown in Figure 4, a dedicated workspace for retroactively rationalizing gaze shifts. This phase is designed specifically to occur after (i.e., rather than alongside) the sensemaking phase to avoid disrupting natural operator gaze patterns. The annotation tool merges valid gaze shifts from Phase 1—cross-robot and cross-view—into a time-sorted list. For each robot, the operator can inspect time-aligned tracks beneath the videos that show contextual information: goal, subgoal, action, motion, joint state, and visual cues.

The first part of annotation is automatic *gaze shift selection*, followed by operator *shift inspection and replay*, which is facilitated by *nonlinear timeline magnification*. The operator then engages in *structured annotation* to retroactively attribute causes to gaze shifts. Note that structured annotation is subject to memory degradation.

Gaze Shift Selection: *Attune* automatically selects and curates gaze shifts for the user to annotate from the set of observed gaze shifts. In order to capture the operator's stable viewing rhythm while reducing early-session novelty effects and recall decay, selection and curation only occur within the most recently viewed segments, specifically the final five minutes of each session. Algorithm 1 depicts the selection procedure. *Attune* first partitions the videos into temporally equal bins (Line 1). *Attune* then samples (Line 6) and attempts to curate (Line 7) a cross-robot and cross-view gaze shift from within each bin, where curation is successful if any of the five rules in Table 2 apply for both shifts. Curated gaze shifts include shifts from one camera view (A) to another (B) where gaze dwells on both the source view A and destination view B for at least 100ms (a typical minimum gaze fixation duration based on Salvucci and Goldberg [56]). Gaze episodes shorter than 100ms on either view are excluded as saccadic pass-throughs or transitional movements. Sometimes, gaze shifts originate from or target portions of the interface that are outside of the six camera views. We designate these as *bridge shifts* to or from an unknown region. Bridge shifts from A⟶ unknown or unknown⟶ B are successfully curated if the time spent in the unknown region is at least 200ms (an operational threshold informed by Manor and Gordon [35]). Most shifts are sampled randomly to preserve ecological validity; a small subset is quality-anchored by selecting shifts with the longest dwell-before durations, ensuring the inclusion of high-confidence, unambiguous shifts alongside temporally representative ones.

Attune divides curated shifts into a set $\mathcal{M}$ for manual annotation (Line 11) and $\mathcal{A}$ for auto-annotation (Line 12, see Phase 4). Each set contains an equal number of cross-robot and cross-view shifts. *Attune* overlays gaze-shift indicators on the timeline to help operators quickly locate moments of attention transition.

Shift Inspection and Replay: *Attune* displays $\mathcal{M}$ on the timeline and in a list at the bottom right side of the screen from which operators can select curated gaze shifts. Selecting a shift opens an annotation card in the right panel (Figure 4 a). *Attune* then presents the shift timestamp (Figure 4 b), the *from* and *to* camera views (Figure 4 c), and the temporally aligned robot context surrounding the shift (Figure 4 d). To support recall, the system offers two replay modes: (1) a *replay shift* that focuses on the transition between *from* and *to* views, and (2) a *replay context* that begins 5s before the shift and ends 0.5s after (Figure 4 e). During both replays, the *from* and *to* views receive directional overlays and the

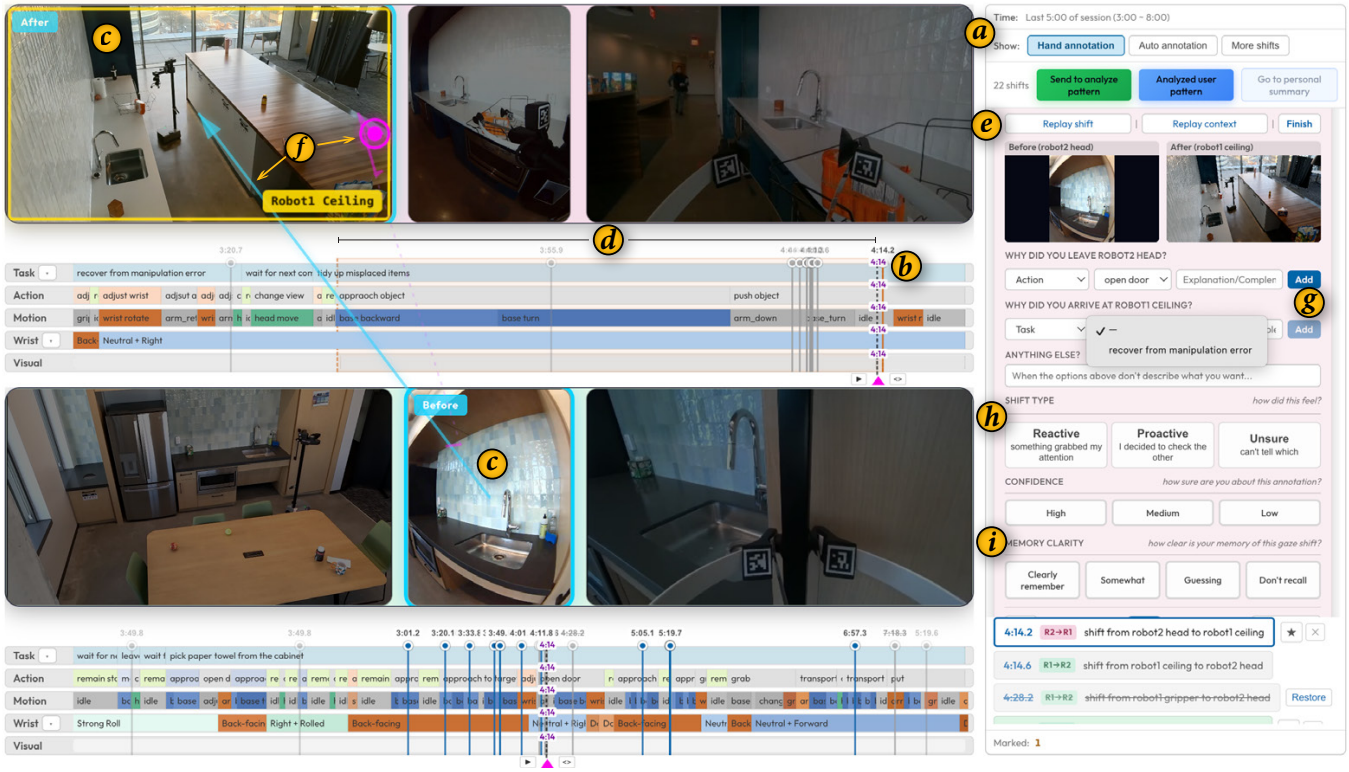


Figure 4: The annotation tool. The annotation card a enables operators to annotate specific shifts b that occur between different robot views c. Operators can control replay during annotation (d and e), view their recent gaze shift (the blue arrow) and gaze point (the pink dot) f, attribute a cause for the shift (g and h) and their confidence i.

Table 2: Rules for curating valid gaze shifts. *A*, B, and *C* are different robot views (cross-view or cross-robot). *Unknown* denotes areas of the screen that do not correspond to a view.

Raw shift	Curated shift
A ($\geq$ 100 ms) $\rightarrow$ B ($\geq$ 100 ms)	$A \rightarrow B$
A ($\geq$ 100 ms) $\rightarrow$ B ($<$ 100 ms) $\rightarrow$ C ($\geq$ 100 ms)	$A \rightarrow C$
A ($\geq$ 100 ms) $\rightarrow$ *unknown* ($<$ 200 ms) $\rightarrow$ B ($\geq$ 100 ms)	$A \rightarrow B$
unknown ($\geq$ 200 ms) $\rightarrow$ B ($\geq$ 100 ms)	*unknown* $\rightarrow B$
A ($\geq$ 100 ms) $\rightarrow$ *unknown* ($\geq$ 200 ms)	$A \rightarrow$ *unknown*

gaze point and recent trajectory are drawn on screen (Figure 4 f). The remaining four views are dimmed.

Nonlinear Timeline Magnification: To further assist with causal attribution, the system applies a piecewise nonlinear temporal mapping over the interval $[t-90\text{s},\ t+30\text{s}]$. The earliest segment $[t-90, t-30]$ occupies 30% of the timeline width; the immediately preceding segment $[t-30, t]$ occupies 60% and is highlighted with an orange overlay; and the post-shift segment $[t, t+30]$ occupies the remaining 10%. *Attune* thus magnifies the 30s most likely to contain the causal event while preserving earlier history and post-shift context in compressed form for a quick overview.

Structured Annotation: Within an annotation card, the operator provides structured explanations for why they left the previous view (*why leave*) and why they arrived at the new view (*why arrive*), retroactively drawing causal links between their gaze shift and the contextual information that may have drawn their shift (Figure 4 g). *Attune* also offers a free-response text box if none of these action or visual categories apply. Operators also label the *shift type*, or whether their gaze shift was driven by "bottom-up" (labeled *reactive* gaze shifts by the interface, reflecting involuntary shifts) or "top-down" (labeled *proactive* gaze shifts by the interface, intended to reflect voluntary gaze shifts) mechanisms (Figure 4 h). Operators additionally provide confidence of their annotations and a rating of their memory clarity (Figure 4 i). Overall, the annotation card organizes operator reasoning and preserves the flexibility to capture nuanced or ambiguous motivations in free text.

For attributing non-robot visual cues to gaze shifts, operators can select YOLO-segmented objects from the relevant video frame, re-label objects if necessary (e.g., if misclassified), and draw masks over unclassified areas of the view to select and label objects not caught by YOLO. Operators can then attach the resulting cue to the annotation card. Figure 5 depicts visual annotation.

Annotation in Action: In our scenario, after the monitoring session, *Attune* guides the operator to review a curated set of their gaze shifts selected from the final few minutes of recording. From the sample, the operator selects a shift in which they moved their

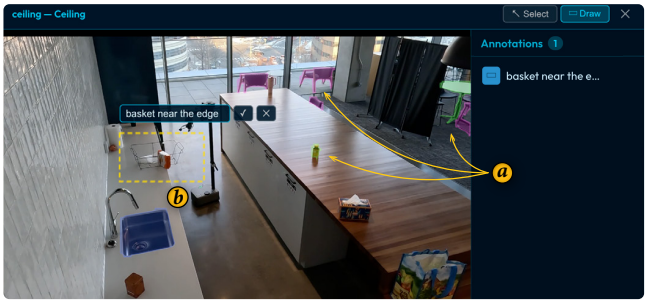
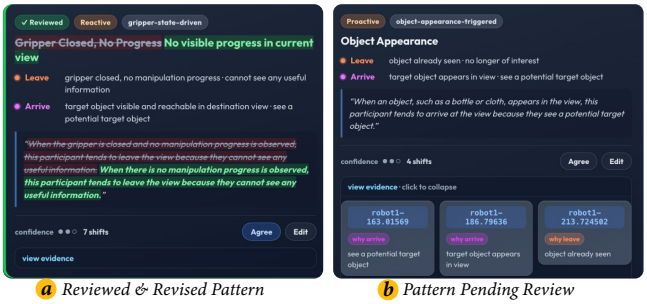


Figure 5: *Attune* highlights YOLO-recognized objects a and enables operators to label unclassified regions b.



a *Reviewed & Revised Pattern* b *Pattern Pending Review*

Figure 6: A pattern reviewed and corrected by the operator a, and a pattern that is pending operator review b.

gaze from one robot's gripper camera to another robot's head camera, and uses the replay tool to re-examine the transition. After watching the segment, the operator explains in the annotation card that they left the gripper view because the manipulation had completed, and arrived at the head view because the second robot appeared to be approaching a person. *Attune* records this structured explanation alongside any robot and visual context at the shift.

3.3 Phase 3: Attention Profile Pattern Analysis

We define a *pattern* as an operator-specific, recurring explanatory structure for multiple gaze shifts—for example, leaving views when manipulation actions complete, or arriving at new views when task-relevant objects become visible. Figure 6 depicts our pattern analysis interface. Once an operator has annotated a set of gaze shifts, the system supports abstraction from individual cases to higher-level viewing patterns.

Pattern Inference: The *Attune* backend organizes the annotated shifts into a structured payload containing the why-leave and why-arrive rationales, robot activities, visual cues, gaze-timing features, and operator ratings, and submits it to a zero-shot LLM-based clustering stage (Gemini-2.5-Flash).[1] The LLM clusters shifts primarily by the semantic similarity of operators' *why-leave* and *why-arrive* explanations, grounds the resulting groups against shared robot and visual context, and produces candidate patterns accompanied by scope descriptions, supporting evidence, and LLM confidence. Memory-clarity and self-confidence ratings from Phase 2 serve as auxiliary reliability signals rather than primary clustering criteria.

Operator Review and Refinement: The system presents candidate patterns to the operator for review and refinement. For each pattern, the interface displays the LLM-generated description alongside the supporting annotation cards, allowing the operator to replay individual gaze shifts and revisit their original reasoning before accepting or modifying the pattern. The operator may revise any patterns and confirm those that accurately represent their monitoring strategy. Approved patterns are saved to a *reviewed pattern library*, which is carried forward in subsequent phases.

Pattern Analysis in Action: In our scenario, *Attune* infers candidate patterns from the operator's annotated shifts and presents each one with its supporting annotation cards. The operator can inspect the evidence, replay the linked gaze shifts, and verify whether the inferred description matches their original reasoning. If needed, they can revise the pattern before saving it. Figure 6b shows a pattern pending review, while Figure 6a shows a reviewed pattern and how the operator can refine the description. Confirmed patterns are added to the attention profile.

[1]The prompts and code for *Attune* are available at https://ari-lab-gmu.github.io/Attune_UIST26/.

3.4 Phase 4: Pattern-Guided Auto-Annotation

As a continuation of the one-time, human-in-the-loop calibration, *Attune* then uses the operator's attention profile pattern to accelerate annotation of additional gaze shifts, using LLM assistance to reduce annotation burden while engaging operators for verification. Specifically, *Attune* propagates the operator-annotated shifts and operator-confirmed patterns to unseen shifts selected from the auto-annotation set $\mathcal{A}$, rather than inferring them from scratch. For each shift that the operator selects from $\mathcal{A}$, *Attune* constructs a structured prompt combining local shift context, pre- and post-shift robot state summaries, YOLO-derived visual cues, gaze timing features, and the operator's reviewed pattern library. This prompt is submitted to a fast, low-cost LLM (Groq Llama-3.3-70B-Versatile),[1] which proposes candidate annotations for *shift type*, *why leave*, and *why arrive*, each accompanied by a confidence level and a brief explanation of the reasoning behind the suggestion.

The generated annotations are treated as editable suggestions rather than final labels. Within each annotation card, the operator can auto-generate or regenerate annotations at any time, inspect the proposed reasoning, and accept, modify, or discard each suggestion. The operator retains full responsibility for the final interpretation and must still provide confidence and memory-clarity judgments after reviewing the generated content.

Auto-Annotation in Action: Auto-annotation is intended to be much faster than manual annotation. In our scenario, the operator selects a new gaze shift from the auto-annotation pool. The system retrieves the shift context and matches it to the reviewed pattern library. The operator inspects the suggested reasoning, confirms, edits, or deletes the *why leave* explanation, and revises the *why arrive* explanation to reflect that the transition was triggered by a visible object rather than task anticipation. The amended annotation is saved alongside its confidence and memory clarity ratings.

3.5 Phase 5: Personal Summary and Reflection

In the final phase, *Attune* synthesizes the operator's complete annotation history (both manually and auto-annotated shifts) and

the reviewed pattern library to produce *Attune*'s final output: an *operator-facing summary* generated with Gemini-2.5-Flash.[1]

The first part is a *reflective annotation summary* that characterizes the operator's annotation style—for example, how consistently they applied structured labels, how often they relied on free-text reasoning, and the distribution of their confidence and memory-clarity ratings across shifts. This gives operators a transparent picture of how they engaged with the annotation process itself. The second part presents the operator's *attention profile summary*, describing the reviewed patterns that characterize their monitoring strategy—for example, which robot states or visual events most reliably triggered attention shifts, and whether the operator tended toward cross-robot or cross-view transitions. This gives operators a consolidated view of their own attention tendencies as captured and verified across Phases 2 and 3. The third part shows explicitly how *Attune* operationalized those patterns during auto-annotation in Phase 4—surfacing which patterns were invoked most frequently, how they mapped onto the auto-generated *why leave* and *why arrive* suggestions, and which suggestions the operator accepted, revised, or discarded. Within the same part, a shift-level view further presents, for each shift, how the operator engaged with the auto-annotation, including whether they accepted, revised, or discarded the suggestions, and what those choices reveal about the coverage and limitations of the inferred pattern library. We aim to help operators understand how the system used their attention profile and where the inferred patterns may need further refinement.

Summary and Reflection in Action: After completing the auto-annotation phase, *Attune* presents the operator with their personal summary. The operator reviews the summary, inspects which patterns drove the auto-annotations in Phase 4, and reflects on where they accepted or revised the generated suggestions. The operator edits the summary to note any refinements, and the updated record is saved to their personal summary, completing the calibration session. This summary is then delivered to the designer.

3.6 Implementation

We implemented *Attune* as a desktop application[1] integrated with a stationary Tobii Pro Spark (60 Hz) eye tracker. The system consists of a React/TypeScript frontend and a FastAPI backend that communicates with the eye tracker through a Tobii SDK. The frontend supports synchronized video playback, gaze replay, annotation, timeline visualization, and summary presentation. In the current version of *Attune*, video playback is pre-recorded. The backend handles gaze recording, annotation persistence, LLM orchestration, and staged output storage. The choice to use two different LLMs rather than one reflects a quality-latency trade-off: Phase 3 clusters each operator's shifts once per session and Phase 5 generates a full operator summary, so they benefit from stronger semantic reasoning (Gemini-2.5-Flash), whereas Phase 4 generates per-shift suggestions at high throughput, favoring a faster, lower-cost model (Llama-3.3-70B). The choice to use LLMs rather than a vision-language model (VLM) is motivated by the visual context being already symbolized—gaze shifts are reasoned over YOLO-derived object labels and annotated robot activity rather than raw pixels. VLM integration, conditioned on operator-reviewed visual patterns, is left for future work.

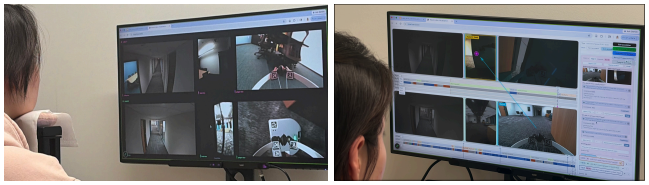


Figure 7: The setup for our evaluation.

4 Evaluation

We evaluated *Attune* with an IRB-approved user study in which participants used *Attune* to monitor robots and annotate their gaze.

4.1 Procedure

Prior to the study, we constructed a video dataset comprising three scenarios involving a physical Stretch 3 robot [30]—*kitchen activities*, *living room activities*, and *hallway missing person search*. Each scenario was recorded twice from three synchronized vantage points—ceiling, robot head, and robot gripper—resulting in three pairs of synchronized video triplets, ranging from 35 to 83 minutes in length. Videos in the same pair were captured on nearby days and at similar times of day to preserve comparable lighting conditions and overall scene appearance. By simultaneously playing videos within a pair, we mimic multi-robot supervision of two robots.

After the consent stage and eye-tracker setup, the study consisted of three phases: *monitoring*, *annotation*, and *review*, depicted in Figure 7. During *monitoring*, each participant used *Attune* to watch eight minutes of video from one scenario. We assigned four participants per scenario. For each participant, the position of each robot on the sensemaking interface (i.e., top or bottom) and the starting timestamps for the eight-minute video duration were randomized, resulting in a unique combination of sensemaking videos per participant. Our choice to have a unique combination per participant reflects a deliberate breadth-over-depth study design that reduced within-scenario statistical power but provided broader qualitative coverage and limited overfitting to any single setting. Participants were instructed to verbalize aloud to the robot in order to keep them engaged with the videos and prevent them from losing focus. Next, during *annotation*, participants reviewed and annotated the rationales behind a subset of curated gaze shifts from $\mathcal{M}$ (5-8 shifts) and $\mathcal{A}$ (5 shifts), including whether the shifts were reactive or proactive, and their annotation confidence. They also rated memory clarity for each annotation. As described in §3, the *annotation* phase was designed to occur *after* (i.e., not co-occur with) the monitoring phase in order to avoid disrupting participants' natural gaze behavior. Finally, during *review*, participants reviewed the AI-generated patterns, auto-annotated gaze shifts, and their participant-facing summary, and then answered questions about their experience.

4.2 Participants

We recruited 12 naïve participants (8 male, 4 female) from the Mason Square campus at George Mason University. Four participants wore corrective lenses; all reported normal or corrected-to-normal vision, and none had prior experience with gaze-based annotation systems. Participants were compensated at $30 USD per hour.

4.3 Data and Measures

Our data consists of: (1) eye gaze recorded continuously via the eye tracker at 60 Hz; (2) user annotations, including why-leave and why-arrive responses, shift types, confidence ratings, and memory clarity; (3) interviews via phase-embedded questions, where after each phase of the study, we asked participants a targeted question corresponding to that phase and, if time permitted, additional questions at the end of the session; (4) audio; and (5) questionnaires.

We checked the audio to characterize participants' verbalization behaviors and analyzed participant gaze patterns to assess individual differences and group-level similarities. We relied on a custom questionnaire, participant annotations, and interview data to assess *Attune*'s ability to elicit causal attributions for gaze shifts. Lastly, we measured usability and usefulness via both the System Usability Scale (SUS) [8] and interview data.

4.4 Results

4.4.1 Verbalization Behaviors. All participants verbalized aloud to comment on both robots during the monitoring session. Participants adopted different strategies: some spoke to the robots continuously as if they were companions, whereas others only verbalized when they noticed something worth warning about. Because speech production and eye gaze behavior occurred concurrently, these verbalization strategies may have affected participants' observed gaze patterns. It is therefore necessary to interpret participant-level differences as conditioned by our monitoring and verbalization protocol rather than as definitive evidence of natural attentional traits. Indeed, one participant noted that focusing on speaking made it harder to later recall gaze shift rationales.

4.4.2 Gaze Patterns. The following analyses characterize observed gaze allocation, but do not adjudicate whether observed individual shifts are primarily task-driven or salience-driven, as our study does not assume a normative gaze pattern for these open-ended monitoring tasks, nor does it include independent task-based or visual-saliency benchmarks. A normative ground truth would require a separate causal model built from larger expert data. Figure 8 shows how participants distributed visual attention across *Attune*'s sensemaking interface. Rather than monitoring all six views uniformly, participants appeared to concentrate on a subset of camera regions within each scenario, forming clear hotspots. Meanwhile, gaze patterns varied across participants, with some participants focusing attention on a few views, whereas others distributed attention more broadly. Across the three scenarios, the spatial distribution of attention shifted as well.

In order to quantify participant gaze patterns, we focus our analysis on how many times each participant "shifts in" to robot actions and motions, normalized by the prevalence of the action or motion. Specifically, we define an *attraction score* as the ratio of the activity's share of valid cross-robot shift-in events to its proportional duration in the session. Scores above 1.0 indicate the behavior attracted gaze shifts more than would be expected from its occurrence frequency alone, while scores below 1.0 indicate under-attraction relative to its prevalence. Our subsequent analyses are cross-robot only, as cross-view transitions involved minimal changes in robot activity and showed high redundancy, whereas cross-robot shifts captured the meaningful allocation of attention across robots in response to their activities.

Our analysis of attraction scores across all participants for each motion and action category found that fine-grained manipulator motions and goal-directed actions scored consistently above baseline (score > 1.0), while gross locomotion and idle states attracted cross-robot shifts below baseline. This indicates that participants were most reliably drawn to the other robot during precise manipulation and task execution and that some robot behaviors may act as shared attentional anchors. Still, participant-level monitoring strategies introduced meaningful variation in cross-robot gaze allocation, highlighting the need for personalized calibration.

We then computed pairwise rank correlation between attraction scores using Kendall's τ, and overall agreement using Kendall's W across all participants and within scenario-based subgroups (P1–P4 for *living room*, P5–P8 for *kitchen*, and P9–P12 for *hallway*). Figure 9 depicts our results. For actions, overall agreement was modest but significant ($W = 0.24$, $\chi^2(21) = 61.1$, $p < .001$), indicating that *participants exhibited some shared patterns in which actions drew attention.* This agreement was substantially stronger within scenario groups ($W = 0.54$–0.70, all $p < .01$), with pairwise τ values showing *more consistent gaze patterns among participants viewing the same scenario and less consistent patterns across scenarios.* For motion, overall agreement was lower and only marginally significant ($W = 0.14$, $\chi^2(12) = 19.9$, $p = .07$), *suggesting greater heterogeneity in how participants attended to motion.* However, this variability was not uniform: participants P9–P12 showed relatively strong agreement ($W = 0.57$, $p = .007$), while agreement was moderate for P5–P8 ($W = 0.40$, $p = .08$) and minimal for P1–P4 ($W = 0.18$, $p = .74$). Pairwise τ values further reveal substantial participant-dependent variation.

Taken together, these exploratory results suggest that participants were most similar to one another in how they attended to actions within a shared scenario, while differences emerged more strongly across individuals and scenarios, particularly in how motion was prioritized, pointing to both scenario-dependent and participant-dependent variation in gaze shift patterns. However, because each participant viewed a different video pairing and participants adopted different verbalization strategies, these differences cannot be attributed solely to stable individual or scenario-based attentional tendencies. We have yet to separate out session-level variation under the present monitoring and verbalization protocol.

4.4.3 Effectiveness of Causal Attribution. Participants generally perceived *Attune* as helping them make their gaze-shift reasoning explicit. On a 7-point Likert scale (1 = strongly disagree, 7 = strongly agree) responding to the question of whether their personal summary is "reflective and helpful," participants' average response was $M = 5.92$ ($SD = 0.67$), indicating agreement that the system captured participants' attention-allocation strategies.

We also examined *Attune*'s ability to distinguish proactive and reactive gaze. Cross-checking proactive/reactive labels against participants' corresponding free-response why-leave and why-arrive rationales from their annotations, there is evidence that both forms of gaze exist in the study task and that participants understood their meaning. Proactive-labeled shifts often invoked deliberate

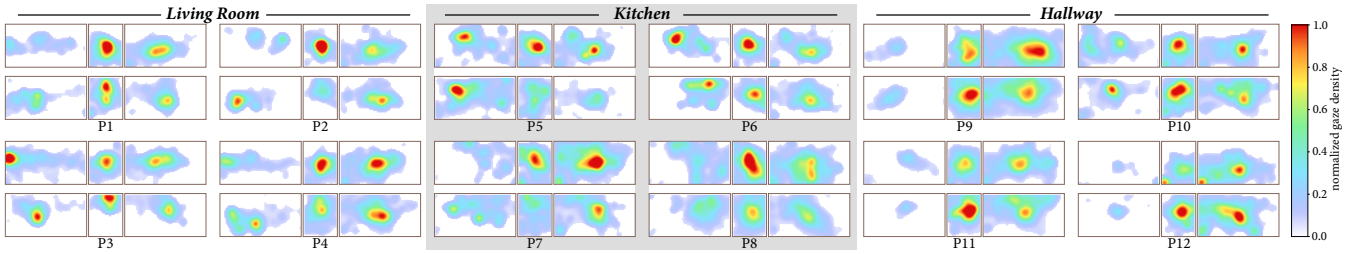


Figure 8: Participant-level gaze heatmaps across the three scenarios: Living Room, Kitchen, and Hallway. Each panel shows one participant's normalized gaze density on the same six-view robot video interface. Warmer colors indicate higher gaze density.

monitoring goals (e.g., *"I just wanted to check how Robot 2 was doing,"* P11), indicating that participants did not describe all gaze shifts in terms of visual salience alone. Reactive-labeled shifts often invoked perceptually salient events (e.g., *"I saw a large bottle-like object at the top of the screen, so I needed to verify,"* P2). Still, self-labels and free-response rationales are inherently retrospective and cannot truly establish whether the underlying attentional mechanism is top-down or bottom-up. We therefore interpret them as participant-endorsed rationales rather than causal classifications.

Using Braun and Clarke's reflexive thematic analysis approach [7], we further analyzed the interviews for *Attune*'s performance as a retrospective causal attribution tool. Four themes emerged.

Theme 1: *Annotation as a Scaffold for Externalizing Monitoring Reasoning.* Participants consistently reported that the annotation process surfaced attentional patterns they had not previously recognized, generating new self-insight rather than simply recording pre-existing knowledge. Several described encountering gaze-shift tendencies for the first time—recognizing an unconscious habit of disengaging when scenes became predictable, or realizing their switching behavior followed patterns never consciously registered (P4, P10). Participants highlighted the value of the *why-leave/why-arrive* structure for reconstructing shifts from both perspectives, and noted that classifying shifts as reactive or proactive further deepened reflection. Reviewing the AI-generated summary helped participants reconnect with thoughts from monitoring (P1, P9).

Theme 2: *Retrospective Reconstruction.* While participants could generally produce plausible explanations for gaze shifts, several questioned whether these faithfully reflected in-the-moment reasoning, noting that knowing the clip outcome made it difficult to disentangle genuine recall from post-hoc justification (P2, P7). P8 captured this ambiguity: *"I'm not sure if it really helped me remember, or if it makes me fake remember—I cannot tell the difference."* The replay function and gaze-shift timeline partially addressed this by restoring attentional context at the point of annotation (P5, P9), while confidence and memory clarity ratings helped participants avoid overclaiming certainty.

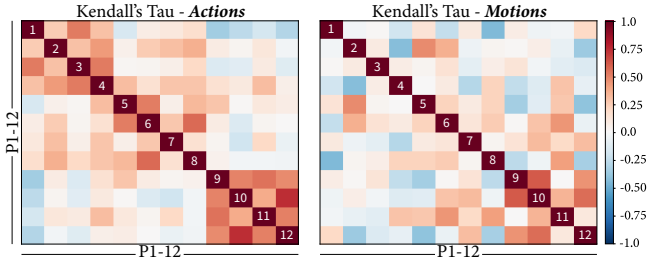


Figure 9: Cross-robot action and motion correlation matrices.

Theme 3: *Robot Intentionality as Gaze Anchor.* To identify the primary drivers of gaze shifts, participants ranked four contextual cues by importance. Using Borda count aggregation ($N = 12$) [20], *Robot Context* ranked highest (6/12 first-place rankings; score = 28), followed by *Visual Context* (4/12; score = 24), *Task Context* (2/12; score = 20), and *Others* (0/12; score = 0), suggesting that robot behavior was the dominant attentional cue, with visual salience having secondary importance. Gaze allocation was anchored to participants' real-time interpretation of robot intent: when behavior was legible, operators directed sustained attention; when intent became ambiguous, gaze lost its anchor and became passive. P11 noted feeling bored when a robot became stuck, but re-engaging the moment it appeared to discover a new object. P1 ranked *Task Context* lowest because *"most of the time I couldn't understand what the robot was trying to do."* These findings suggest that operator attention profiles are not stable personal traits, but are co-constructed with the legibility of robot behavior.

Theme 4: *Accuracy Threshold for AI Annotation as Cognitive Scaffold.* Participants responded positively to AI-generated summaries when the summaries compressed behavior into an interpretable profile. P3 described the result as feeling *"like my profile,"* and P5 noted it *"aligns perfectly with the summary the annotation generated for me."* However, participants noted that current outputs were more descriptive than interpretive (P9) and sometimes conflated object- and scene-level attention (P10). A clear accuracy threshold emerged: P9 estimated roughly *"80 or 90% accuracy"* would be needed before preferring AI annotation over manual effort, and also noted a cognitive tradeoff—*"although I take less time with AI annotation…I feel more cognitive load."* Participants favored a validation-oriented workflow (P6) over full automation, suggesting AI support should be framed around efficient review rather than replacement.

4.4.4 Usability and Usefulness. *Attune* received a mean SUS score of 75.21 ($SD = 14.48$), corresponding approximately to Bangor et al. [3]'s *Good* anchor. Several *Attune* annotation components were widely valued: the replay function and gaze-shift timeline helped restore attentional context at the point of annotation (P1, P5, P9),

and the why-leave/why-arrive structure was preferred over single-question formats for its ability to reconstruct a shift from both perspectives (P6, P8). The primary usability complaint was information density—P10 proposed progressive disclosure via step-by-step pop-ups, noting that presenting all options at once confused new users. Participants rarely struggled to describe gaze-shift reasoning in natural language; difficulty emerged when mapping explanations onto predefined categories. P9 found fine-grained labels—such as safety type, role, or task—difficult to distinguish, while P5 noted that the full option set appeared regardless of clip context and argued that the system should dynamically narrow viable choices.

5 Discussion

Attune is a pre-deployment elicitation tool that helps operators retrospectively reconstruct and characterize their gaze shifts during multi-robot monitoring. It is conceived as an initial step toward a broader calibration workflow in which designers may ultimately use operator-confirmed attention profiles to inform robot behavior and interface design before deployment. Our evaluation helps reveal the viability of this vision and surfaces three design implications for calibration systems. We discuss these implications below.

5.1 Implication: Annotation Elicits Rationales

A central finding is that retrospective, calibration-oriented gaze-shift attribution does not merely *record* operator behavior; it can also help *elicit* operator knowledge. Participants often recognized attentional tendencies during annotation that they had not explicitly articulated before, such as disengaging when a scene became predictable or switching views in response to cues they had not consciously registered (P4, P10). This suggests that calibration may elicit tacit knowledge that is practice-based, context-dependent, and not readily accessible through direct introspection alone [50].

Attune's annotation workflow appears to support this process in several ways. The *why-leave/why-arrive* structure helped participants interpret gaze shifts from both the source and destination perspectives, while the reactive/proactive classification encouraged reflection on broader top-down and bottom-up patterns rather than isolated events. Together, these findings suggest that calibration-oriented annotation supports event-level explanation rather than merely recording where operators look.

At the same time, retrospective annotation introduces a methodological tension: participants may reconstruct plausible explanations rather than retrieve original in-the-moment reasoning, especially when clip outcomes are already known (P8). This post-hoc approach was nevertheless deliberate, as requesting explanations during monitoring would interrupt the monitoring task and could alter the gaze behavior being observed. One alternative that avoids directly interrupting the monitoring task is passive inference over substantially more data than a short calibration session can provide. Retrospective annotation therefore trades a risk of recall decay for gaze behavior that remains uncontaminated by the act of measurement. Our findings suggest that replay, nonlinear timeline magnification, and confidence or memory-clarity ratings may help mitigate this risk, but do not eliminate it. Video length also remained a consistent source of recall degradation. More broadly, these results suggest that calibration should be treated not as passive data collection, but as a structured elicitation process that helps operators externalize otherwise implicit monitoring strategies while accounting for the limits of retrospective recall.

Design Recommendation: Calibration systems should explicitly support operator memory during retrospective annotation, for example, through visual signposting, rapid replay around shift boundaries, and explicit uncertainty marking for later review.

5.2 Implication: Robot Legibility and Attention

Our findings suggest that operator attention profiles should not be interpreted as purely internal preferences, but are co-shaped through interaction with robot behavior and how legibly it communicates state and intent. What operators choose to monitor may partially depend on how clearly the robot communicates its state, intention, and ongoing activity to the operator.

This matters because a gaze shift can reflect different causes. In some cases, it indicates purposeful monitoring of meaningful robot activity. In others, it reflects an effort to resolve uncertainty caused by ambiguous motion, weak signaling, or limited visual evidence. Seen this way, calibration does not simply capture where an operator prefers to look; it can also reveal where robot behavior or multi-view presentation fails to support efficient supervision.

More broadly, these findings suggest that calibration should be treated as a relational process rather than a one-sided profiling method. An attention profile may reflect the fit between operator expectations and robot legibility. Repeated shifts toward moments of ambiguity, verification, or missed cues can indicate not only attentional tendencies, but also opportunities to redesign robot behavior for more interpretable supervision.

Design Recommendation: Calibration systems should account for robot behavioral legibility when interpreting attention profiles, so that attention patterns are understood as a function of both operator strategy and the clarity of robot intent.

5.3 Implication: AI Assistance Scaffolds and Shapes Reflection

Our findings suggest that AI assistance can make calibration more scalable by reducing annotation effort and helping operators recognize broader attention patterns that may be difficult to extract from raw annotations alone. Participants often found AI-generated summaries useful for making their monitoring tendencies more legible, suggesting that the pipeline can transform fragmented annotations into meaningful inputs for future design work. At the same time, the value of AI assistance appears contingent on suggestion quality, where participants expressed that AI usefulness depended on whether a suggestion matched their own recollections. Suggestions that matched participants' recollections appeared to reduce effort and support reflection, whereas mismatches increased perceived verification burden and could undermine the validity of the calibration itself. This risk was especially pronounced when participants no longer remembered the original gaze event clearly enough to confidently confirm or reject a suggestion. In such cases, AI assistance risks introducing plausible but unverifiable interpretations that could shape, rather than merely support, the operator's explanation of their own attentional reasoning. This concern motivates bolstering human-in-the-loop calibration, in which inadequately

calibrated models may produce plausible but spurious rationales. In our case, rather than inferring subjective intent from scratch, Phase 4 propagates rationales anchored in the operator's manual annotations and reviewed patterns. This calibration incurs upfront effort; future mixed-initiative approaches could reduce that cost by prioritizing ambiguous or informative shifts for manual review.

Overall, our findings suggest that AI-assisted calibration should be designed as a human-in-the-loop, verification-oriented workflow, thus scaffolding reflection, rather than a purely automated workflow. Automated gaze analysis answers where and when the eye moved, but the rationales participants surfaced were tacit and not present in any signal. Annotation and pattern analysis provides this signal, suggesting why an operator's gaze shifts. As such, we view calibration as necessary for telling a designer what to change. We argue that the initial cost of acquiring this signal is outweighed by its benefit, as it grounds later automation in the operator's own confirmed rationales rather than in inferences from external cues. After manual annotation, *Attune* positions the operator as the final interpretive authority, thus preserving a verification-oriented workflow. To ease future verification burdens, low-confidence AI suggestions with low information gain can be suppressed.

Finally, participants' privacy concerns suggest that this reflective support must also be bounded in its long-term use. Repeated calibration across deployments may enable increasingly fine-grained behavioral profiles of individual operators, raising concerns about secondary use beyond the original calibration context. Calibration systems should therefore make data retention limits, access boundaries, deletion after the calibration session, limits on sharing raw gaze video, and intended uses explicit, particularly given the risk of repurposing calibration data for workplace surveillance, so that the personalization that makes calibration valuable does not become a liability for the people it is meant to support.

Design Recommendation: AI-assisted calibration systems must prioritize and support human-in-the-loop verification over pure automation by surfacing high-confidence and high-information-gain suggestions, preserving operator authority over interpretation, and supporting operator privacy.

5.4 Limitations and Future Work

This work has several limitations. While our two-robot, six-view, prerecorded setup represents the minimal multi-robot configuration needed to elicit cross-robot gaze shifts and enables a foundational feasibility test, it does not establish transfer to larger fleets, live deployments, different robot platforms, or alternative display layouts, all of which may change workload and gaze behavior. Future work should examine different fleet sizes, platforms, and display layouts.

Much additional analysis remains to be done, including evaluating the objective quality of summaries produced by *Attune*. Crucially, we did not leverage our data beyond analysis, and as a result, our contributions stop short of demonstrating how attention profiles can be operationalized in practice. Important next steps are to (1) evaluate these profiles with interface designers, (2) inform the development of multi-robot video sensemaking tools tailored to individual operator attentional tendencies, and (3) adapt robot behaviors directly in response to these profiles. Future work should also integrate *Attune* with alert-driven workflows and investigate how our proposed calibration step can sustain attention at the alert source while avoiding distraction elsewhere.

Additional limitations remain. First, *Attune* supports retrospective rather than live annotation. While retroactive annotation facilitates reflection, it also risks memory decay and inaccurate rationalization. Future work should explore live or near-real-time annotation despite its potential cognitive cost. Second, we did not systematically evaluate LLM accuracy. Our findings therefore focus on participants' use of AI assistance rather than model performance itself. Future work should examine how LLM accuracy affects calibration validity, verification effort, and trust. Third, the study did not include task-based or visual-saliency baselines, so it cannot objectively distinguish task-driven from salience-driven shifts; future work should compare operator-endorsed rationales with such baselines. Fourth, *Attune* is a purpose-built research prototype rather than an extension of established behavioral annotation tooling. Building a custom interface was necessary due to the lack of existing tools that couple gaze-shift detection, robot replay, and annotation in a single workflow (§2.3). It nevertheless raises an adoption cost: practitioners cannot access these capabilities from the tools they already use, and our usability findings are specific to our interface rather than to gaze-rationale annotation in general. Future work should integrate our approach into existing tools. Fifth, *Attune* does not assume a specific monitoring duration after calibration—its elicited profiles are intended to apply both to sustained monitoring and shorter windows. Though shorter, alert-driven windows were not evaluated here, alerts and gaze calibration are a promising research direction. Alerts indicate where attention is needed, whereas *Attune* characterizes how attention is sustained or lost once a feed is inspected. Integrating alert logs as a secondary timeline is therefore a natural extension. For longer monitoring durations, the manual cost of calibration could be reduced through further mixed-initiative approaches. Lastly, our study design has other limitations. Verbalization increases cognitive burden and divergent verbalization strategies may have altered natural attention allocation. Future work should explore non-verbalized designs. Additionally, we recruited naïve participants, whereas domain experts are needed to assess applicability to operational multi-robot contexts. Larger per-scenario samples and additional measures beyond our current study are also needed to further strengthen our analyses and fully validate *Attune*'s profiling capability.

6 Conclusion

We present *Attune*, a pre-deployment elicitation tool that helps operators construct attention profiles for the future calibration of multi-robot monitoring interfaces and robot behavior. *Attune* records operator eye gaze when watching robots perform different tasks and assists the operator in attributing possible causes to their gaze shifts. *Attune* then uses individual shifts to extrapolate the operator's gaze patterns. We conducted a user study to evaluate *Attune* and uncovered several design implications.

Acknowledgments

This work was supported by George Mason University and in part by 4-VA, a higher education collaborative partnership for advancing the Commonwealth of Virginia.

References

[1] Johanna Ahlskog, Maria-Theresa Bahodi, Artur Lugmayr, and Timothy Merritt. 2024. Fostering trust through user interface design in multi-drone search and rescue. In *Proceedings of the Second International Symposium on Trustworthy Autonomous Systems*. 1–11. doi:10.1145/3686038.3686052

[2] ATLAS.ti. 2023. *ATLAS.ti Scientific Software Development GmbH*. https://atlasti.com

[3] Aaron Bangor, Philip T Kortum, and James T Miller. 2008. An empirical evaluation of the system usability scale. *Intl. Journal of Human–Computer Interaction* 24, 6 (2008), 574–594. doi:10.1080/10447310802205776

[4] Michael Barz, Omair Shahzad Bhatti, Hasan Md Tusfiqur Alam, Duy Minh Ho Nguyen, Kristin Altmeyer, Sarah Malone, and Daniel Sonntag. 2025. eyeNotate: Interactive Annotation of Mobile Eye Tracking Data Based on Few-Shot Image Classification. *Journal of Eye Movement Research* 18, 4 (2025), 27. doi:10.3390/jemr18040027

[5] Steve Benford, Pepita Barnard, Sarah Sharples, Helena Webb, Clara Mancini, Ayse Kucukyilmaz, Simon Castle Green, Eike Schneiders, Victor Ngo, Alan Chamberlain, et al. 2025. Charting the Ecosystem of Trust in Cat Royale Or What It Takes to Trust a Robot to Play with Cats. In *International Conference on Social Robotics*. Springer, 623–636. doi:10.1007/978-981-95-2398-6_42

[6] Steve Benford, Rachael Garrett, Christine Li, Paul Tennent, Claudia Núnez-Pacheco, Ayse Kucukyilmaz, Vasiliki Tsaknaki, Kristina Höök, Praminda Caleb-Solly, Joe Marshall, et al. 2025. Tangles: Unpacking extended collision experiences with soma trajectories. *ACM Transactions on Computer-Human Interaction* 32, 4 (2025), 1–34. doi:10.1145/3723875

[7] Virginia Braun, Victoria Clarke, Nikki Hayfield, and Gareth Terry. 2019. Thematic analysis 48. *Handbook of research methods in health social sciences* (2019), 843–860. doi:10.1007/978-981-10-5251-4_103

[8] John Brooke. 1996. SUS-A 'quick and dirty' usability scale. *Usability Evaluation in Industry* 189, 194 (1996), 4–7. doi:10.1201/9781498710411-35

[9] Shih-Yi Chien, Huadong Wang, Michael Lewis, Siddharth Mehrotra, and Katia Sycara. 2011. Effects of alarms on control of robot teams. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 55. Sage Publications Sage CA: Los Angeles, CA, 434–438. doi:10.1177/1071181311551089

[10] Erin K Chiou, Mustafa Demir, Verica Buchanan, Christopher C Corral, Mica R Endsley, Glenn J Lematta, Nancy J Cooke, and Nathan J McNeese. 2022. Towards human–robot teaming: Tradeoffs of explanation-based communication strategies in a virtual search and rescue task. *International Journal of Social Robotics* 14, 5 (2022), 1117–1136. doi:10.1007/s12369-021-00834-1

[11] Manolis Chiou, Nick Hawes, and Rustam Stolkin. 2021. Mixed-initiative variable autonomy for remotely operated mobile robots. *ACM Transactions on Human-Robot Interaction (THRI)* 10, 4 (2021), 1–34. doi:10.1145/3472206

[12] Jacob W Crandall, Mary L Cummings, Mauro Della Penna, and Paul MA De Jong. 2010. Computing the effects of operator attention allocation in human control of multiple robots. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 41, 3 (2010), 385–397. doi:10.1109/tsmca.2010.2084082

[13] Oliver Deane, Eszter Toth, and Sang-Hoon Yeo. 2023. Deep-SAGA: a deep-learning-based system for automatic gaze annotation from eye-tracking data. *Behavior Research Methods* 55, 3 (2023), 1372–1391. doi:10.3758/s13428-022-01833-4

[14] Kerry Dhakal. 2022. NVivo. *Journal of the Medical Library Association: JMLA* 110, 2 (2022), 270. doi:10.5195/jmla.2022.1271

[15] Carla Diana. 2024. Designing robots that work and matter. In *Designing Interactions with Robots*. Chapman and Hall/CRC, 140–147. doi:10.1201/9781003371021-7

[16] Jill L Drury, Jean Scholtz, and David Kieras. 2007. Adapting GOMS to model human-robot interaction. In *Proceedings of the ACM/IEEE international conference on Human-robot interaction*. 41–48. doi:10.1145/1228716.1228723

[17] Jill L Drury, Jean Scholtz, and David Kieras. 2007. Modeling Human-Robot Interaction with GOMS. (2007).

[18] Kamak Ebadi, Lukas Bernreiter, Harel Biggie, Gavin Catt, Yun Chang, Arghya Chatterjee, Christopher E Denniston, Simon-Pierre Deschênes, Kyle Harlow, Shehryar Khattak, et al. 2023. Present and future of SLAM in extreme environments: The DARPA SubT challenge. *IEEE Transactions on Robotics* 40 (2023), 936–959. doi:10.1109/tro.2023.3323938

[19] Saad Elbeleidy, Terran Mott, Dan Liu, Ellen Do, Elizabeth Reddy, and Tom Williams. 2023. Beyond the session: Centering teleoperators in socially assistive robot-child interactions reveals the bigger picture. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–33. doi:10.1145/3610175

[20] Peter Emerson. 2013. The original Borda count and partial voting. *Social Choice and Welfare* 40, 2 (2013), 353–358. doi:10.1007/s00355-011-0603-9

[21] Cyrus K Foroughi, Shannon Devlin, Richard Pak, Noelle L Brown, Ciara Sibley, and Joseph T Coyne. 2023. Near-perfect automation: Investigating performance, trust, and visual attention allocation. *Human factors* 65, 4 (2023), 546–561. doi:10.1177/00187208211032889

[22] Maria Fox and Derek Long. 2003. PDDL2. 1: An extension to PDDL for expressing temporal planning domains. *Journal of artificial intelligence research* 20 (2003), 61–124. doi:10.1613/jair.1129

[23] Olivier Friard and Marco Gamba. 2016. BORIS: a free, versatile open-source event-logging software for video/audio coding and live observations. *Methods in ecology and evolution* 7, 11 (2016), 1325–1330. doi:10.1111/2041-210x.12584

[24] Mafalda Gamboa, Sofia Thunberg, Patricia Alves-Oliveira, and Meagan B Loerakker. 2025. We Are the Robots: Tapping Into the Lived Experiences of Wizards of Oz. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–8. doi:10.1145/3706599.3720149

[25] Fei Gao, Mary L. Cummings, and Erin Treacy Solovey. 2014. Modeling Teamwork in Supervisory Control of Multiple Robots. *IEEE Transactions on Human-Machine Systems* 44, 4 (2014), 441–453. doi:10.1109/THMS.2014.2312391

[26] Cindy M Grimm and Kristen Thomasen. 2021. On the practicalities of robots in public spaces. *WeRobot 2021* (2021).

[27] Daniel Höller, Gregor Behnke, Pascal Bercher, Susanne Biundo, Humbert Fiorino, Damien Pellier, and Ron Alford. 2020. HDDL: An extension to PDDL for expressing hierarchical planning problems. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 9883–9891. doi:10.1609/aaai.v34i06.6542

[28] Khadija Iddrisu, Waseem Shariff, Maciej Stec, Noel O'Connor, and Suzanne Little. 2026. Eye Movement Classification Using Neuromorphic Vision Sensors. *Journal of Eye Movement Research* 19, 1 (2026), 17. doi:10.3390/jemr19010017

[29] Omer Mutlu Turk Kaya and Gokhan Erdemir. 2023. Design of an eight-wheeled mobile delivery robot and its climbing simulations. In *SoutheastCon 2023*. IEEE, 895–900. doi:10.1109/southeastcon51012.2023.10115114

[30] Charles C. Kemp, Aaron Edsinger, Henry M. Clever, and Blaine Matulevich. 2022. The Design of Stretch: A Compact, Lightweight Mobile Manipulator for Indoor Human Environments. In *2022 International Conference on Robotics and Automation (ICRA)*. 3150–3157. doi:10.1109/ICRA46639.2022.9811922

[31] Hee Rin Lee, Sarah Fox, EunJeong Cheon, and Samantha Shorey. 2025. Minding the Stop-Gap: Attending to the "Temporary," Unplanned, and Added Labor of Human-Robot Collaboration in Context. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 34–44. doi:10.1109/hri61500.2025.10973955

[32] Séverin Lemaignan, Fernando Garcia, Alexis Jacq, and Pierre Dillenbourg. 2016. From real-time attention assessment to "with-me-ness" in human-robot interaction. In *2016 11th ACM/IEEE international conference on human-robot interaction (HRI)*. Ieee, 157–164. doi:10.1109/hri.2016.7451747

[33] Michael Lewis, Shi-Yi Chien, Siddarth Mehortra, Nilanjan Chakraborty, and Katia Sycara. 2014. Task switching and single vs. multiple alarms for supervisory control of multiple robots. In *International conference on engineering psychology and cognitive ergonomics*. Springer International Publishing, 499–510. doi:10.1007/978-3-319-07515-0_50

[34] R Barry Lewis and Steven M Maas. 2007. QDA Miner 2.0: Mixed-model qualitative data analysis software. *Field methods* 19, 1 (2007), 87–108. doi:10.1177/1525822x06296589

[35] Barry R Manor and Evian Gordon. 2003. Defining the temporal threshold for ocular fixation in free-viewing visuocognitive tasks. *Journal of neuroscience methods* 128, 1-2 (2003), 85–93. doi:10.1016/s0165-0270(03)00151-1

[36] Max Planck Institute for Psycholinguistics. 2026. ELAN (Version 7.1) [Computer software]. https://archive.mpi.nl/tla/elan.

[37] Paul McIlvenny. 2024. Guide for DOTEbase users: An online help guide. (2024).

[38] Paul McIlvenny, Jacob Gorm Davidsen, and Alexander Stein. 2024. DOTEbase: software tools for qualitative analysis. (2024).

[39] Paul McIlvenny, Jacob Gorm Davidsen, Alexander Stein, and Artúr Barnabás Kovács. 2022. DOTE: distributed open transcription environment. (2022).

[40] Hyu Miyashita and Yasuto Nakanishi. 2025. FrameLight: Guiding Attention Through Motion and Arrangement in Shape-Changing Furniture Robots. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 1498–1502. doi:10.1109/hri61500.2025.10974138

[41] Diederick C. Niehorster, R. S. Hessels, Marcus Nyström, J. S. Benjamins, and I. T. C. Hooge. 2025. gazeMapper: A tool for automated world-based analysis of gaze data from one or multiple wearable eye trackers. *Behavior Research Methods* (2025). doi:10.3758/s13428-025-02704-4

[42] Ataol Burak Özsu, Tuğçe Nur Pekçetin, Defne Şiir Faydali, and Burcu A Urgen. 2025. Distraction by a Human or a Robot: Effects of Perceptual Load and Action Type. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 1171–1175. doi:10.1109/hri61500.2025.10974021

[43] Sergio Passero, Hannah Rm Pelikan, Mathias Broth, and Barry Brown. 2024. Honkable Gestalts: Why autonomous vehicles get honked at. In *Proceedings of the 16th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. 317–328. doi:10.1145/3640792.3675732

[44] Yunqiang Pei, Renming Huang, Mingfeng Zha, Guoqing Wang, Peng Wang, Qiao Kang, Yang Yang, and Heng Tao Shen. 2025. AttentionAR: AR Adaptation and Warning for Real-World Safety via Attention Modeling and MLLM Reasoning. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*. 1–19. doi:10.1145/3746059.3747674

[45] Hannah Pelikan, Katie Winkle, and David Porfirio. 2024. Designing human-robot interactions: a StEER tutorial. In *Adjunct Proceedings of the 2024 Nordic Conference on Human-Computer Interaction*. 1–4. doi:10.1145/3677045.3685467

[46] Hannah RM Pelikan, Fanjun Bu, and Wendy Ju. 2025. The people behind the robots: How wizards wrangle robots in public deployments. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–21. doi:10.1145/3706598.3713237

[47] Hannah RM Pelikan, Bilge Mutlu, and Stuart Reeves. 2025. Making sense of public space for robot design. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 152–162. doi:10.1109/hri61500.2025.10973847

[48] Hannah RM Pelikan, Stuart Reeves, and Marina N Cantarutti. 2024. Encountering autonomous robots on public streets. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. 561–571. doi:10.1145/3610977.3634936

[49] Jeffrey R Peters, Vaibhav Srivastava, Grant S Taylor, Amit Surana, Miguel P Eckstein, and Francesco Bullo. 2015. Human supervisory control of robotic teams: Integrating cognitive modeling with engineering design. *IEEE Control Systems Magazine* 35, 6 (2015), 57–80. doi:10.1109/mcs.2015.2471056

[50] Michael Polanyi. 2009. The tacit dimension. In *Knowledge in organisations*. Routledge, 135–146.

[51] Daniel J Rea, Stela H Seo, Neil Bruce, and James E Young. 2017. Movers, shakers, and those who stand still: visual attention-grabbing techniques in robot teleoperation. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. 398–407. doi:10.1145/2909824.3020246

[52] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 779–788. doi:10.1109/cvpr.2016.91

[53] Giacomo Rizzolatti, Lucia Riggio, Isabella Dascola, and Carlo Umiltá. 1987. Reorienting attention across the horizontal and vertical meridians: evidence in favor of a premotor theory of attention. *Neuropsychologia* 25, 1 (1987), 31–40. doi:10.1016/0028-3932(87)90041-8

[54] Sayanti Roy, Trey Smith, Brian Coltin, and Tom Williams. 2023. I need your help... or do i? maintaining situation awareness through performative autonomy. In *Proceedings of the 2023 ACM/IEEE international conference on human-robot interaction*. 122–131. doi:10.1145/3568162.3576954

[55] Elie Saad, Mark A Neerincx, and Koen V Hindriks. 2019. Welcoming robot behaviors for drawing attention. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 636–637. doi:10.1109/hri.2019.8673325

[56] Dario D. Salvucci and Joseph H. Goldberg. 2000. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications* (Palm Beach Gardens, Florida, USA) *(ETRA '00)*. Association for Computing Machinery, New York, NY, USA, 71–78. doi:10.1145/355017.355028

[57] Yi Ting Sam, Erin Hedlund-Botti, Manisha Natarajan, Jamison Heard, and Matthew Gombolay. 2024. The impact of stress and workload on human performance in robot teleoperation tasks. *IEEE Transactions on Robotics* 40 (2024), 4725–4744. doi:10.1109/tro.2024.3484630

[58] Fabian Schirmer, Philipp Kranz, Chad G Rose, Volker Willert, Jan Schmitt, and Tobias Kaupp. 2025. Utilizing eye gaze for human-robot collaborative assembly. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 1603–1607. doi:10.1109/hri61500.2025.10974041

[59] Han Sloetjes and Peter Wittenburg. 2008. Annotation by category-ELAN and ISO DCR. In *6th international Conference on Language Resources and Evaluation (LREC 2008)*. doi:10.63317/454o8uo6f7z8

[60] Vidya Somashekarappa, Christine Howes, and Asad Sayeed. 2020. An annotation approach for social and referential gaze in dialogue. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. 759–765.

[61] Vaibhav Srivastava, Ruggero Carli, Cédric Langbort, and Francesco Bullo. 2014. Attention allocation for decision making queues. *Automatica* 50, 2 (2014), 378–388. doi:10.1016/j.automatica.2013.11.028

[62] Miguel Valdez and Matthew Cook. 2023. Humans, robots and artificial intelligences reconfiguring urban life in a crisis. *Frontiers in Sustainable Cities* 5 (2023), 1081821. doi:10.3389/frsc.2023.1081821

[63] Lisa-Marie Vortmann, Jannes Knychalla, Sonja Annerer-Walcher, Mathias Benedek, and Felix Putze. 2021. Imaging time series of eye tracking data to classify attentional states. *Frontiers in Neuroscience* 15 (2021), 664490. doi:10.3389/fnins.2021.664490

[64] Michael E Walker, Maitrey Gramopadhye, Bryce Ikeda, Jack Burns, and Daniel Szafir. 2024. The cyber-physical control room: A mixed reality interface for mobile robot teleoperation and human-robot teaming. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. 762–771. doi:10.1145/3610977.3634981

[65] Zicheng Wang and Beno Benhabib. 2025. Concurrent Multi-Robot Search of Multiple Missing Persons in Urban Environments. *Robotics* 14, 11 (2025), 157. doi:10.3390/robotics14110157

[66] Christopher Wickens. 2021. Attention: Theory, principles, models and applications. *International Journal of Human–Computer Interaction* 37, 5 (2021), 403–417. doi:10.1080/10447318.2021.1874741

[67] Christopher D Wickens. 2015. *Noticing events in the visual workplace: The SEEV and NSEEV models*. Cambridge University Press. doi:10.1017/cbo9780511973017.046

[68] Christopher D Wickens, Juliana Goh, John Helleberg, William J Horrey, and Donald A Talleur. 2017. Attentional models of multitask pilot performance using advanced display technology. In *Human Error in Aviation*. Routledge, 155–175. doi:10.4324/9781315092898-10

[69] Jeremy M Wolfe and W Gray. 2007. Guided search 4.0. *Integrated models of cognitive systems* (2007), 99–119.

[70] Nan Wu, Kaiyan Liu, Ruizhi Cheng, Bo Han, and Puqi Zhou. 2024. Theia: Gaze-driven and perception-aware volumetric content delivery for mixed reality headsets. In *Proceedings of the 22nd Annual International Conference on Mobile Systems, Applications and Services*. 70–84. doi:10.1145/3643832.3661858

[71] Fumitaka Yamaoka, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. 2009. Developing a model of robot behavior to identify and appropriately respond to implicit attention-shifting. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*. 133–140. doi:10.1145/1514095.1514120

[72] Yan Zhang, Haoqi Li, Ramtin Tabatabaei, and Wafa Johal. 2025. ROSAnnotator: A Web Application for ROSBag Data Analysis in Human-Robot Interaction. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 1099–1103. doi:10.1109/hri61500.2025.10974254

[73] Yu Zhang and Songdong Xue. 2025. A New Trend In the Warehousing: A Review of Human Robot Collaboration. *IEEE Access* (2025). doi:10.1109/access.2025.3606967

[74] Puqi Zhou, Ali Asgarov, Aafiya Hussain, Wonjoon Park, Amit Paudyal, Sameep Shrestha, Chia-Wei Tang, Michael Lighthiser, Michael Hieb, Xuesu Xiao, et al. 2026. Designing Multi-Robot Ground Video Sensemaking with Public Safety Professionals. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. 1–22. doi:10.1145/3772318.3790679