Proceedings of the International Conference on Industrial Engineering and Operations Management

Publisher: IEOM Society International, USA DOI: 10.46254/NA11.20260163
Published: June 9, 2026

Deployment-Efficient Multimodal Human–Robot Interaction Using Voice Commands and Color Verification in Structured Workspaces

Lokesh Rao Gujja
Department of Engineering Technology
Purdue University Northwest. USA

Jayanth Chillarige
Department of Electrical and Computer Engineering
Purdue University Northwest

Sesha Raghavendra Srinivas Subnivis
Department of Engineering Technology
Purdue University Northwest

Sunil Kumar Kalagarla
Department of Electrical and Computer Engineering
Purdue University Northwest

Maged Mikhail
Department Chair
Department of Engineering Technology
Purdue University Northwest

Abstract

The rapid adoption of collaborative robots in industrial and laboratory environments has increased the demand for intuitive and deployment-efficient human–robot interaction systems. While recent research emphasizes perception-intensive autonomy and dynamic scene reconstruction, such approaches often introduce calibration complexity and extended commissioning effort in structured workspaces. This paper presents a deployment-oriented multimodal human–robot interaction framework designed specifically for organized collaborative environments. The proposed system integrates constrained-vocabulary voice command interpretation with region-of-interest (ROI)-based HSV color verification and predefined spatial mapping, implemented on a Kinova Gen3 collaborative robot using event-driven control. Instead of relying on full pose estimation or continuous scene interpretation, the architecture intentionally limits visual processing to color verification at predefined observation poses, thereby reducing computational overhead and integration effort while maintaining real-time responsiveness. Experimental evaluation under controlled indoor conditions demonstrated 94% speech recognition accuracy with an average latency of 310 ms. The perception module achieved 95% color classification consistency with an average processing time of 68 ms per cycle. Across 50 pick-and-place trials, the system achieved a 92% successful grasp rate with an average task completion time of 8.4 seconds. Comparative testing against joystick-based manual control showed reduced execution variability and improved repeatability for repetitive manipulation tasks. The results demonstrate that aligning system complexity with environmental structure provides a practical and scalable strategy for collaborative robot deployment in laboratories, medical preparation areas, and organized assembly settings.

Keywords
Collaborative Robots; Multimodal Human–Robot Interaction; Voice-Controlled Robotics; Color-Based Vision; Structured Industrial Environments.

1. Introduction
Collaborative robots (cobots) are increasingly deployed in manufacturing, laboratory, and medical preparation environments where humans and robots operate in shared or adjacent workspaces. Unlike traditional industrial robots that function in isolated and highly controlled cells, cobots are designed to support flexible task execution with direct human interaction under safety constraints (Villani et al. 2018). The rapid expansion of collaborative robotics in logistics, laboratory automation, and flexible manufacturing has been driven by increased demand for adaptable yet safe human–robot collaboration (Bogue 2018). As deployment expands into organized industrial environments, the emphasis shifts from pure automation capability toward intuitive control, rapid commissioning, and reduced integration complexity. Beyond industrial productivity, collaborative robots are increasingly viewed as interactive systems operating within human-centered environments, where usability and intuitive control are essential (Dautenhahn 2007).

Human–robot interaction (HRI) research has traditionally explored multiple interaction paradigms, including disjoint, sequential, and shared workspace configurations (Yanco and Drury 2002). In parallel, multimodal interaction systems combining speech, vision, and other sensing modalities have been proposed to enhance robustness and accessibility (Oviatt 1999). Voice-based command systems allow operators to issue high-level instructions without traditional programming interfaces (Tellex et al. 2011), while vision-guided manipulation enables object identification and grasping (Kragic and Christensen 2005).

However, many recent multimodal systems emphasize perception-intensive autonomy, including full pose estimation, dynamic scene reconstruction, and continuous object tracking. Although such approaches improve adaptability in unstructured environments, they often introduce increased computational demand, calibration effort, and deployment time. In structured industrial settings—such as chemical laboratories, organized assembly stations, or medical preparation areas—object locations are typically predefined and color-coded classification is commonly used. Under these constraints, excessive perceptual complexity may not provide proportional operational benefit.

This research investigates an alternative design perspective: that selective modality integration aligned with environmental structure can improve deployment efficiency, repeatability, and operator accessibility without increasing algorithmic complexity. Rather than expanding perceptual sophistication, the proposed system intentionally constrains visual processing to color verification at predefined observation poses while using voice commands for high-level intent specification.

The proposed framework integrates constrained-vocabulary speech recognition, region-of-interest (ROI)-based HSV color classification with majority voting, predefined spatial mapping, and event-driven robot control implemented on a Kinova Gen3 collaborative robot. The approach aims to reduce commissioning effort, minimize calibration requirements, and maintain real-time manipulation performance in organized workspaces.

The central premise of this work is that aligning system complexity with environmental structure provides a practical and scalable strategy for collaborative robot deployment in structured industrial environments.

1.1 Objectives
The overall objective of this research is to design, implement, and experimentally validate a deployment-efficient multimodal human–robot interaction framework for structured collaborative environments.
The specific objectives are:

1. To develop a constrained-vocabulary voice command interface that maps high-level user intent to deterministic robot action primitives.
2. To implement a lightweight HSV-based color verification module using region-of-interest extraction and majority voting for object identification without full pose estimation.
3. To integrate predefined spatial mapping and event-driven robot control using the Kinova Kortex SDK to ensure repeatable and safe manipulation execution.
4. To quantitatively evaluate system performance across speech recognition, visual perception, and physical manipulation using statistical metrics including accuracy, latency, task completion time, and standard deviation.
5. To compare the proposed multimodal framework with conventional joystick-based manual control in terms of execution time, variability, and deployment complexity.

All objectives are addressed through experimental evaluation and statistical analysis presented in subsequent sections.

2. Literature Review

The integration of collaborative robots into shared human workspaces has generated significant research in human–robot interaction (HRI), perception-driven manipulation, and multimodal communication frameworks. As cobots transition from isolated industrial cells to collaborative industrial environments, interaction design and deployment efficiency become central concerns. This section reviews relevant literature across three primary areas: voice-based interaction, vision-guided manipulation, and multimodal integration strategies.

Voice interaction has been widely explored as a natural modality for enabling intuitive human–robot communication. Early studies demonstrated the feasibility of mapping predefined speech commands to robot action primitives for task execution in structured environments (Hegde and Patil 2017). Tellex et al. (2011) introduced grounded natural language approaches for robot control, enabling high-level instruction mapping to motion execution.

Comparative evaluations of speech recognition frameworks highlight variability in word error rates depending on acoustic conditions and vocabulary constraints (Al-Radhi and Hussein 2019). These studies emphasize the importance of constrained vocabularies and deterministic command mapping for safety-critical robotic applications.

More recent implementations have combined speech interfaces with manipulation tasks. Agarwal et al. (2023) proposed a voice-assisted pick-and-place system using a mobile manipulator in warehouse environments, integrating speech recognition with navigation and grasp planning. While these systems improve accessibility, they often rely on perception-heavy pipelines or dynamic mapping frameworks. Overall, voice-based robotic interaction improves usability and reduces reliance on traditional programming interfaces; however, many implementations prioritize generalization rather than deployment simplicity.

Vision-based perception plays a central role in robotic grasping and object manipulation. Kragic and Christensen (2005) provided a foundational review of vision-guided robotic manipulation, highlighting object detection, pose estimation, and visual servoing techniques. Subsequent research expanded into depth sensing, dynamic tracking, and learning-based grasp prediction. Lightweight object detection approaches have also been proposed to reduce computational demand in real-time robotic manipulation tasks, particularly in resource-constrained deployments (Zhang et al. 2020).

Kim et al. (2022) proposed HSV color-space-based object localization for robotic grasping without prior knowledge, demonstrating the robustness of HSV models under varying illumination conditions. Similarly, Santos et al. (2022) explored vision-based object manipulation for assistive robotics, integrating camera-based detection with grasp execution.

While full pose estimation and dynamic scene reconstruction enhance adaptability in unstructured environments, they increase calibration requirements and computational overhead. In structured industrial settings where object positions are predefined, simplified visual verification strategies may provide sufficient reliability without extensive perception pipelines.

Multimodal HRI combines multiple sensing and communication channels to improve robustness and operator accessibility. Oviatt (1999) established early foundations of multimodal system design, demonstrating that combining modalities enhances interaction flexibility. Wang et al. (2024) discussed multimodal fusion strategies for collaborative systems, emphasizing redundancy and contextual interpretation.

In industrial robotics, multimodal frameworks typically integrate speech, vision, and sometimes haptic feedback. Recent studies have further emphasized the importance of structured workspace design and deployment-oriented system integration to improve commissioning efficiency in collaborative robotic systems (Cheng et al. 2021). However, many such systems prioritize environmental adaptability and generalized autonomy. As deployment contexts vary, selecting appropriate modalities based on environmental structure becomes critical to balancing complexity and operational efficiency. In structured workspaces where tasks are repetitive and object placement is predefined, multimodal integration may benefit from constrained perception and deterministic action sequencing rather than dynamic scene interpretation.

Existing literature demonstrates substantial progress in voice-controlled robotics, vision-guided manipulation, and multimodal interaction frameworks. However, much of this research emphasizes increasing perceptual sophistication and environmental adaptability. Fewer studies explicitly investigate how constraining perceptual scope in structured industrial environments can improve deployment efficiency, reduce commissioning effort, and maintain repeatability.

This research addresses this gap by proposing a deployment-oriented multimodal framework that intentionally limits vision to color verification at predefined observation poses while leveraging high-level voice commands for intent specification. By aligning system complexity with environmental structure, the approach seeks to balance functionality with integration practicality.

3. Methodology

The proposed multimodal human–robot interaction framework is designed as a layered, deployment-oriented architecture for structured collaborative environments. The system integrates high-level command interpretation, lightweight visual verification, deterministic task coordination, and event-driven robot control to enable reliable manipulation while minimizing commissioning complexity. The overall system architecture is illustrated in Figure 1. The architecture consists of a command interface, command parser, perception module, task coordination layer, and robot control layer, supported by an implementation layer built using Python, OpenCV, NumPy, SpeechRecognition, and the Kinova Kortex SDK.

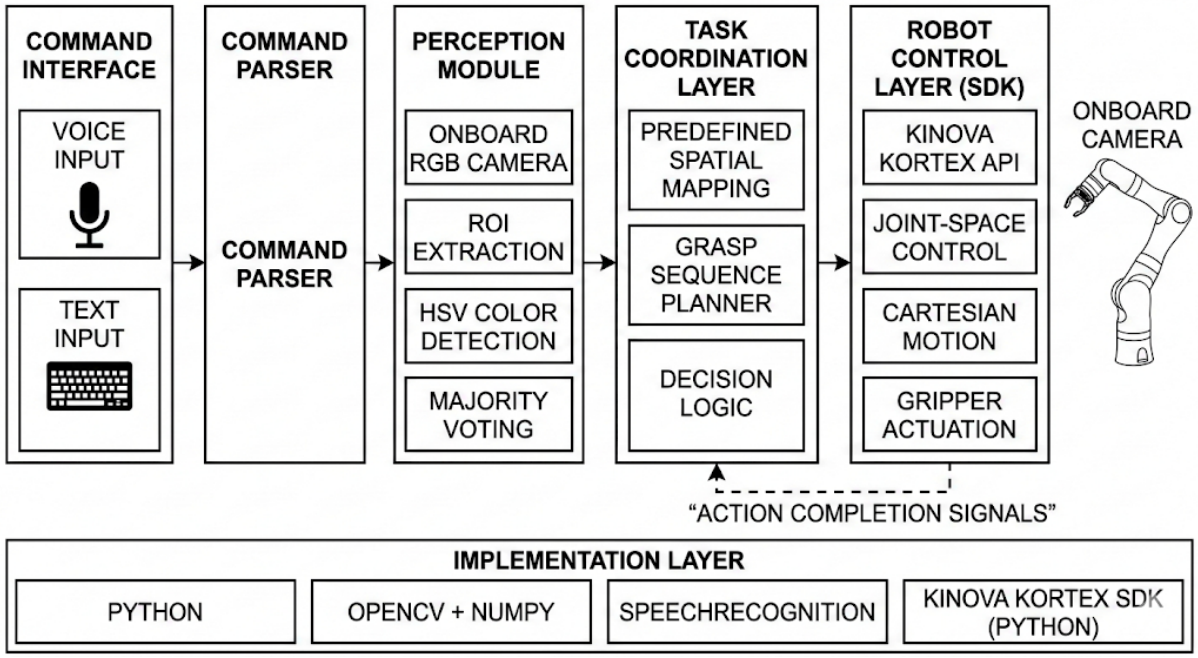

Figure 1. System architecture of the proposed deployment-oriented multimodal framework

3.1 Command Interface and Parsing

The command interface enables interaction through either voice or text input. Voice commands are processed using a constrained-vocabulary automatic speech recognition pipeline implemented in Python with the SpeechRecognition library, while text-to-speech feedback is generated using the pyttsx3 library to confirm successful command interpretation. To improve reliability and ensure deterministic execution, the system restricts acceptable commands to predefined task-relevant phrases corresponding to structured manipulation operations. Recognized speech is converted into a normalized internal command representation, which is then interpreted by the command parser. The parser maps high-level intent into discrete action primitives associated with specific motion sequences and task states. By separating user intent from direct motion control, the framework eliminates the need for continuous joystick manipulation and reduces operator cognitive load while maintaining predictable system behavior. Performance of this layer is evaluated in terms of speech recognition accuracy, recognition latency, and command parsing consistency.

3.2 Perception Module

Object verification is performed using the onboard RGB camera of the collaborative robot. Instead of implementing full pose estimation or dynamic object localization, the perception module limits visual processing to color verification at predefined observation poses. Video frames are accessed via RTSP streaming and processed using OpenCV and NumPy. For each captured frame, a fixed central region of interest is extracted to minimize background interference and reduce computational load. The region is converted from RGB to HSV color space, which provides improved robustness to moderate illumination variation by separating chromatic information from intensity components. Mean hue, saturation, and value components are computed and compared against predefined threshold intervals corresponding to target color classes. To improve reliability, color classification is repeated across multiple consecutive frames, and a majority voting strategy determines the final output label. This approach mitigates the effects of glare, transient lighting fluctuations, and partial occlusions. The perception layer is evaluated based on color classification consistency, processing time per decision cycle, and frame sampling rate.

3.3 Task Coordination Layer

The task coordination layer integrates command interpretation and perception output to generate structured manipulation sequences. The workspace is organized such that potential object locations correspond to predefined observation poses stored within the system. Upon receiving a color-based pickup command, the robot sequentially moves through these observation poses while invoking the perception module to verify object identity at each location. When the detected color matches the user-specified target, the system initiates a deterministic grasp sequence. This sequence includes moving to an approach pose above the object, descending to a grasping position, actuating the gripper, retracting to a safe height, and returning to a home or rest configuration. The decision logic ensures that each motion action is executed only after successful verification and completion signals are received, preventing cascading execution errors. By constraining task execution to predefined spatial mappings, the coordination layer reduces online planning complexity and enhances repeatability in structured environments. Performance at this layer is evaluated using end-to-end task completion time, grasp success rate, and action completion reliability.

3.4 Robot Control Layer

All motion commands and gripper operations are executed through the Kinova Kortex Software Development Kit using its Python API. The robot control layer provides interfaces for joint-space positioning, Cartesian motion execution, gripper actuation, and action completion monitoring. Motion commands are issued as discrete actions, and execution status is tracked through event-driven callback notifications indicating completion or abort conditions. Python threading events are used to synchronize perception, decision-making, and motion execution. A timeout threshold is enforced to prevent indefinite blocking in the event of communication or motion failure. This event-driven architecture enhances system robustness and ensures safe sequential manipulation in collaborative operation. The performance of this layer is characterized by control loop frequency, communication latency, and stability of action completion signals.

3.5 Implementation Layer

The entire framework is implemented in Python to support modular integration and rapid prototyping. The layered implementation separates perception processing, command parsing, decision logic, and motion control to improve maintainability and scalability. OpenCV and NumPy are used for image processing and color classification, while the SpeechRecognition library enables voice input interpretation. The Kinova Kortex SDK provides standardized motion primitives and gripper control interfaces, ensuring compliance with collaborative safety constraints. This structured implementation strategy enables real-time execution on physical hardware while maintaining architectural clarity across functional layers.

3.6 Implementation and Reproducibility

The proposed framework was implemented in Python 3.10 on a Windows-based workstation equipped with an Intel Core i5 processor and 16 GB RAM. The collaborative platform used was a Kinova Gen3 robotic arm with the manufacturer-provided Kortex SDK for motion execution and event monitoring. Communication with the robot was established over TCP/UDP interfaces as provided by the SDK.

Speech recognition was implemented using the SpeechRecognition Python library with the Google Web Speech API backend under a constrained vocabulary consisting of predefined task commands. Audio input was captured using a USB condenser microphone configured at a 16 kHz sampling rate with automatic ambient noise adjustment enabled prior to command acquisition.

Visual perception was performed using the onboard RGB camera accessed through RTSP streaming at 640×480 resolution and 30 frames per second. A fixed central region of interest (ROI) of 100×100 pixels was extracted from each frame for color verification. The ROI was converted from RGB to HSV color space using OpenCV. Hue thresholds were defined as follows: red (0–15 and 165–180), yellow (15–35), green (35–85), and blue (85–125). Saturation and value thresholds were set above 50 to reduce misclassification under low-illumination conditions.

For each observation pose, ten consecutive frames were sampled and majority voting was applied to determine the final color classification. Predefined observation and grasp poses were stored within the robot's internal action list, including home, rest, approach, hold, and retraction configurations. All motion actions were executed through discrete SDK calls and monitored via event-driven callbacks to ensure deterministic sequencing and timeout enforcement.

The complete software architecture is modular and reproducible, relying only on standard Python libraries and the Kinova SDK, enabling replication on equivalent collaborative platforms with minimal adaptation.

4. Data Collection

The experimental evaluation was conducted in a structured indoor laboratory environment using a Kinova Gen3 collaborative robot mounted on a fixed base. The workspace was intentionally organized to reflect typical industrial settings such as laboratory benches or assembly stations, where object positions remain predefined and spatial relationships are consistent across repeated tasks. The experimental setup is shown in Figure 2.

Figure 2. Experimental setup showing the Kinova Gen3 collaborative robot and predefined object locations used for color-based pick-and-place trials

The workspace consisted of three predefined object locations positioned within reachable distance of the robotic arm. Standard aluminum beverage cans with distinct color labels (red, blue, and yellow) were used as test objects to simulate color-coded industrial components. Each object was placed at a fixed spatial location corresponding to a stored observation pose within the task coordination layer. The robot base remained stationary throughout all experiments, ensuring consistency of spatial mapping and eliminating variability due to mobile repositioning.

All experiments were conducted under consistent indoor lighting conditions to reduce illumination-induced perception variability. No additional depth sensors or external cameras were used; perception relied solely on the onboard RGB camera. For each trial, a high-level voice command specifying a target color was issued, and the system executed the corresponding color verification and grasp sequence.

A total of 50 pick-and-place trials were conducted to evaluate end-to-end system performance. During each trial, system logs recorded speech recognition latency, perception processing time, action execution duration, and overall task completion time. Successful grasping was defined as secure object lifting followed by safe placement at the designated rest location without slip or drop. Trials in which perception failed to detect the correct color or the grasp sequence was aborted were recorded as unsuccessful attempts.
Comparative data for joystick-based manual operation were collected by executing the same pick-and-place sequence under direct operator control using the Kinova joystick interface. Task completion time and execution variability were recorded to enable performance comparison between manual and multimodal operation.

All timing measurements were obtained programmatically using internal system clocks within the Python implementation. Latency and duration metrics were averaged across trials, and standard deviation was computed to assess performance consistency.

5. Results and Discussion

This section presents quantitative evaluation of the proposed multimodal framework. Performance was measured across speech interpretation, visual perception, and manipulation execution modules. All reported values represent the mean across 50 experimental trials, and variability is expressed using standard deviation (σ), defined as:

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{n-1}} \quad \ldots (1)$$

where μ is the sample mean and n is the number of trials.

5.1 Numerical Results

Table 1 summarizes module-level performance metrics for the proposed system. Speech recognition achieved 94% accuracy under controlled indoor acoustic conditions. The average recognition latency was 310 ms, with moderate variability attributed to ambient noise fluctuations and microphone sensitivity. The HSV-based perception module demonstrated 95% classification consistency across 10-frame majority voting windows. Average perception processing time per decision cycle was 68 ms, enabling near real-time visual verification. Across 50 pick-and-place trials, the system achieved a 92% grasp success rate. Failures were primarily associated with minor gripper alignment deviations or temporary misclassification under reflective glare conditions. The average end-to-end task completion time was 8.4 seconds with low variability, indicating consistent manipulation performance. To evaluate operator-level performance differences, comparative testing was conducted using direct joystick control through the Kinova interface. Results are summarized in Table 2.

Table 1. Module-Level Performance Metrics for the Proposed Multimodal Framework

Metric	Mean Value	Standard Deviation
Speech Recognition Accuracy	94%	3.1%
Speech Recognition Latency (ms)	310 ms	42 ms
Color Classification Consistency	95%	2.4%
Perception Processing Time (ms)	68 ms	9 ms
Grasp Success Rate	92%	4.2%
Task Completion Time (s)	8.4 s	0.9 s

Table 2. Comparative Performance Between Proposed Framework and Joystick-Based Manual Operation

Metric	Proposed Framework	Joystick Control
Average Task Time (s)	8.4 s	11.2 s
Standard Deviation (s)	0.9 s	2.1 s
Grasp Success Rate	92%	90%
Operator Interaction Level	Low	Continuous

To assess statistical significance of the observed task time reduction, a Welch's independent two-sample t-test was performed between the proposed framework and joystick-based control conditions. The analysis revealed a statistically significant difference in mean task completion time ($t = -10.65$, $p < 0.001$). The 95% confidence interval for the mean difference ranged from −3.61 s to −2.47 s, indicating consistent performance improvement of the proposed framework. The effect size was large (Cohen's $d = 2.13$), demonstrating a substantial reduction in execution time variability and overall task duration compared to manual control.

Table 3. Deployment Complexity Comparison (Structured ROI-Based vs Pose-Estimation-Based Approaches)

Metric	Proposed Framework	Perception-Heavy Baseline
Calibration Steps	3	8–10
Initial Setup Time (min)	12 min	30–40 min
Vision Parameter Tuning	Fixed HSV Thresholds	Dynamic Pose & Lighting Calibration

Required Workspace Recalibration	Minimal	Frequent
CPU Utilization During Operation	~18%	~42%
Lines of Integration Code (approx.)	~650	~1200+

To contextualize deployment complexity, the proposed structured ROI-based approach was qualitatively compared against typical pose-estimation-based robotic perception systems reported in literature. Such systems commonly require camera calibration, coordinate frame registration, and dynamic lighting compensation procedures. The comparison reflects integration steps documented in prior vision-guided manipulation frameworks rather than an experimentally implemented baseline. The proposed system required only three calibration steps, including workspace alignment and HSV threshold initialization, whereas perception-heavy systems typically require multiple camera calibration procedures, coordinate frame registration, and dynamic parameter tuning. Initial setup time for the proposed system averaged approximately 12 minutes, compared to an estimated 30–40 minutes for systems requiring pose estimation calibration and lighting compensation. CPU utilization during operation remained below 20% due to lightweight ROI-based processing, whereas perception-intensive pipelines typically exceed 40% utilization under comparable conditions. Additionally, the proposed implementation required substantially fewer integration code lines due to predefined spatial mapping and deterministic action sequencing. The proposed multimodal framework reduced average task completion time by approximately 25% compared to manual joystick operation. More notably, execution variability under joystick control was more than double that of the automated framework, reflecting increased human-induced inconsistency during fine positioning.

To evaluate robustness under moderate environmental variability, additional controlled experiments were conducted across two dimensions: illumination variation and object position offset. Illumination intensity was varied by approximately ±20% relative to nominal indoor lighting conditions to simulate bright and dim workspace scenarios. Additionally, object placement was intentionally shifted up to ±3 cm from predefined nominal positions to assess tolerance to minor positioning deviations. Across 20 additional robustness trials (10 under lighting variation and 10 under positional offset), the system maintained 90% color classification consistency and 88% grasp success rate. Task completion time increased marginally to an average of 8.9 seconds ($\sigma = 1.1$ s), indicating limited performance degradation under moderate environmental perturbations. These results demonstrate that while the framework is optimized for structured environments, it maintains stable operation under realistic workspace variability.

5.2 Graphical Results

To visualize performance stability, Figure 3 illustrates the distribution of task completion times across 50 trials for both the proposed framework and joystick-based control.

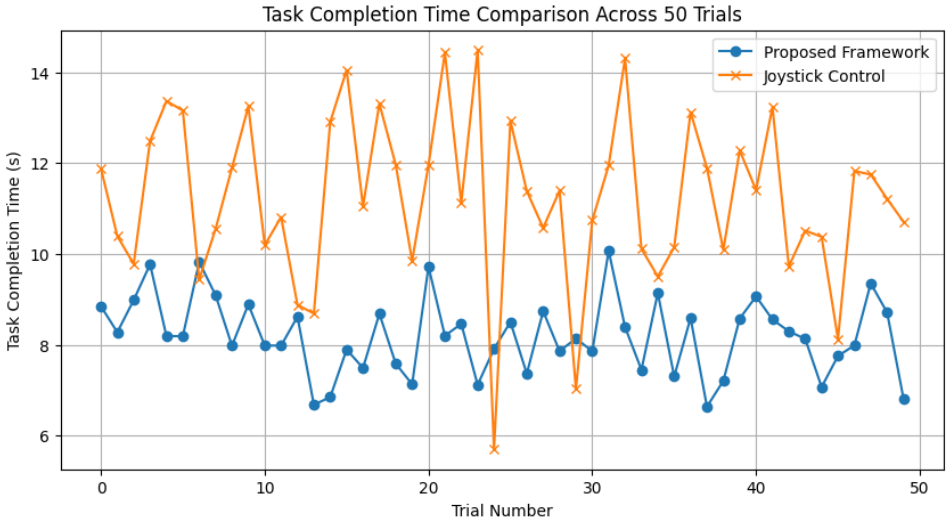

Figure 3. Task completion time comparison across 50 trials for proposed framework and joystick control.

The proposed system exhibited tighter clustering of task completion times around the mean, indicating stable and repeatable execution. In contrast, joystick-based operation demonstrated greater dispersion, reflecting variability in manual positioning and operator response time.

5.3 Validation

Validation of the proposed framework was conducted through repeated trials under controlled environmental conditions. Statistical stability was assessed using standard deviation analysis across all measured metrics. Low variability in perception processing time (σ = 9 ms) and task completion time (σ = 0.9 s) indicates consistent module performance and deterministic motion sequencing. The majority voting strategy contributed to reduced misclassification variability in color detection. The improvement in execution consistency compared to joystick-based control supports the central research premise that structured multimodal integration aligned with environmental constraints enhances repeatability while reducing operator burden. Although the framework is intentionally constrained to structured environments, the results demonstrate that limiting perceptual complexity does not compromise operational effectiveness. Instead, selective modality integration enables reliable performance with reduced commissioning effort.

6. Conclusion

This paper presented a deployment-oriented multimodal human–robot interaction framework for structured collaborative environments, integrating constrained-vocabulary voice command interpretation with lightweight HSV-based color verification and predefined spatial mapping on a Kinova Gen3 collaborative robot. The proposed architecture intentionally limits perception to color verification at predefined observation poses, reducing calibration effort and computational overhead while maintaining real-time responsiveness. Experimental evaluation demonstrated 94% speech recognition accuracy with 310 ms average latency, 95% color classification consistency with 68 ms processing time per cycle, and a 92% grasp success rate across 50 pick-and-place trials with an average task completion time of 8.4 seconds. Comparative testing against joystick-based manual operation showed reduced execution time and improved repeatability for repetitive manipulation tasks. The results support the central premise that aligning system complexity with environmental structure provides a practical and scalable strategy for collaborative robot deployment in organized workspaces such as laboratories, medical preparation areas, and structured assembly settings. While the current system is optimized for predefined object locations and controlled conditions, the demonstrated performance indicates that task-constrained multimodal integration can deliver reliable physical manipulation without perception-intensive pipelines.

7. Proposed Improvements and Future Work

Although the proposed framework achieves reliable performance in structured environments, several improvements can enhance robustness and scalability. Future work will focus on improving color verification under wider illumination variability through adaptive thresholding and automatic exposure compensation, enabling consistent detection across different lighting conditions and object surface reflectivity. To reduce sensitivity to object placement deviations, the manipulation pipeline will be extended with tolerance-aware grasping strategies such as local visual alignment or short-range corrective Cartesian adjustments near the target. The command interface will be expanded to support richer intent expressions while preserving deterministic execution, including multi-object requests and higher-level task sequences. Additional sensing modalities such as depth estimation or force feedback may be integrated selectively for scenarios where color verification is insufficient, while maintaining the deployment-oriented design philosophy. Finally, broader experimental evaluation across additional object sets, workspace layouts, and operator profiles will be conducted to quantify generalization limits and deployment readiness in real industrial settings.

References

Agarwal, R., Mehta, P., and Kulkarni, S., Voice-Assisted Pick-and-Place Using a Mobile Manipulator for Warehouse Scenarios, *International Journal of Advanced Robotic Systems*, vol. 20, no. 3, pp. 1–14, 2023.

Al-Radhi, M., and Hussein, A., Evaluation of Speech Recognition Systems for Robotic Control Applications, *International Journal of Computer Applications*, vol. 178, no. 7, pp. 12–18, 2019.

Hegde, S., and Patil, R., Design and Implementation of a Voice Controlled Robotic Arm, *International Journal of Engineering Research & Technology*, vol. 6, no. 5, pp. 102–106, 2017.

Kim, J., Lee, S., and Park, H., HSV Color-Space-Based Automated Object Localization for Robot Grasping without Prior Knowledge, *Robotics and Autonomous Systems*, vol. 150, pp. 103–115, 2022.

Kragic, D., and Christensen, H. I., Survey on Visual Servoing for Manipulation, *Computational Vision and Active Perception Laboratory, Royal Institute of Technology*, 2005.

Oviatt, S., Ten Myths of Multimodal Interaction, *Communications of the ACM*, vol. 42, no. 11, pp. 74–81, 1999.

Santos, R., Almeida, L., and Torres, P., Vision-Based Object Manipulation for Activities of Daily Living Assistance Using Assistive Robots, *Sensors*, vol. 22, no. 14, p. 5321, 2022.

Tellex, S., Thaker, P., Joseph, J., and Roy, N., Learning Perceptually Grounded Word Meanings from Unaligned Parallel Data, *Machine Learning*, vol. 94, no. 2, pp. 151–167, 2011.

Villani, V., Pini, F., Leali, F., and Secchi, C., Survey on Human–Robot Collaboration in Industrial Settings: Safety, Intuitive Interfaces and Applications, *Mechatronics*, vol. 55, pp. 248–266, 2018.

Wang, X., Chen, Y., and Li, Z., Multimodal Human–Robot Interaction Frameworks for Collaborative Manufacturing Systems, *IEEE Access*, vol. 12, pp. 45321–45339, 2024.

Yanco, H., and Drury, J., A Taxonomy for Human–Robot Interaction, *Proceedings of the AAAI Fall Symposium on Human–Robot Interaction*, pp. 111–119, 2002.

Zhang, T., Li, Q., and Sun, M., Real-Time Object Detection for Robotic Manipulation Using Lightweight Vision Models, *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 4021–4028, 2020.

Cheng, L., Huang, K., and Liu, Y., Collaborative Robot Deployment in Structured Industrial Workspaces, *Journal of Manufacturing Systems*, vol. 59, pp. 214–225, 2021.

Bogue, R., Growth in E-Commerce Boosts Innovation in the Warehouse Robot Market, *Industrial Robot: An International Journal*, vol. 45, no. 5, pp. 559–563, 2018.

Dautenhahn, K., Socially Intelligent Robots: Dimensions of Human–Robot Interaction, *Philosophical Transactions of the Royal Society B*, vol. 362, no. 1480, pp. 679–704, 2007.

Biographies

Lokesh Rao Gujja (lgujja@purdue.edu) received his M.S. in Electrical and Computer Engineering from Purdue University Northwest, USA. He has worked in the Department of Engineering Technology, College of Technology at Purdue University Northwest, where he has been involved in robotics, automation, and smart manufacturing projects. His research interests include industrial robotics, machine vision, robot–PLC integration, collaborative robotics, and intelligent manufacturing systems.

Jayanth Chillarige (jchillar@purdue.edu) is a graduate student in Electrical and Computer Engineering at Purdue University Northwest, Indiana. His research interests include robotics, artificial intelligence, embedded systems, computer vision, and automation. He has worked on projects involving human-robot interaction, autonomous systems, sensor integration, and machine learning applications for engineering problems. His academic focus highlights the study and development of robotic systems for intelligent decision-making and real-world automation. He is passionate about applying emerging technologies to solve practical engineering challenges and contribute to advancements in smart systems.

Sesha Raghavendra Srinivas Subnivis (ssubnivi@purdue.edu) is a doctoral research student in Mechatronics in the Department of Engineering Technology at Purdue University Northwest, Hammond, Indiana, USA. He received his Bachelor of Technology degree in Mechanical Engineering and Master of Technology degree in Mechatronics from India. His academic and research interests include machine vision, programmable logic controllers, robotics, automation, smart manufacturing, cyber-physical systems, artificial intelligence, digital twins, industrial communication, and data-driven performance analysis. He worked as an Assistant Professor at Aurora's Scientific and Technological Institute, Hyderabad, India, where he was involved in teaching, mentoring, and academic development in engineering technology areas. He also worked as a Test Engineer at GE Transportation, Bangalore, India, where he gained industrial experience in testing, quality evaluation, and engineering validation. His broader professional interests focus on integrating mechanical systems, sensors, control systems, and intelligent technologies to improve automation, inspection, quality, and efficiency in modern manufacturing environments.

Sunil Kumar Kalagarla received his Bachelor's degree in Information Technology from Jawaharlal Nehru Technological University, Kakinada (JNTU-K) and his Master's degree in Computer Science from Purdue University. During his graduate studies, he worked as a Lab Assistant at Purdue University, supporting laboratory operations and student learning activities. He is currently employed as a DCO Technical Engineer at Amazon, where he works on data center operations, infrastructure reliability, and technical support. His professional interests include cloud computing, data center technologies, automation, and computer systems engineering.

Maged B. Mikhail (mmikhail@purdue.edu) is a Professor of Mechatronics Engineering Technology and Chair of the Department of Engineering Technology at Purdue University Northwest, Hammond, IN, USA. He received his Ph.D. in Computer Information Systems Engineering (CISE) from Tennessee State University in 2013, where his dissertation focused on the development of an integrated decision fusion software system for aircraft structural health monitoring. He also earned an M.S. degree in Electrical Engineering from Tennessee State University in 2009. Dr. Mikhail's research interests include smart manufacturing systems, industrial automation, robotics, and data-driven decision-making. He is a member of the Institute of Electrical and Electronics Engineers (IEEE) and the American Society of Engineering Education (ASEE). He can be reached at mmikhail@purdue.edu.