AI-assisted Voice Control in Human–Robot Collaboration during Scaffold Assembly Tasks: A Controlled Evaluation Using Questionnaires and Motion-Tracking Indicators

Tanghan Jiang[1], Erol Cemiloglu[1], Yihai Fang[1]

[1]Monash University, Australia
tanghan.jiang1@monash.edu, ecem0001@student.monash.edu, yihai.fang@monash.edu

Abstract

AI-assisted voice control is increasingly used in hands-busy human–robot collaboration, yet its effects on trust and task-grounded behaviours remain rarely explored. This study evaluated AI-assisted voice control in a scaffold-assembly handover task, comparing it against a No-voice Baseline while keeping robot paths and task layout constant. Workload (NASA-TLX), safety, performance, and trust were assessed using questionnaires, alongside motion-tracking indicators such as reaction time and head-orientation-based attention. Results showed that while AI-assisted voice control significantly improved perceived task performance and effectiveness, it also increased mental and temporal workload. Objective metrics indicated faster reaction times during handovers and a non-significant tendency toward reduced robot-directed monitoring. These findings suggest that voice-based AI agents can enhance operational efficiency and human-robot synchrony, but may impose higher cognitive demands that should be balanced in HRC system design.

Keywords –

Human-robot collaboration; AI-assisted voice control; Scaffold assembly; Motion tracking

1 Introduction

Construction work continues to face challenges related to labour availability and worker exposure to physical hazards during physically intensive activities [1]. Therefore, human–robot collaboration (HRC) has been explored to assist with such tasks, particularly when robots operate in proximity to humans within a shared workspace [2]. In shared-space collaboration, outcomes depend not only on robot motion performance but also on how humans anticipate robot actions, maintain safety margins, and allocate attention between the robot and the workpiece [3]. These demands are salient in handover and assembly workflows, where timing uncertainty can increase conservative waiting and readiness behaviour [4].

Recent advances in large language models have enabled AI-enabled voice interfaces for robotic systems [5]. In hands-busy settings, voice interaction provides a hands-free channel to pacing control (e.g., pause, resume, speed adjustment) and can provide status feedback during task execution [6]. Such capabilities may increase perceived controllability and predictability, shaping task-grounded behaviours and perceptions relevant to trust assessment and perceived safety [7]. However, voice interaction can also introduce cognitive demands due to attentional switching, command formulation, and potential misrecognition, potentially increasing perceived workload even when coordination improves [8].

As voice interfaces become feasible for shared-space collaboration, it is necessary to quantify their influence on trust-related perceptions and task-grounded behaviours during controlled robot execution [9]. The incremental effect of adding a voice-based pacing and status-feedback channel remains insufficiently evidenced when robot task plans and nominal paths are held constant, limiting causal interpretation of voice-control-driven effects, especially on human perceptions and behaviours [6]. To address this gap, this study uses a controlled laboratory scaffold-assembly task in which participants collaborate with a mobile manipulator to install four horizontal PVC pipes under two conditions: an AI-voice-controlled (AI-assisted) condition and a no-AI (Baseline) condition. The task layout and planned robot motion were controlled across conditions, and the manipulation was the availability of the voice interaction channel. Trust-related outcomes were operationalised using a three-layer measurement framework comprising (i) self-report measures (workload, perceived safety, perceived task performance/effectiveness, trust, and check-all-that-apply (CATA) attributions), (ii) a motion-tracking attention indicator (head-orientation proxy relative to the end-effector), and (iii) a motion-tracking

temporal-anticipation indicator (stance-based readiness duration from stance onset to robot release across repeated handovers). The study examines whether voice-enabled interaction alters these subjective evaluations and behavioural indicators in a shared-space scaffold assembly task.

2 Literature Review

Human–robot collaboration (HRC) in construction is motivated by the need to improve productivity under labour constraints while reducing workers' exposure to safety and ergonomic risks in physically intensive operations [5]. Collaborative robotic systems are increasingly explored for assembly and material-handling tasks where human dexterity and situational judgement complement robotic repeatability and load-bearing capability [10]. In shared-workspace collaboration, task outcomes depend not only on robot motion performance but also on coordination processes that shape how humans anticipate robot actions, maintain safety margins, and allocate attention between the robot and the task during repeated handovers and sequential assembly [2], [11].

Within such close-proximity settings, trust is a critical human factor because it influences reliance–supervision decisions that directly affect both efficiency and safety management [12]. Trust in HRC has been conceptualised as a multidimensional construct encompassing perceived safety, predictability, and alignment with the human's mental model, and is commonly assessed using complementary self-report and task-based indicators [11], [13]. In handover-centric workflows, trust-relevant adaptation is often reflected in strategy choices (e.g., monitoring intensity and conservatism as the robot approaches), motivating measurement approaches that combine self-reported judgements with behavioural indicators observed during interaction [14].

Given trust's role in coordination, interaction modalities and information support emerge as primary levers for shaping expectations and managing timing uncertainty in shared workspaces. Transparency cues that communicate robot state, intent, and timing reduce uncertainty and improve coordination, while perceived controllability (e.g., pausing, resuming, or pace adjustment) can mitigate timing mismatches during critical handovers and support trust calibration [14], [15]. In hands-busy shared-space tasks, voice interaction offers a practical channel to deliver such transparency and controllability because hands are often occupied and visual attention is required for manipulation, alignment, and quality verification. Recent AI-based voice interfaces can provide contextual status feedback and support constrained natural-language commands for pacing regulation, but may entail trade-offs such as cognitive overhead and disruption from misrecognition or ambiguity during time-sensitive actions [8], [16].

To interpret modality effects in shared-space collaboration, trust assessment benefits from pairing adoption-relevant self-reports with task-grounded behavioural indicators sensitive to supervision, readiness, and attention across repeated handovers [17]. Self-reports could capture indicators such as perceived safety, workload, task effectiveness, and trust-relevant perceptions, helping to distinguish coordination benefits from interface burden, while behavioural indicators characterise interaction strategy by monitoring allocation and temporal readiness around critical events [11], [13]. Monitoring can be approximated using head- or gaze-based orientation proxies when eye tracking is impractical, provided the proxy is defined relative to task-relevant robot targets during handovers [18]. Complementarily, temporal readiness indicators quantify the timing of preparatory activity around predictable robot actions, capturing a conservatism-efficiency trade-off in sustained preparatory behaviour [19]. Nevertheless, controlled evidence isolating voice-based pacing and status feedback under fixed robot task plans and nominal motion paths using complementary subjective and behavioural indicators aligned with handover structure remains limited [13]. Additionally, modality evaluations are often confounded with changes in autonomy, speed policy, or motion planning that directly affect perceived safety and coordination difficulty in shared-space settings [11]. Accordingly, the present study pairs adoption-relevant self-reports with task-grounded indicators of monitoring and readiness during repeated handovers, thereby providing controlled evidence of how a voice-based pacing channel reshapes coordination strategy under fixed-robot execution.

3 Methods

A controlled human–robot scaffold-assembly task was used to isolate the effects of AI-assisted voice control (restricted pacing commands and status feedback) on human perceptions and behaviours while holding robot motion plans and task layout constant. Section 3.1 outlines the robotic platform, sensing architecture, and task procedure, Section 3.2 describes the conditions and participant administration, and Section 3.3 specifies subjective and objective measures.

3.1 Experimental Setup and Task Procedure

The experiment was conducted in an indoor robotics laboratory using a simplified scaffold-assembly scenario. A MiR100 mobile base transported PVC pipes between a loading rack and a fixed scaffold mock-up, and a UR-10e arm mounted on the base executed pre-programmed

trajectories from the rack to a handover pose and into the target clamps at the assembly zone. The base followed a fixed route to a docking pose under conservative collaborative speed limits, with motion executed under a ROS2-based control architecture. In the AI-assisted condition, a microphone–speaker pair enabled an Autogen-based voice pipeline that delivered brief status messages and accepted four restricted pacing commands (i.e., pause, resume, slower, faster) mapped to discrete speed adjustments and temporary halts within predefined limits, with waypoint geometry and trajectory shape unchanged. Safety was ensured through conservative speed settings, software safety zones, and accessible emergency-stop buttons, all monitored by experimenters.

Objective motion data were collected using OptiTrack (for the robot) and Xsens Awinda (for the human). OptiTrack comprised eight cameras surrounding the scaffold mock-up, with markers on the UR-10e links and gripper to reconstruct end-effector kinematics at 120 Hz. Participants wore an upper-body Xsens suit with sensors on the head, torso, and upper limbs, providing segment kinematics at 60 Hz for deriving head orientation and upper-body movement features. At the start of each recording, a brief wand-waving event produced salient peaks that were used to align timestamps and register the two data streams within a common task frame. The experimental layout, key task locations, and sensing configuration are summarized below (Figure 1).

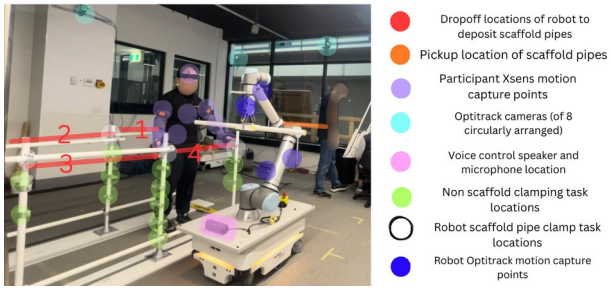

Figure 1. Experimental layout and sensing configuration

The collaborative task comprised four repeated robot–human handover–assembly cycles per trial, each involving pipe delivery, handover/alignment, participant installation, and robot return for the next pipe (Figure 2). Participants tightened clamps with a spanner after each handover; during robot return, they attached non-load-bearing clamps to maintain comparable activity across the trial. Participants received a standardised briefing and video instruction and completed an unrecorded familiarisation round at reduced speed (including voice commands where applicable). Critically, robot routes and nominal trajectories were held constant across conditions, with the voice channel manipulating only pacing/status interaction. Across the Baseline (No-AI) and AI-assisted conditions, robot route, docking poses, and nominal trajectories were held constant, and only the availability of voice-based pacing/status interaction differed.

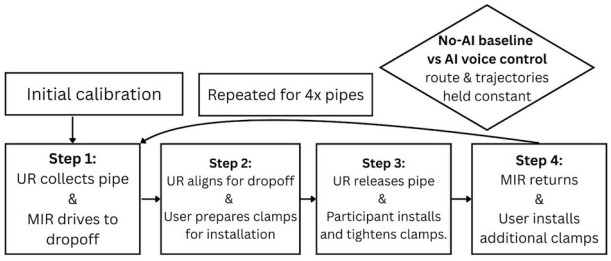

Figure 2. Trial workflow and controlled conditions

3.2 Control Conditions and Participants

Two conditions were designed to isolate the interface effect while holding the physical task and robot execution constant. In the Baseline (No-AI) condition, the MiR100 (robot base) and UR-10e (robot arm) executed the predefined route, docking pose, and alignment trajectories without voice input or verbal status feedback. In the AI-assisted condition, the same motion plans and layout were used, but an Autogen-based assistant provided status announcements and accepted the four pacing commands that adjusted speed or inserted temporary halts without altering trajectory geometry; thus, condition differences are attributable to the voice interface.

Participants were recruited from the university community (n = 21), and kinematic indicators were analysed using indicator-specific valid pairs after signal-quality exclusions. Eligibility required adults with normal or corrected-to-normal vision and no self-reported musculoskeletal conditions affecting upper-body movement; age, gender, and prior construction/robotics exposure were recorded. Each participant completed one trial per condition in a counterbalanced order with brief rests between conditions. The experiment was approved by the Monash University Ethics Committee (ID: 47827), and written consent was obtained prior to the experiment starting.

3.3 Measurements

Objective kinematic indicators were derived from synchronised Xsens (human) and OptiTrack (robot) recordings to characterise attention towards the robot and temporal coordination during handover. The data streams were temporally aligned using the wand-waving event at the start of recording and resampled to a common sampling rate of 60 Hz. Four robot–human handover episodes per trial were segmented using robot motion logs and verified against motion-capture signals. Following the behavioural trust-indicator framework of

Campagna et al. [14], analogous attention-based and temporal measures were adopted for the Xsens–OptiTrack scaffold-assembly task.

Attention was defined as the proportion of trial time for which head orientation lay within a distance-adaptive cone centred on the robot end-effector (Figure 3). This head-orientation proxy captures overt robot-directed monitoring toward the end-effector during handover, providing a task-grounded attention indicator without requiring eye tracking. To preserve comparability under varying head–end-effector distance, the cone was derived from a fixed spatial region of radius r = 1 m around the end-effector, such that the angular threshold widens at closer distances and narrows when farther away, reducing distance-driven classification artefacts. For each frame t, the angle $\theta(t)$ between the head orientation vector and the head-to-end-effector direction was computed, and the cone half-angle was set as

$$\varphi_{cone}(t) = arc\sin\left(min\left(1, \frac{r}{d(t)}\right)\right) \quad (1)$$

with r=1 m and $d(t)$ the head–end-effector distance. A binary attending label was assigned as A(t)=1 when $\theta(t) \le \varphi_{cone}(t)$ (else 0), and the trial-level score was

$$Attention = \frac{1}{T}\sum_{t=1}^{T} A(t) \quad (2)$$

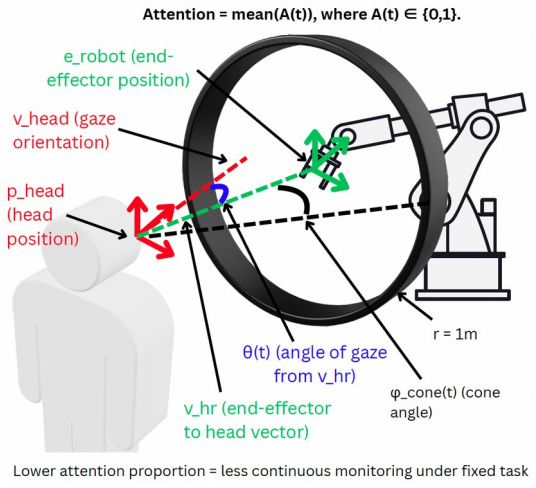

Figure 3. Attention operationalisation (distance-adaptive cone)

Stance-based reaction time captured sustained readiness duration during each handover. For handover i ($i=1,\ldots,4$), RT_i was defined as the interval from stance onset $t^{(i)}_{stance,on}$ to robot release $t^{(i)}_{release}$:

$$RT_i = t^{(i)}_{release} - t^{(i)}_{stance,on} \quad (3)$$

and the trial-level indicator was averaged across the four handovers. Here, shorter stance durations were interpreted as greater temporal confidence in robot behaviour, whereas longer durations reflected more conservative sustained readiness during approach.

Alongside motion-capture data, questionnaires were administered to characterise participants' background and condition-specific perceptions of workload, safety, trust, and task performance/effectiveness. Participants first completed a demographic form that included age, gender, prior construction experience, and robot exposure. After each condition, workload was assessed using the NASA-TLX (mental, physical, temporal demand, performance, effort, and frustration) and aggregated into an overall workload score. Safety perception and perceived task performance/effectiveness were measured using two study-specific five-item Likert scales, and trust in safe operation was captured with a single-item rating. These items were measured on a 1 (strongly disagree) to 5 (strongly agree) Likert scale, with item means forming composite indices (higher values indicating more positive perceptions). Participants also completed a check-all-that-apply (CATA) attribution checklist on perceived influences on performance/effectiveness and trust. Together, the framework comprised three layers for Baseline (No-AI) vs. AI-Assisted comparisons: (i) self-reports (TLX, safety, performance/effectiveness, trust, and CATA endorsements), (ii) task-grounded monitoring allocation (attention), and (iii) task-grounded temporal readiness (stance-based reaction time).

4 Results

4.1 Data Validity and Pre-Processing Procedure

Questionnaire and kinematic datasets were screened to ensure valid paired observations under Baseline and AI-assisted conditions. The questionnaire items were checked for completeness and coding, scale scores were computed, and internal consistency (Cronbach's α) was evaluated for multi-item scales. CATA responses were coded as option-level binary endorsements per condition to support paired/repeated-measures modelling. For kinematic indicators, synchronised Xsens and OptiTrack recordings were aligned using the wand-waving event, resampled to 60 Hz, and segmented into four robot–human handover episodes per trial, with the segments verified against motion-capture signals and robot motion logs. Participants were included on an indicator-by-indicator basis when valid paired observations were

available (stance-based reaction time: n = 19; attention: n = 11), where the reduced attention sample reflects indicator-specific signal-quality exclusions rather than condition-dependent missingness.

4.2 Subjective Results

The AI-assisted voice control yielded a clear subjective trade-off: perceived task performance/effectiveness increased with perceived workload, whereas perceived safety and the global trust-in-safe-operation rating remained stable (Table 1). Workload (overall NASA-TLX score) increased substantially under AI-assisted conditions relative to Baseline, whereas Performance/Effectiveness showed a moderate improvement. Safety and Trust did not differ reliably, supporting the interpretation that the intervention altered perceived efficiency and demand without shifting perceived safety or confidence in safe operation.

Table 1 Primary subjective outcomes (paired Baseline vs. AI-assisted conditions, n = 21)

Construct	Baseline M(SD)	AI M(SD)	t, p, d_z
Workload (TLX overall)	43.35 (16.85)	61.37 (23.11)	4.22, 0.000419, 0.92
Performance/ Effectiveness	3.51 (0.85)	4.10 (0.74)	2.90, 0.00902, 0.63
Safety perception	3.90 (0.89)	4.07 (0.90)	0.82, 0.420, 0.18
Trust (safe operation)	4.10 (0.77)	4.14 (0.79)	0.24, 0.815, 0.05

Notes: Two-tailed paired t-tests were used for all constructs (df = n − 1). Effect size is the paired sample standardised mean difference (d_z). Multi-item scales showed high internal consistency: Performance/Effectiveness ($\alpha_{Baseline}$=0.89; $\alpha_{AI-assisted}$ =0.85) and Safety perception ($\alpha_{Baseline}$=0.92; $\alpha_{AI-assisted}$=0.84).

Workload changes were not uniform across TLX dimensions; the subscale breakdown indicates that Temporal Demand was the most sensitive component to the intervention, consistent with the voice channel's pacing function. Temporal Demand (weighted) increased from Baseline to AI-assisted (Baseline: M=5.17, SD=4.47; AI-assisted: M=14.70, SD=10.53; t=4.25, p=0.000272, d_z=0.96) and remained significant after Holm correction across TLX subscales (p_{Holm}=0.00163). By contrast, Effort showed only nominal evidence (p=0.0416) and did not survive correction (p_{Holm} =0.208), while the remaining TLX components did not reach corrected significance. Therefore, a workload elevation concentrated on time-pressure or pace regulation rather than a broad increase across all demand dimensions was observed.

Perceived task performance and effectiveness improved under AI-assisted conditions, and item-level tests (Holm-adjusted within the five items) indicated that this improvement was primarily driven by perceived support for tempo. The strongest driver was "The robot's speed helped with my performance" (Baseline M=3.00; AI-assisted M=4.19; t=4.86, p=0.000096, d_z=1.06; p_{Holm}=0.000478), whereas "The workflow of the task felt productive" was weaker and marginal after correction (p_{Holm}=0.067) and the remaining items did not show reliable change. This pattern suggests that perceived gains were explicitly attributed to pacing/speed support rather than an undifferentiated improvement across all facets of task execution.

Attributional results further clarify the perceived mechanism by indicating that participants more frequently attributed both performance and trust outcomes to the interaction strategy under AI-assisted voice control (Figure 4). After false-discovery-rate control within each checklist, endorsement patterns were broadly comparable across most factors, but "The way I interacted (observing, voice control, etc.) with the robot" emerged as the only robust differential factor in both domains. A clustered GEE model indicated higher odds of endorsing this factor under AI-assisted conditions for both Trust attributions (OR=6.25, 95% CI [1.84, 21.22], p_{FDR}=0.0198) and Performance attributions (OR=6.25, 95% CI [1.84, 21.22], p_{FDR}=0.0264). Collectively, these findings support the interpretation that AI voice control reshaped perceived coordination strategy and pacing-related experience, producing efficiency gains with a measurable increase in workload cost while leaving perceived safety and overall trust in safe operation unchanged.

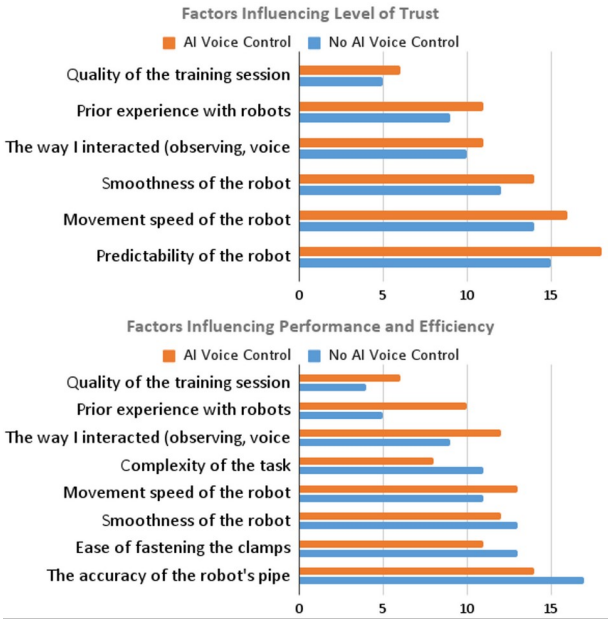

Figure 4. Attributional shifts in perceived determinants of performance and trust across conditions

4.3 Objective Results

Objective kinematic indicators provided task-grounded evidence of coordination-strategy adaptation under AI-assisted voice control, complementing the subjective pattern of increased workload alongside improved perceived performance/effectiveness. Two complementary indicators were analysed: stance-based reaction time (sustained readiness duration during handover), which captures temporal coordination around the release event, and attention (proportion of trial time classified as attending), which captures overt monitoring allocation under workflow constraints. The paired comparisons show a large improvement in temporal coordination and a trend toward reduced continuous robot-directed monitoring (Table 2).

Table 2 Objective kinematic indicators (paired Baseline vs. AI-assisted conditions)

Indicator	Baseline M(SD)	AI M(SD)	$\Delta(B-$	t, p, d_z
Stance-based reaction time (s)	25.15 (5.60)	15.23 (3.73)	9.92	8.43, 1.15×10^{-7}, 1.93
Attention (proportion)	0.2318 (0.1183)	0.1794 (0.1149)	0.0524	1.86, 0.0927, 0.56

Notes: Paired t-tests (two-tailed). Sample sizes: stance-based reaction time n=19; Attention n=11 (indicator-specific valid pairs). Directionality: RT decreased in 19/19 participants; attention decreased in 9/11 participants. 95% CI for Δ: reaction time [7.45, 12.39] s; attention [−0.0104, 0.1152]. Δ(B−AI) denotes the paired mean difference (Baseline minus AI-assisted); d_z is the paired-sample standardised mean difference.

Stance-based reaction time showed a substantial, directionally consistent reduction under AI-assisted voice control (Figure 5). All participants with valid paired data showed shorter sustained readiness duration under AI-assisted conditions (19/19), indicating a systematic contraction of the preparatory window before release and, therefore, tighter temporal coordination around handover events. Although the magnitude of improvement varied across individuals, the consistent direction supports a robust behavioural shift rather than isolated participant-specific changes.

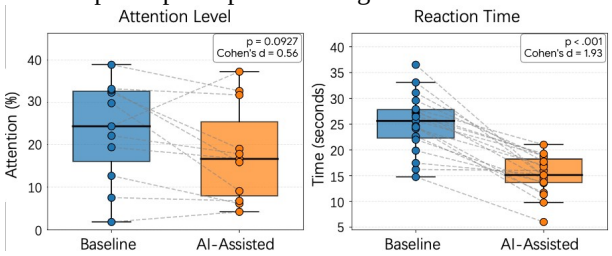

Figure 5. Objective behavioural indicators of coordination strategy

Attention showed a moderate decrease under AI-assisted, with predominantly negative within-participant changes (9/11), but the statistical evidence did not reach conventional significance. This indicator should be interpreted with the smaller valid-pair sample (n=11) due to indicator-specific signal-quality exclusions. Given the task structure, Attention should be interpreted as a task-constrained monitoring strategy: phases involving reception, alignment, and safe proximity impose a functional lower bound on robot-facing orientation. Accordingly, lower attention under AI-assisted is interpreted as reduced continuous monitoring demand within those constraints, while recognising the smaller valid-pair sample for this indicator.

Method transparency for attention classification is provided by representative traces that visualise how the attending decision is derived under dynamic head–robot proximity (Figure 6). The distance profile shows repeated approach–withdrawal cycles aligned with the handover workflow, indicating that head–robot geometry varies systematically across phases rather than remaining fixed. The lower panel overlays the head-to-robot angle $\theta(t)$ with the distance-adaptive cone threshold $\phi_{cone}(t)$, from which the binary attending series $A(t)$ is obtained ($A(t)$=1 when $\theta(t) \le \phi_{cone}(t)$). Because $\phi_{cone}(t)$ expands at closer distances and narrows when the robot is

farther away, the criterion preserves a consistent behavioural meaning (i.e., orientation within a fixed 1 m region around the end-effector) across changing proximity. This supports the interpretability of the trial-level Attention comparison by reducing the risk that condition differences reflect proximity-driven classification artefacts rather than monitoring strategy differences, which is particularly relevant given the smaller valid-pair sample for attention.

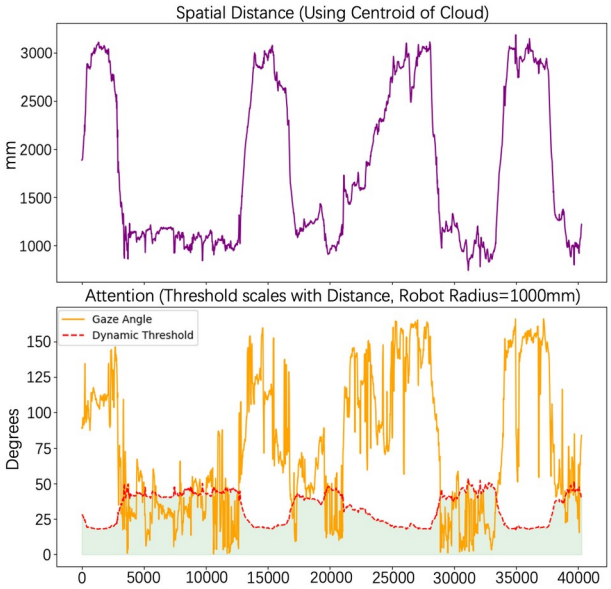

Figure 6. Method transparency examples for Attention classification under dynamic proximity

5 Discussion

AI-assisted voice control yielded convergent evidence of an efficiency–demand trade-off in close-proximity scaffold assembly. Subjectively, participants reported higher Task Performance and Effectiveness alongside increased workload, with the most robust workload change concentrated in NASA-TLX Temporal Demand (Table 1). Objectively, stance-based reaction time was consistently shorter under AI-assisted than under the Baseline condition, indicating tighter temporal coordination around the release event, whereas Attention showed a modest downward trend, consistent with reduced continuous robot-directed monitoring within task constraints (Table 2; Figure 4). In contrast, safety perception and the single-item trust-in-safe-operation rating remained stable, thereby limiting the interpretation to coordination strategy and perceived effectiveness rather than to a shift in global safety or trust appraisals (Table 1).

The stance-based reduction in reaction time provides a plausible behavioural pathway linking interface support to perceived effectiveness. Because this indicator captures sustained readiness rather than stimulus–response latency, shorter durations under AI-assisted reflect a contraction of the preparatory window before release and therefore more efficient timing alignment during handover. This behavioural profile aligns with item-level evidence indicating that participants attributed performance gains primarily to speed/tempo support rather than to broad improvements across all facets of execution. Notably, the directionally uniform reduction (19/19) indicates a systematic shift in coordination strategy rather than an effect driven by a small subset of responders.

Workload increases were not diffuse across demand dimensions but were concentrated in time-related pressure, suggesting that the coordination benefit was achieved through tighter temporal coupling rather than by making the task uniformly easier. The increase in Temporal Demand aligns with objective evidence of compressed readiness timing: maintaining shorter preparatory windows can impose stronger timing requirements and increase perceived time pressure, even when overall performance appears improved. This pattern highlights a key design implication for voice-enabled pacing: interface support that accelerates coordination may simultaneously increase temporal demand, motivating future work to test whether alternative cue timing, predictability, or confirmation strategies can preserve coordination gains while reducing perceived time pressure.

The stability of safety perception and the global trust rating indicate that efficiency gains did not translate into detectable changes in broad safety or trust evaluations in this task and sample. Nevertheless, attribution results suggest that participants perceived the intervention's effect primarily through interaction strategy rather than through changes in robot capability or safety margins: "the way I interacted with the robot" was the only endorsement that differed robustly across conditions for both performance and trust attributions (Figure 3). The Attention indicator provides tentative behavioural support for reduced monitoring under AI-assisted, but should be interpreted cautiously given the smaller valid-pair sample (n=11) and the task-imposed lower bounds on robot-facing orientation during reception and alignment (Table 2). The method-transparency example further supports interpretability by demonstrating that the attending criterion adapts to dynamic proximity, reducing the likelihood that condition differences reflect distance-driven classification artefacts (Figure 5). Overall, the results indicate that AI-assisted voice control primarily reshaped temporal coordination and interaction strategy, improving perceived task effectiveness with a measurable time-related workload cost while leaving global safety and trust-in-safe-operation ratings almost unchanged.

6 Conclusions

This study investigated how AI-assisted voice control influences coordination and trust-related outcomes in a controlled, construction-motivated scaffold-assembly handover task while holding robot task plans, nominal motion paths, and task layout constant. Results indicate an efficiency–demand trade-off. Under AI-assisted voice control, participants reported higher Task Performance and Effectiveness, alongside higher perceived workload. The most significant increase was in weighted NASA-TLX Temporal Demand, whereas safety perception and trust in safe operation remained relatively stable. Motion-tracking evidence converged with these perceptions, showing a large, directionally uniform reduction in stance-based reaction time (improved temporal coordination around release) and a non-significant tendency toward lower attention. CATA attributions further locate differences in conditions primarily in interaction strategy/coordination mode rather than in broad changes in safety or global trust. Overall, voice-enabled pacing and status feedback can measurably reshape coordination strategy and perceived task effectiveness in shared-space scaffold assembly tasks, but may increase time-related demand, motivating interface designs that preserve coordination gains while mitigating temporal-pressure costs.

References

[1] Y. Fu, W. Lu, and J. Chen, "A virtual reality-based ergonomic assessment approach for human-robot collaboration workstation design in modular construction manufacturing," *Advanced Engineering Informatics*, vol. 64, p. 103054, 2025.

[2] V. Villani, F. Pini, F. Leali, and C. Secchi, "Survey on human–robot collaboration in industrial settings: Safety, intuitive interfaces and applications," *Mechatronics*, vol. 55, pp. 248–266, Nov. 2018, doi: 10.1016/j.mechatronics.2018.02.009.

[3] A. Nourmohammadi, M. Fathi, and A. H. Ng, "Balancing and scheduling assembly lines with human-robot collaboration tasks," *Computers & Operations Research*, vol. 140, p. 105674, 2022.

[4] K. E. Stecke and M. Mokhtarzadeh, "Balancing collaborative human–robot assembly lines to optimise cycle time and ergonomic risk," *International Journal of Production Research*, vol. 60, no. 1, pp. 25–47, Jan. 2022, doi: 10.1080/00207543.2021.1989077.

[5] J. Liu, H. Luo, and D. Wu, "Human–robot collaboration in construction: robot design, perception and interaction, and task allocation and execution," *Advanced Engineering Informatics*, vol. 65, p. 103109, 2025.

[6] E. Mendez *et al.*, "Integration of deep learning and collaborative robot for assembly tasks," *Applied Sciences*, vol. 14, no. 2, p. 839, 2024.

[7] B. Sadrfaridpour and Y. Wang, "Collaborative assembly in hybrid manufacturing cells: An integrated framework for human–robot interaction," *IEEE Transactions on Automation Science and Engineering*, vol. 15, no. 3, pp. 1178–1192, 2017.

[8] M. Romeo, I. Torre, S. Le Maguer, A. Sleat, A. Cangelosi, and I. Leite, "The Effect of Voice and Repair Strategy on Trust Formation and Repair in Human-Robot Interaction," *J. Hum.-Robot Interact.*, vol. 14, no. 2, pp. 1–22, June 2025, doi: 10.1145/3711938.

[9] W.-C. Chang and S. Hasanzadeh, "Toward a Framework for Trust Building between Humans and Robots in the Construction Industry: A Systematic Review of Current Research and Future Directions," *J. Comput. Civ. Eng.*, vol. 38, no. 3, p. 03124001, May 2024, doi: 10.1061/JCCEE5.CPENG-5656.

[10] A. Keshvarparast, D. Battini, O. Battaia, and A. Pirayesh, "Collaborative robots in manufacturing and assembly systems: literature review and future research agenda," *J Intell Manuf*, vol. 35, no. 5, pp. 2065–2118, June 2024, doi: 10.1007/s10845-023-02137-w.

[11] S. Hopko, J. Wang, and R. Mehta, "Human factors considerations and metrics in shared space human-robot collaboration: A systematic review," *Frontiers in Robotics and AI*, vol. 9, p. 799522, 2022.

[12] M. Diab and Y. Demiris, "TICK: A Knowledge Processing Infrastructure for Cognitive Trust in Human–Robot Interaction," *Int J of Soc Robotics*, Jan. 2025, doi: 10.1007/s12369-024-01206-1.

[13] G. Campagna and M. Rehm, "A Systematic Review of Trust Assessments in Human–Robot Interaction," *J. Hum.-Robot Interact.*, vol. 14, no. 2, pp. 1–35, June 2025, doi: 10.1145/3706123.

[14] G. Campagna, M. Lagomarsino, M. Lorenzini, D. Chrysostomou, M. Rehm, and A. Ajoudani, "Estimating trust in human-robot collaboration through behavioral indicators and explainability," *IEEE Robotics and Automation Letters*, 2025, Accessed: Dec. 18, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/11129656/

[15] A. Ajoudani, A. M. Zanchettin, S. Ivaldi, A. Albu-Schäffer, K. Kosuge, and O. Khatib, "Progress and prospects of the human–robot collaboration,"

Autonomous robots, vol. 42, no. 5, pp. 957–975, 2018.

[16] M. Norda, C. Engel, J. Rennies, J.-E. Appell, S. C. Lange, and A. Hahn, “Evaluating the efficiency of voice control as human machine interface in production,” *IEEE Transactions on Automation Science and Engineering*, vol. 21, no. 3, pp. 4817–4828, 2023.

[17] H. Chauhan, A. Pakbaz, Y. Jang, and I. Jeong, “Analyzing trust dynamics in human–robot collaboration through psychophysiological responses in an immersive virtual construction environment,” *Journal of Computing in Civil Engineering*, vol. 38, no. 4, p. 04024017, 2024.

[18] A. Meynard, G. Seneviratna, E. Doyle, J. Becker, H.-T. Wu, and J. S. Borg, “Predicting trust using automated assessment of multivariate interactional synchrony,” in *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, IEEE, 2021, pp. 1–8. Accessed: Dec. 17, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9667082/

[19] K. Hald and M. Rehm, “Determining movement measures for trust assessment in human-robot collaboration using imu-based motion tracking,” in *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, IEEE, 2023, pp. 1267–1272. Accessed: Dec. 17, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10309497/