

Embodied AI in the Operating Room: A Voice-Interactive Robotic Scrub Technician for Surgical Instrument Handoffs

Mustafa Khan

The University of Texas Health Science Center at Houston

Joao Bombardelli

`joao.m.bombardelli@uth.tmc.edu`

The University of Texas Health Science Center at Houston

Research Article

Keywords: artificial intelligence, embodied AI, vision-language-action models, surgical robot

Posted Date: August 6th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10438744/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Journal of Robotic Surgery on August 27th, 2026. See the published version at <https://doi.org/10.1007/s11701-026-03866-9>.

Abstract

Background

Embodied artificial intelligence may extend surgical robotics beyond teleoperation by enabling robots to interpret natural-language commands, perceive dynamic environments, and generate adaptive physical actions. Surgical instrument exchange provides a clinically recognizable test case requiring communication, visual localization, dexterity, and bidirectional interaction. We developed a voice-interactive robotic scrub technician capable of instrument handoff and take-back using a vision-language-action model.

Methods

A bimanual robotic platform was trained by fine-tuning the pretrained $\pi 0.5$ model on 350 human-teleoperated demonstrations. The system integrated real-time multicamera observations, Whisper large-v2 speech recognition, language-conditioned robotic action generation, and text-to-speech acknowledgment. Performance was evaluated in 100 autonomous trials across five tasks: needle-driver handoff, forceps handoff, needle-and-suture handoff, needle-driver take-back, and forceps take-back. Each task included 10 seen and 10 held-out tray configurations. Outcomes included command interpretation, stage-wise completion, end-to-end success, completion time, and failure mode.

Results

The system completed 81% of trials successfully. Success was 95% for both take-back tasks, 85% for forceps handoff, 80% for needle-driver handoff, and 50% for needle-and-suture handoff. Performance decreased from 92% in seen layouts to 70% in held-out layouts. Speech-command interpretation succeeded in 99% of trials. Once a secure grasp was achieved, downstream manipulation succeeded in 81 of 82 trials. Missed grasp was the predominant failure mode.

Conclusions

A unified learned policy enabled voice-conditioned bidirectional surgical instrument exchange and demonstrated meaningful generalization to held-out layouts. Improved grasp acquisition, expanded instrument diversity, quantitative recovery testing, and evaluation in realistic operating-room workflows are needed before clinical translation or integration with complete back-table workflows.

INTRODUCTION

As we enter the fourth industrial revolution, artificial intelligence (AI) and robotics are poised to bring transformative changes across multiple fields, and surgery will be no exception. The operating room has historically evolved through the adoption of new technologies, from improved anesthesia and sterile

technique to laparoscopy, microsurgery, and robotic-assisted surgery [1]. However, despite these advances, most surgical robots remain dependent on direct human control and are limited in their ability to perceive, reason, and interact autonomously within the dynamic surgical environment [2, 3].

The emergence of embodied AI represents a new frontier in surgical robotics. Embodied AI refers to the integration of artificial intelligence into physical systems, allowing machines not only to process information, but also to interact with the real world through perception, decision-making, and action. Recent advances in vision-language-action (VLA) models have enabled robotic systems to interpret visual information, respond to natural language commands, and perform complex motor tasks in less structured environments [4–6]. These developments suggest that future robotic systems may be capable of assisting surgical teams in ways that extend beyond traditional teleoperation, shifting the intraoperative paradigm toward co-adaptive human-machine partnerships [7, 8].

One area with significant potential for embodied AI is intraoperative assistance [7]. Surgical procedures rely heavily on coordinated teamwork, anticipation of surgeon needs, and precise instrument handling. The scrub technician plays a critical role in maintaining operative flow by preparing, organizing, and handing off instruments at the appropriate time [9]. Although this task may appear routine, it requires visual recognition, contextual understanding, manual dexterity, and real-time interaction with the surgeon [10]. Furthermore, standardizing these workflows can mitigate systemic risks related to miscommunication and operational inefficiencies during instrument handling [10, 11]. Although classical computer vision and visual-servoing systems have attempted to anticipate or manipulate tools using visual input [12, 13], linking perception directly to physically grounded task execution provides an appropriate model for studying embodied AI in the operating room.

Several robotic scrub nurse systems have been developed to automate aspects of this workflow. Early prototypes relied on rule-based computer vision and speech or gesture recognition to identify and deliver a fixed set of instruments from a static tray [11, 14]. More recent laparoscopic systems have focused on predicting instrument needs from live video or optimizing the mechanics of the handover itself [9, 12]. Most recently, RoboNurse-VLA demonstrated the first vision-language-action-based robotic scrub nurse, combining a pretrained OpenVLA/Llama 2 backbone with dedicated detection and segmentation modules to achieve accurate instrument handover to a surgeon’s hand [15]. However, to our knowledge, existing systems, including RoboNurse-VLA, have focused exclusively on instrument delivery and have not demonstrated take-back behavior or recovery from interference or initial failure. These capabilities would allow a robot to retrieve a used instrument, return it to the surgical tray, or recover from interference without complete task reinitialization. Our system uses a different foundation model, $\pi 0.5$, fine-tuned end to end on our teleoperated demonstrations, and extends this work by performing both delivery and take-back within a single unified policy.

In this article, we demonstrate how recent advances in machine learning and robotics can be applied to surgical workflow by developing a robotic scrub technician capable of performing surgical instrument handoff, retrieval, and recovery. By combining visual perception, language-guided instruction, and robotic manipulation, this proof-of-concept platform represents an early step toward intelligent robotic assistants

that may one day improve efficiency, reduce workload, and enhance coordination within surgical teams [7, 8, 16].

METHODS

We developed a robotic scrub technician platform designed to identify, select, hand off, and take back surgical instruments in response to surgeon commands, replicating common instrument-passing tasks performed by a scrub technician in the operating room. The system was built using a robotic manipulation platform integrated with visual perception, natural-language understanding, and action generation through a pretrained VLA model (Fig. 1).

The robotic system was trained using imitation learning, in which robot actions are learned from human demonstrations rather than manually programmed step by step [17, 18]. Demonstrations were collected through teleoperation while the robot performed surgical instrument handoff and take-back tasks. During each demonstration, synchronized video observations, surgeon commands, and robot joint actions were recorded. These demonstrations provided the model with paired examples of visual input, language instruction, and corresponding physical actions. The final task-specific fine-tuning dataset comprised 350 human-teleoperated episodes spanning instrument handoff and take-back behaviors (Video 1 at <https://surgicalintelligencelabs.com/asia>, Teleoperation and recording).

The learning architecture was based on a pretrained VLA model, $\pi 0.5$, developed by Physical Intelligence [4, 19, 20] (Fig. 2). VLA models combine visual information, natural-language prompts, and robot action outputs within a single framework. In this setting, the model receives image observations of the surgical tray and a language command, such as a request for a specific surgical instrument, and generates the robot actions needed to locate, grasp, hand off, or take back the instrument. The model was fine-tuned on our surgical instrument handoff and take-back dataset to adapt its pretrained manipulation capabilities to this task.

A simulated surgical tray environment was created to evaluate the feasibility of robotic instrument recognition, selection, handoff, and retrieval. Surgical instruments were arranged on the tray in variable positions to test the system's ability to respond to different visual configurations. The robot was prompted with surgeon-style verbal commands requesting specific instruments. The task was considered successful when the robot correctly identified the requested instrument, grasped or retrieved it, transported it toward the surgeon's hand or the tray, and released it in a controlled manner.

The robotic workflow consisted of four primary steps: command interpretation, instrument localization, robotic grasping, and instrument handoff [9, 21]. In addition to instrument delivery, the system was trained to perform take-back behavior, allowing the robot to retrieve an instrument after use and return it to the surgical tray. The teleoperation, data recording, data conversion, model training, and inference pipelines required extensive new coding and previous code modification to interface the pretrained model with our specific robotic hardware, camera configuration, and surgical task environment.

Performance was evaluated in a controlled, simulated operating room setting using repeated instrument handoff and take-back trials. Outcomes included recognition of the requested instrument, completion of the handoff or take-back task, recovery from interference or initial errors, and responses to different instrument commands. This proof-of-concept evaluation assessed whether an embodied AI system could perform basic scrub technician instrument exchanges using vision, language, and robotic manipulation.

Speech Interface

To create a more natural surgeon-robot interaction, the system was integrated with a speech interface that allowed the surgeon to issue verbal instrument requests rather than relying solely on typed commands. Spoken commands were captured through a microphone and transcribed using Whisper large-v2, a large-scale automatic speech recognition model [22]. The transcribed text was then passed to the VLA model as the language instruction for the robotic task.

This speech-to-action pipeline allowed the surgeon to request instruments using natural language, such as “hand me the forceps” or “take back the needle driver.” After transcription, the command was processed by the robotic control system and paired with the real-time visual input from the surgical tray. The model then generated the corresponding robot actions required to identify, grasp, deliver, or retrieve the requested instrument.

To improve human-robot interaction and provide feedback to the surgical operator, the system was also integrated with text-to-speech synthesis using the Cartesia API [23]. This allowed the robot to generate spoken responses confirming task interpretation, execution status, or task completion. For example, after receiving a command, the system could verbally acknowledge the requested instrument before initiating the handoff sequence.

The speech-to-text and text-to-speech modules were incorporated into the inference pipeline as an interface layer between the surgeon and the VLA model. This design allowed the core robotic policy to remain focused on visual perception and action generation, while the speech interface translated natural verbal communication into machine-readable instructions and returned audible feedback to the operator. This multimodal communication framework was intended to more closely resemble the verbal interaction that occurs between a surgeon and scrub technician in the operating room.

Use of generative artificial intelligence

Generative artificial intelligence was used to assist with language editing, manuscript organization, and formatting. All scientific content, interpretations, references, and revisions were reviewed and approved by the authors, who take responsibility for the final manuscript.

RESULTS

Evaluation Overview

The final evaluation included 100 autonomous trials across five scrub-technician tasks: needle driver handoff, forceps handoff, needle-and-suture handoff, needle driver take-back, and forceps take-back. Each task was evaluated in 20 independent trials, comprising 10 trials using seen tray configurations and 10 trials using held-out configurations. Across all tasks, the robotic system completed 81 of 100 trials successfully (81%; Wilson 95% confidence interval [CI], 72%–87%). Among successful trials, the mean completion time was 7.1 ± 1.4 seconds.

Task-Specific Performance

Task-level performance is summarized in Table 1. The robot successfully completed 16 of 20 needle driver handoffs (80%; 95% CI, 58%–92%), 17 of 20 forceps handoffs (85%; 95% CI, 64%–95%), and 10 of 20 needle-and-suture handoffs (50%; 95% CI, 30%–70%). Successful handoff trials were completed in 8.4 ± 0.2 seconds for the needle driver, 8.4 ± 0.2 seconds for forceps, and 8.4 ± 0.2 seconds for the needle-and-suture task.

Table 1
Task-level performance of the robotic scrub technician.

Task	Successful trials, n/N (%)	Wilson 95% CI	Seen	Held-out	Time, s, mean $\pm$ SD
Needle driver handoff	16/20 (80%)	58%-92%	10/10	6/10	8.4 ± 0.2
Forceps handoff	17/20 (85%)	64%-95%	10/10	7/10	8.4 ± 0.2
Needle-and-suture handoff	10/20 (50%)	30%-70%	6/10	4/10	8.4 ± 0.2
Needle driver take-back	19/20 (95%)	76%-99%	10/10	9/10	6.0 ± 0.8
Forceps take-back	19/20 (95%)	76%-99%	10/10	9/10	5.5 ± 0.2
Overall	81/100 (81%)	72%-87%	46/50	35/50	7.1 ± 1.4
Completion time was calculated among successful trials only. CI indicates confidence interval; SD, standard deviation.					

Take-back behavior was more reliable than instrument handoff. The system completed 19 of 20 needle driver take-back trials (95%; 95% CI, 76%–99%) and 19 of 20 forceps take-back trials (95%; 95% CI, 76%–99%). Mean completion times were 6.0 ± 0.8 seconds and 5.5 ± 0.2 seconds, respectively. When pooled by task type, the robot completed 43 of 60 handoff trials (71.7%; 95% CI, 59%–81%) and 38 of 40 take-back trials (95.0%; 95% CI, 83%–99%).

Generalization to Held-Out Tray Configurations

Performance decreased when the robot was evaluated on tray configurations not represented during training (Fig. 3). The system completed 46 of 50 trials using seen layouts (92%; 95% CI, 81%–97%)

compared with 35 of 50 trials using held-out layouts (70%; 95% CI, 56%–81%), corresponding to a 22-percentage-point decrease in end-to-end success.

For needle driver handoff, forceps handoff, and needle-and-suture handoff, success decreased from 10/10 to 6/10, 10/10 to 7/10, and 6/10 to 4/10, respectively. Take-back performance remained comparatively stable, decreasing from 10/10 to 9/10 for both the needle driver and forceps. Across all handoff tasks, success was 26/30 (86.7%) on seen layouts and 17/30 (56.7%) on held-out layouts. Across take-back tasks, success was 20/20 (100%) on seen layouts and 18/20 (90%) on held-out layouts.

Stage-Wise Performance

Stage-wise performance is shown in Table 2. Speech-command interpretation was successful in 99 of 100 trials (99%). The correct instrument or target was identified in 89 trials (89%), grasp acquisition was successful in 82 trials (82%), and transport was completed in 81 trials (81%). Handoff or placement, final orientation, and overall task completion were each successful in 81 trials (81%).

Table 2 Stage-wise success across autonomous trials.							
Task	Command/ASR	Correct target	Grasp	Transport	Handoff/place	Orientation	Overall
Needle driver handoff	100%	100%	80%	80%	80%	80%	80%
Forceps handoff	100%	90%	85%	85%	85%	85%	85%
Needle-and-suture handoff	95%	65%	55%	50%	50%	50%	50%
Needle driver take-back	100%	95%	95%	95%	95%	95%	95%
Forceps take-back	100%	95%	95%	95%	95%	95%	95%
Overall	99%	89%	82%	81%	81%	81%	81%
ASR indicates automatic speech recognition.							

The largest stage-wise decline occurred during the needle-and-suture handoff task: command interpretation succeeded in 19/20 trials, correct target identification in 13/20, grasp acquisition in 11/20, and successful transport and final handoff in 10/20. In contrast, command interpretation was successful in all 40 take-back trials, and each subsequent stage was successfully completed in 38/40 trials.

Failure Analysis

Nineteen trials did not meet the predefined end-to-end success criterion. Missed grasp was the most frequent failure mode, accounting for 11 of 19 failures (57.9%). Four failures (21.1%) resulted from policy timeout or freezing, two (10.5%) involved selection of an incorrect instrument, one (5.3%) reflected failure to recognize the requested instrument, and one (5.3%) resulted from dropping the instrument during transport.

Fifteen of the 19 failures (78.9%) occurred during held-out-layout testing. No primary failures were categorized as unstable grasp, poor final orientation, failed release, or safety stop, and no trial required emergency termination for unsafe robot motion.

Voice-Interactive Autonomous Demonstration

The integrated speech interface enabled voice-initiated execution throughout the evaluation. The automatic speech-recognition pipeline correctly interpreted 99 of 100 commands, after which the transcribed instruction was passed to the vision-language-action policy for autonomous execution. The single command-interpretation failure occurred during the needle-and-suture handoff evaluation.

Representative autonomous instrument handoffs and take-back tasks with full system assistance are shown in Supplemental Video 2 at <https://surgicalintelligencelabs.com/asia>, Full policy (The fully autonomous robot provides complete assistance by responding to specific commands, handing instruments to the surgeon, and taking them back). Autonomous recovery from interference and execution failure is shown in Supplemental Video 3 at <https://surgicalintelligencelabs.com/asia>, Recoveries (The fully autonomous robot recovers from interference and execution failures).

DISCUSSION

Principal Findings and Technical Interpretation

This proof-of-concept study quantitatively evaluated a voice-interactive embodied artificial intelligence system for bidirectional surgical instrument exchange. Across 100 autonomous trials spanning five scrub-technician tasks, the system achieved an overall end-to-end success rate of 81% (95% confidence interval [CI], 72%-87%). Take-back was completed successfully in 38 of 40 trials (95%), and the speech interface correctly interpreted 99 of 100 commands. Once a stable grasp was achieved, the robot completed the remaining transport, orientation, handoff or placement, and release sequence in 81 of 82 trials. These findings localize the main performance limitation to visual localization and grasp acquisition rather than downstream trajectory execution.

The principal contribution is not a single isolated handoff but a more complete interaction loop. The same learned policy delivered instruments, received them back, returned them to the tray, and provided audible confirmation of interpreted commands. This combination extends prior VLA-based scrub assistance beyond one-way delivery and makes robot intent observable to the operative team [15]. The post-grasp

reliability further suggests that the action-generation component converted combined visual and language representations into stable manipulation trajectories. Achieving these behaviors after fine-tuning on 350 human-teleoperated episodes supports the feasibility of task-specific adaptation while leaving substantial room for improvement through broader training data.

Task Performance, Generalization, and Grasping

Performance varied by task. Forceps and needle-driver handoffs succeeded in 85% and 80% of trials, respectively, whereas needle-and-suture handoff succeeded in 50%. Both take-back tasks achieved 95% success. The lower needle-and-suture performance is consistent with a target that has a small visual footprint, narrow graspable surface, reflective curved geometry, variable orientation, and an attached deformable suture. These properties increase the difficulty of localization and stable grasp acquisition compared with larger, rigid, ring-handled instruments. Once the needle was securely grasped, the remaining transport and delivery sequence was generally completed successfully.

The comparison of seen and held-out tray layouts provides the clearest assessment of current generalization. Overall success decreased from 92% in seen layouts to 70% in held-out layouts. The gap was concentrated in handoff behavior, which decreased from 86.7% to 56.7%, while take-back remained high at 100% and 90%, respectively. Fifteen of the 19 failures occurred in held-out configurations. These results indicate that the policy learned the fundamental exchange behaviors but remained sensitive to unfamiliar instrument positions, orientations, and object-context combinations.

The held-out gap should be interpreted in the context of a training set containing only 350 task-specific episodes. Robotic foundation-model studies show that scene diversity, object diversity, semantic augmentation, cross-embodiment data, and action-chunking strategies can improve transfer and generalization [19]. The present findings do not establish a fixed ceiling for $\pi 0.5$. They identify incomplete coverage of spatial and object variation as a direct and testable limitation.

Future demonstrations should emphasize spatial randomization, variable instrument orientation, clutter, occlusion, lighting variation, neighboring-object configurations, and failed-grasp recovery. Needle-specific data should vary curvature orientation, suture arrangement, visual contrast, overlap, and grasp point. Higher-resolution or multiscale visual processing may also be needed because downsampling can remove the fine geometric features required to distinguish a small needle from its attached suture. Learning from autonomous rollouts, corrective interventions, and recovery examples may further improve grasp reliability [31–34].

Failure Recovery and Closed-Loop Control

Representative demonstrations showed the robot recovering from unsuccessful approaches and external interference without complete task reinitialization (Video 3 at <https://surgicalintelligencelabs.com/asia>, Recoveries and Fig. 4). After the physical scene changed, the policy incorporated subsequent visual observations, adjusted its movement, reacquired the target, and continued toward the original objective. This behavior is relevant to the operating room, where instruments can shift, grasps can become unstable, and human team members may temporarily enter the robot's workspace.

Conventional automation generally performs best when objects and trajectories remain consistent with predefined conditions. A closed-loop VLA policy instead conditions each action chunk on updated observations and the task instruction [4, 18, 30, 31, 35–37]. This permits corrective action after movement has begun rather than requiring continued execution of an obsolete open-loop trajectory. For instrument exchange, the policy must account for changes in the instrument, gripper, receiving hand, and surrounding workspace throughout the interaction.

Recovery is clinically important because performance under ideal, undisturbed conditions is insufficient for collaboration in a variable human environment. A useful assistant must remain predictable while responding to a displaced instrument, missed approach, interrupted trajectory, or changing hand position. The present demonstrations suggest that the policy learned more than a memorized sequence, but they do not establish unrestricted autonomous fault management.

Future studies should replace qualitative examples with standardized perturbation testing. Prespecified disturbances could include instrument displacement, temporary visual occlusion, trajectory interruption, failed grasp attempts, and changes in receiving-hand position. Recovery success, recovery time, number of corrective actions, and need for human assistance should be reported separately. These measurements would define which disturbances the policy can manage reliably and when surgeon or technician intervention is required.

Comparison With RoboNurse-VLA

RoboNurse-VLA is the most closely related prior system and represents an important milestone in voice-commanded VLA-based scrub assistance [15]. It combines a custom YOLOv8 detector, SAM 2 segmentation, an MLP projector, and a pretrained Llama 2/OpenVLA backbone. The present platform instead fine-tuned an end-to-end $\pi 0.5$ policy directly from 350 synchronized teleoperated episodes and did not require a separately trained instrument detector, segmentation model, or manually annotated detection dataset. The different designs illustrate complementary modular and unified approaches rather than directly comparable experimental conditions.

The present workflow also extends beyond one-way delivery. Handoff and take-back were learned within the same policy, reflecting the repeated bidirectional exchanges performed during operative care. Whisper large-v2 converted spoken requests into task instructions, while Cartesia text-to-speech generated audible acknowledgments. This feedback makes the robot's interpreted intent observable and provides a foundation for future clarification requests, uncertainty statements, warnings, and human override.

Execution architecture represents another distinction. RoboNurse-VLA reports VLA inference at 5 Hz with lower-level joint control at 30 Hz. The $\pi 0.5$ policy predicts sequential action chunks executed at 50 Hz, providing dense action updates during localization, grasping, transport, handoff, retrieval, and tray replacement [4, 31]. The near-universal completion of the post-grasp sequence is consistent with reliable downstream action execution, although the current study was not designed to attribute performance differences to control frequency alone.

The present evaluation additionally separated command interpretation, target identification, grasp acquisition, transport, placement, orientation, and end-to-end completion; compared seen and held-out layouts; included a free needle with attached suture; and demonstrated qualitative recovery after interference. Table 4 summarizes these differences. Because the studies used different hardware, instruments, training procedures, and evaluation protocols, the comparison should be interpreted as a distinction in scope and architecture, not as a statistical claim of superiority.

Data Efficiency and Scalability

The model's performance after 350 teleoperated episodes demonstrates task-specific learning from a comparatively compact dataset, but it should be viewed as an early point on a larger scaling trajectory [19]. The dataset was sufficient to acquire voice-conditioned task selection, localization, grasping, delivery, take-back, and tray replacement. It was not large enough to represent the full variability of an operating-room back table.

A complete back table may contain dozens of instrument families and approximately 100 instruments, sizes, and variants. Many share a metallic appearance and differ only in subtle geometry. They may be stacked, partially occluded, placed at variable angles, or returned to different locations during a procedure. Scaling will therefore require more demonstrations across instrument identity, tray composition, camera perspective, lighting, clutter, occlusion, surgeon hand position, and longer sequential workflows.

Clinical Role and Human-Robot Collaboration

Instrument exchange is an appropriate early application of embodied AI because it is structured and clinically recognizable but still requires communication, visual recognition, spatial reasoning, dexterity, and shared situational awareness [9–11, 14, 21, 38, 39]. Prior robotic scrub-nurse systems demonstrated portions of this workflow using fixed locations, rule-based control, gesture input, or video-predictive pipelines [11, 12, 14]. Challenges in workflow anticipation, spatial-temporal relationships, and tool context remain central to effective collaboration [40, 41].

Bidirectional exchange is important for meaningful team integration. A scrub technician repeatedly passes instruments, receives them after use, organizes the back table, and prepares subsequent items. A delivery-only robot addresses one part of this loop. Demonstrating both handoff and take-back provides a foundation for longer sequences in which the system tracks instrument location and use while remaining responsive to the surgical team.

Communication is equally important. A collaborative robot must make its interpreted intent understandable, move predictably, accept correction, and avoid disrupting operative flow [21, 38, 39]. Spoken acknowledgment allows the team to hear which task the system inferred before or during movement. Future interfaces could request clarification, express uncertainty, warn of an ambiguous instrument, or report that a grasp cannot be completed safely.

The intended role of this technology is assistive rather than substitutive. Human scrub technicians provide sterile-field management, inventory control, procedural anticipation, troubleshooting, and clinical judgment

that extend far beyond instrument transfer. Robotic assistance may be most useful for repetitive or ergonomically demanding exchanges while trained personnel retain responsibility for sterility, unexpected events, and operative coordination. This approach is consistent with co-adaptive human-machine teams and defined levels of human supervision [2, 3, 7, 8, 39].

More advanced surgical autonomy remains dependent on structured tasks, specialized systems, and human oversight. STAR demonstrated conditional autonomy for intestinal anastomosis, while microsurgical platforms and language-conditioned surgical policies illustrate progress in precision assistance and learned task execution [42–44]. Sequence-based imitation learning, specialized action policies, and automated workflow analysis may support future integration [35, 40, 45]. Nevertheless, instrument exchange is a more appropriate near-term target because it does not require independent diagnostic or operative decisions and can be developed within an explicitly supervised collaborative workflow.

Limitations and Future Directions

This study has several limitations. The platform was evaluated in a simulated surgical tray environment rather than a live operating room. Lighting, background variation, instrument diversity, sterility constraints, and human interaction were more controlled than they would be during surgery. The evaluation included isolated handoff and take-back tasks rather than a complete procedure, and no human scrub-technician comparator was included. Completion times therefore describe system performance but do not establish clinical efficiency.

Technical limitations were concentrated in target localization and grasp acquisition under held-out layouts and for the needle-and-suture task. The recovery evidence was qualitative, the instrument set was small, and the policy was not evaluated during prolonged sequential use. Future work should expand instrument and scene diversity, assess high-resolution perception, quantify standardized recovery, and evaluate longer workflows. Complementary learning approaches include real-time action conditioning, autonomous rollouts with human intervention, precision-focused reinforcement learning, and multiscale embodied memory [31–34].

Clinical translation will require sterility-compatible hardware, collision avoidance, workspace monitoring, surgeon and technician override, uncertainty communication, and clearly defined stopping rules. Testing should progress from controlled simulations to realistic operating-room environments and supervised workflow studies before any clinical deployment. Regulatory, ethical, and human-factors evaluation will be necessary to define responsibility, acceptable autonomy, and safe integration with the operative team [2, 3, 7, 8].

CONCLUSION

This proof-of-concept demonstrates that a $\pi 0.5$ -based robotic scrub technician can perform voice-conditioned bidirectional instrument exchange. After fine-tuning on 350 teleoperated episodes, the system achieved 81% overall success, 95% take-back success, and 99% speech-command interpretation across

100 autonomous trials. Once a secure grasp was achieved, downstream manipulation was highly reliable, identifying grasp acquisition in unfamiliar layouts and for small targets as the principal remaining limitation. These findings support further development of embodied AI for supervised operating-room assistance. Expanded training data, realistic workflow evaluation, and systematic safety and regulatory validation will be required before extension to complete back-table management or clinical use [4, 7, 8].

Declarations

Funding

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Competing interests

The authors have no relevant financial or non-financial interests to disclose.

Author contributions

M.K. conceived the study, contributed to system development, analyzed the data, and drafted the manuscript. J.B. conceived the study, contributed to system development, analyzed the data, and drafted the manuscript. M.B. contributed to study conception and manuscript drafting. All authors reviewed and approved the final manuscript.

Data availability

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Code availability

The custom code supporting the findings of this study is available from the corresponding author upon reasonable request.

Ethics approval

This study was conducted in a simulated operating-room environment and did not involve patients, human research participants, identifiable human data, animals, or biological specimens. Institutional review board approval and informed consent were therefore not required.

References

1. Morrell ALG, Morrell-Junior AC, Morrell AG, Mendes JM, Tustumi F, De-Carli LA (2023) The history of robotic surgery and its evolution: when illusion becomes reality. *Rev Col Bras Cir*. <https://doi.org/10.1590/0100-6991e-20202798>
2. Yang GZ, Cambias J, Cleary K et al (2017) Medical robotics—regulatory, ethical, and legal considerations for increasing levels of autonomy. *Sci Robot* 2:eaam8638. <https://doi.org/10.1126/scirobotics.aam8638>
3. Attanasio A, Scaglioni B, De Momi E, Fiorini P, Valdastrì P (2021) Autonomy in surgical robotics. *Annu Rev Control Robot Auton Syst* 4:651–679. <https://doi.org/10.1146/annurev-control-062420-090543>
4. Black K, Brown N, Driess D et al (2024) $\pi 0$: A vision-language-action flow model for general robot control. *arXiv*. <https://doi.org/10.48550/arXiv.2410.24164>. arXiv:2410.24164
5. Brohan A, Brown N, Carbajal J et al (2023) RT-1: robotics transformer for real-world control at scale. In: *Robotics: Science and Systems*. arXiv:2212.06817
6. Brohan A, Brown N, Carbajal J et al (2023) RT-2: vision-language-action models transfer web knowledge to robotic control. In: *Conference on Robot Learning*. arXiv:2307.15818
7. Granados A et al (2026) Evolving surgical teams in the age of artificial intelligence and robotics. *Front Sci*. <https://doi.org/10.3389/fsci.2026.1783803>
8. Long Y, Lin A, Kwok DHC et al (2025) Surgical embodied intelligence for generalized task autonomy in laparoscopic robot-assisted surgery. *Sci Robot*. <https://doi.org/10.1126/scirobotics.adt3093>
9. Scheppach C, Wagner L, Spinner C et al (2024) Innovating instrument handover techniques for robotic scrub nurses. *Curr Dir Biomed Eng*. <https://doi.org/10.1515/cdbme-2024-1052>
10. Wachs JP, Frenkel B, Dori D (2014) Operation room tool handling and miscommunication scenarios: an object-process methodology conceptual model. *Artif Intell Med* 62:153–163. <https://doi.org/10.1016/j.artmed.2014.10.006>
11. Jacob MG, Li YT, Akingba GA, Wachs JP (2013) Collaboration with a robotic scrub nurse. *Commun ACM* 56:68–75. <https://doi.org/10.1145/2447976.2447993>
12. Wagner L, Jourdan S, Mayer L et al (2024) Robotic scrub nurse to anticipate surgical instruments based on real-time laparoscopic video analysis. *Commun Med*. <https://doi.org/10.1038/s43856-024-00581-0>
13. Pérez C, Morales R et al (2008) Autonomous surgical robot using visual servoing in the extended image plane. *IFAC Proc Vol*. <https://doi.org/10.3182/20080706-5-KR-1001.02314>
14. Perez-Vidal C, Carpintero E, Garcia-Aracil N et al (2012) Steps in the development of a robotic scrub nurse. *Robot Auton Syst* 60:1544–1557. <https://doi.org/10.1016/j.robot.2012.01.005>
15. Li S, Wang J, Dai R et al (2024) RoboNurse-VLA: robotic scrub nurse system based on a vision-language-action model. *arXiv*. <https://doi.org/10.48550/arXiv.2409.19590>. arXiv:2409.19590
16. Maier-Hein L, Vedula SS, Speidel S et al (2017) Surgical data science for next-generation interventions. *Nat Biomed Eng* 1:691–696. <https://doi.org/10.1038/s41551-017-0132-7>

17. Kim JW, Zhao TZ, Schmidgall S et al (2024) Surgical Robot Transformer: imitation learning for surgical tasks. In: Conference on Robot Learning. <https://doi.org/10.48550/arXiv.2407.12998>
18. Zhao TZ, Kumar V, Levine S, Finn C (2023) Learning fine-grained bimanual manipulation with low-cost hardware. In: Robotics: Science and Systems. <https://doi.org/10.15607/RSS.2023.XIX.016>
19. Black K, Brown N, Darpinian J et al (2025) $\pi 0.5$: a vision-language-action model with open-world generalization. arXiv preprint
20. Driess D, Springenberg JT, Ichter B et al (2026) Knowledge insulating vision-language-action models: train fast, run fast, generalize better. arXiv preprint
21. Strabala K, Lee MK, Dragan A et al (2013) Toward seamless human-robot handovers. J Hum Robot Interact 2. <https://doi.org/10.5898/JHRI.2.1.Strabala>
22. Radford A, Kim JW, Xu T, Brockman G, McLeavey C, Sutskever I (2023) Robust speech recognition via large-scale weak supervision. In: International Conference on Machine Learning
23. Cartesia (2026) Sonic text-to-speech models: technical documentation
24. Vaswani A, Shazeer N, Parmar N et al (2017) Attention is all you need. Adv Neural Inf Process Syst. 30
25. Radford A, Kim JW, Hallacy C et al (2021) Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. 139:8748–8763
26. Alayrac JB, Donahue J, Luc P et al (2022) Flamingo: a visual language model for few-shot learning. Adv Neural Inf Process Syst 35:23716–23736
27. Beyer L, Steiner A, Pinto AS et al (2024) PaliGemma: a versatile 3B vision-language model for transfer. arXiv. arXiv:2407.07726.
28. Open X-E, Collaboration, Padalkar A, Pooley A et al (2024) Open X-Embodiment: Robotic learning datasets and RT-X models. In: IEEE International Conference on Robotics and Automation. <https://doi.org/10.48550/arXiv.2310.08864>
29. Kim MJ, Pertsch K, Karamcheti S et al (2024) OpenVLA: an open-source vision-language-action model. arXiv. <https://doi.org/10.48550/arXiv.2406.09246>. arXiv:2406.09246
30. Chi C, Feng S, Du Y et al (2023) Diffusion policy: visuomotor policy learning via action diffusion. In: Robotics: Science and Systems. arXiv:2303.04137
31. Black K, Galliker MY, Levine S (2025) Real-time execution of action chunking flow policies. In: Advances in Neural Information Processing Systems. arXiv:2506.07339
32. Amin A, Aniceto R, Balakrishna A et al (2025) $\pi^*0.6$: a VLA that learns from experience. arXiv. arXiv:2511.14759.
33. Xu C, Springenberg JT, Equi M, Amin A, Esmail A, Levine S, Ke L (2026) RL Token: bootstrapping online reinforcement learning with vision-language-action models. arXiv preprint
34. Torne M, Pertsch K, Walke H et al (2026) MEM: multi-scale embodied memory for vision-language-action models. arXiv preprint
35. Mazza L et al (2026) Supervised mixture-of-experts for surgical grasping and retraction. arXiv. <https://doi.org/10.48550/arXiv.2601.21971>. arXiv:2601.21971
36. Belkhale S, Ding T, Xiao T et al (2024) RT-H: action hierarchies using language. arXiv. arXiv:2403.01823.

37. Shi LX, Ichter B, Equi M et al (2025) Hi Robot: open-ended instruction following with hierarchical vision-language-action models. In: International Conference on Machine Learning. PMLR. 267
38. Ortenzi V, Cosgun A, Pardi T, Chan WP, Croft E, Kulić D (2021) Object handovers: a review for robotics. IEEE Trans Robot 37:1855–1873. <https://doi.org/10.1109/TRO.2021.3075365>
39. Younis R, Yamlahi A, Bodenstedt S et al (2024) A surgical activity model of laparoscopic cholecystectomy for cooperation with collaborative robots. Surg Endosc. <https://doi.org/10.1007/s00464-024-10958-w>
40. Khan DZ, Luengo I et al (2022) Automated operative workflow analysis of endoscopic pituitary surgery using machine learning. J Neurosurg. <https://doi.org/10.3171/2021.6.JNS21923>
41. Zhang FX, Al Moubayed N, Shum HPH (2022) Towards graph representation learning based surgical workflow anticipation. In: IEEE-EMBS International Conference on Biomedical and Health Informatics. <https://doi.org/10.1109/BHI56158.2022.9926801>
42. Saeidi H, Opfermann JD, Kam M et al (2022) Autonomous robotic laparoscopic surgery for intestinal anastomosis. Sci Robot 7:eabj2908. <https://doi.org/10.1126/scirobotics.abj2908>
43. Chen WF, Yamamoto T, Dayicioglu D, Shaked A (2025) Robot-assisted lymph node-to-vein anastomosis using a next-generation microsurgical robot. Curr Oncol 32. <https://doi.org/10.3390/curroncol32070377>
44. Kim JW, Chen JT, Hansen P et al (2025) SRT-H: a hierarchical framework for autonomous surgery via language-conditioned imitation learning. Sci Robot. <https://doi.org/10.1126/scirobotics.adt5254>
45. Furnari G, Secchi C, Ferraguti F (2025) Sequence-based imitation learning for surgical robot operations. Artif Intell Surg. <https://doi.org/10.20517/ais.2024.32>

Figures

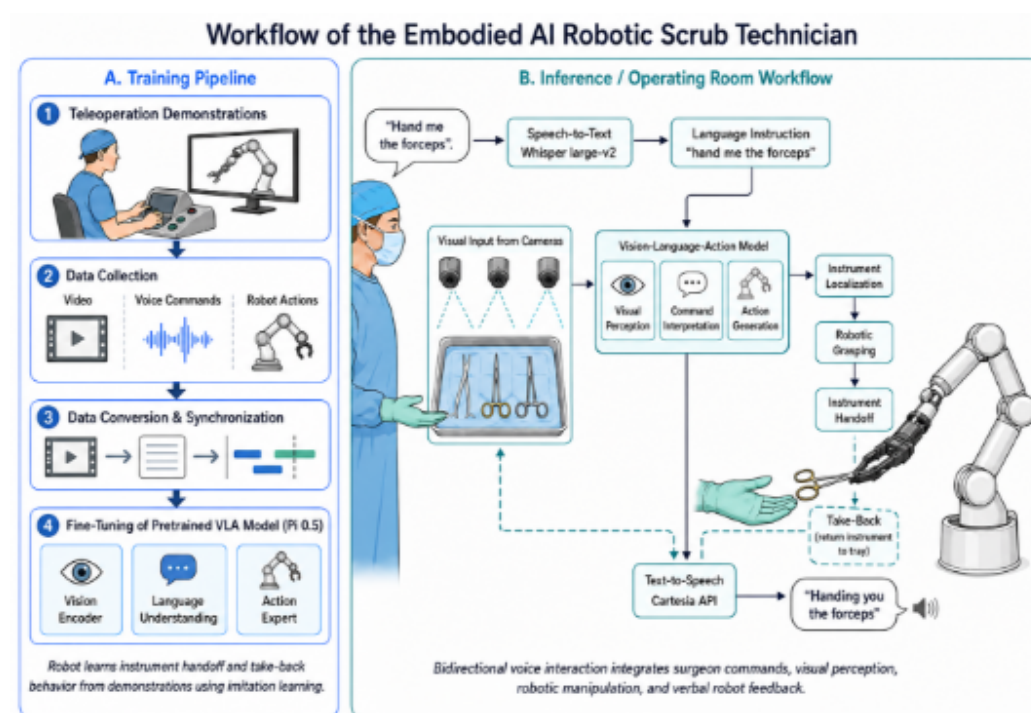


Figure 1

Workflow of the embodied AI robotic scrub technician. (A) Training pipeline: Surgical instrument handoff and take-back demonstrations are collected through human teleoperation. Synchronized video observations, spoken task instructions, and robot actions are recorded, converted, and aligned into a training dataset. These demonstrations are then used to fine-tune the pretrained $\pi 0.5$ VLA model, enabling the system to learn visual perception, language-conditioned task understanding, and robotic action generation. (B) Inference and operating-room workflow: The surgeon issues a natural-language instrument request, which is transcribed by Whisper large-v2 and passed to the VLA model together with real-time visual observations from multiple cameras. The model interprets the command, localizes the requested instrument, and generates the action sequence required for grasping, handoff, or take-back. A Cartesia text-to-speech module provides audible acknowledgment and task feedback, creating a bidirectional surgeon-robot interaction loop

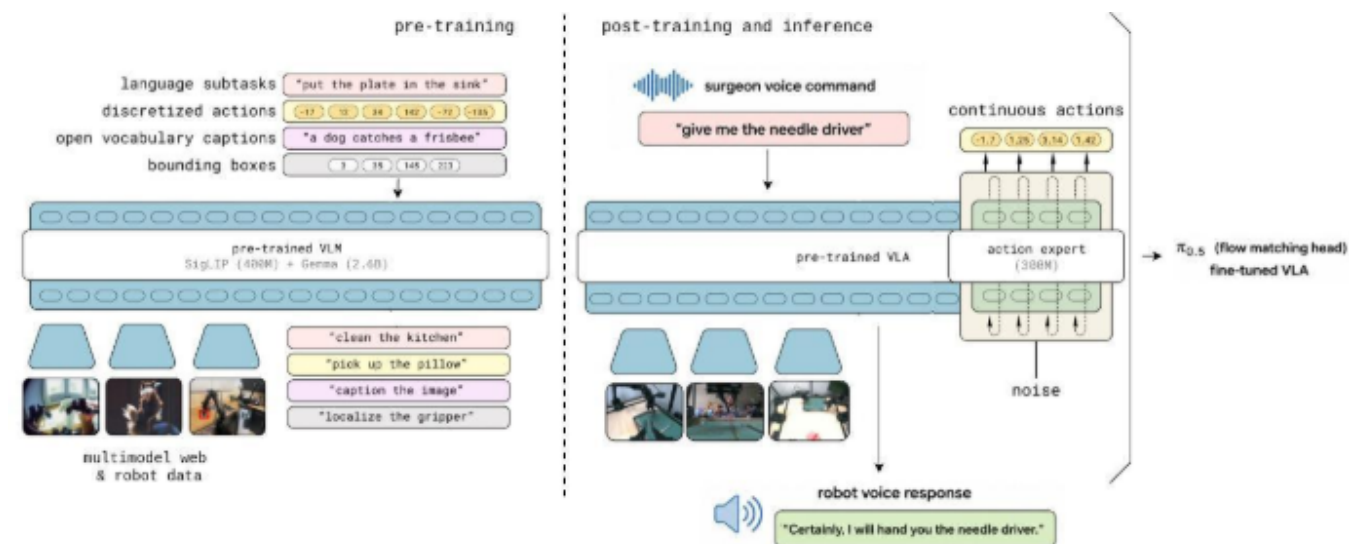


Figure 2

Pretraining, fine-tuning, and inference architecture of the $\pi 0.5$ vision-language-action model. During pretraining, the model learns from large-scale multimodal web and robotic datasets containing language instructions, visual observations, object-localization information, and robot actions. The pretrained vision-language backbone combines a SigLIP visual encoder with a Gemma language model to develop general visual, semantic, and task-level representations. For the present application, $\pi 0.5$ was fine-tuned on task-specific demonstrations of surgical instrument handoff and take-back. During inference, the system receives a surgeon's spoken request, such as "give me the needle driver," together with real-time camera observations. The robot then provides an audible response confirming the requested task. The fine-tuned VLA model interprets the visual scene and language instruction, while the 300-million-parameter continuous action expert generates action chunks that are executed at 50 Hz

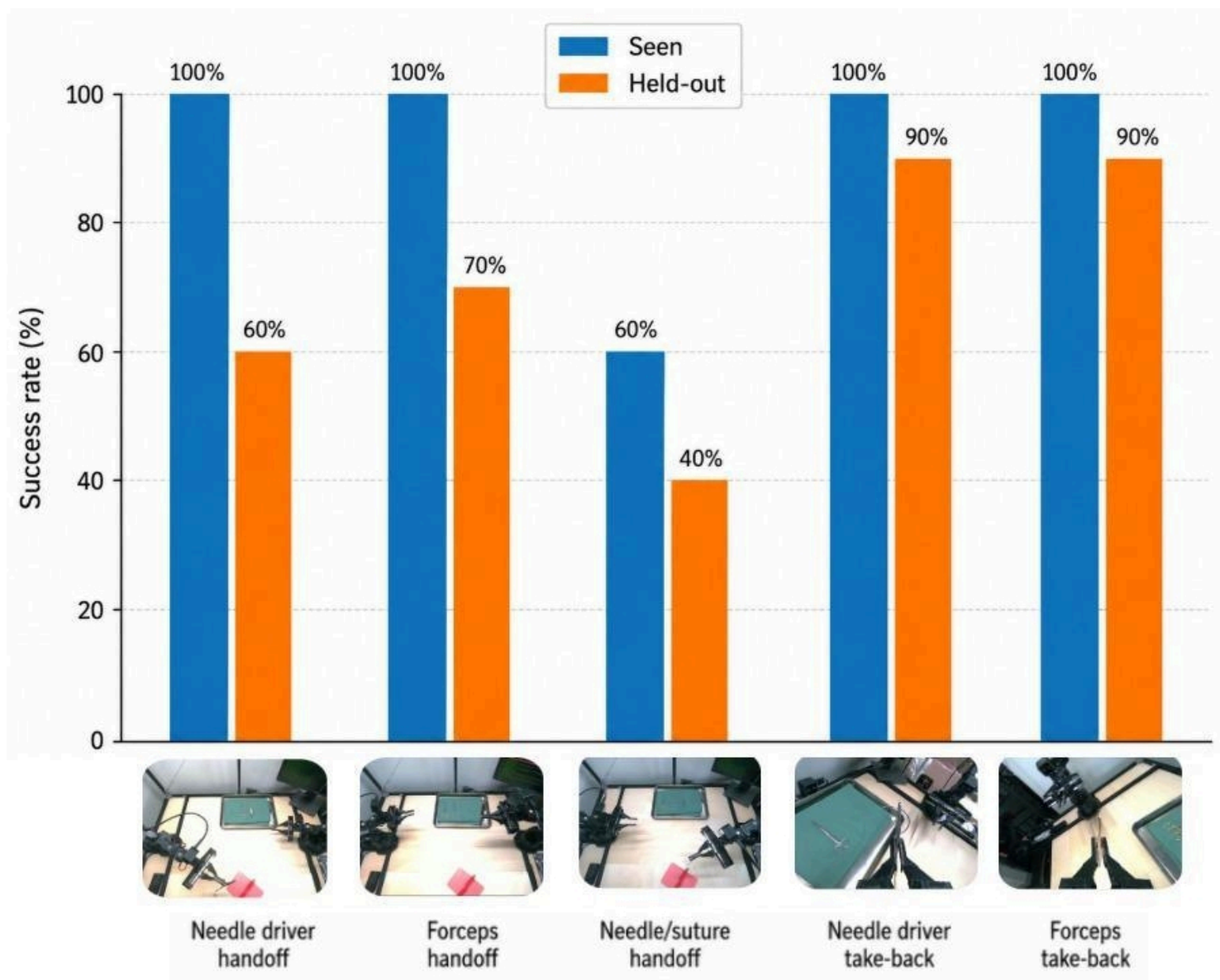


Figure 3

Generalization across seen and held-out surgical tray configurations. Task-specific success rates are shown for tray layouts represented during training and for held-out layouts containing novel instrument positions or orientations. Performance decreased most notably for handoff tasks, whereas take-back remained comparatively robust, achieving 90% success in held-out configurations for both instruments

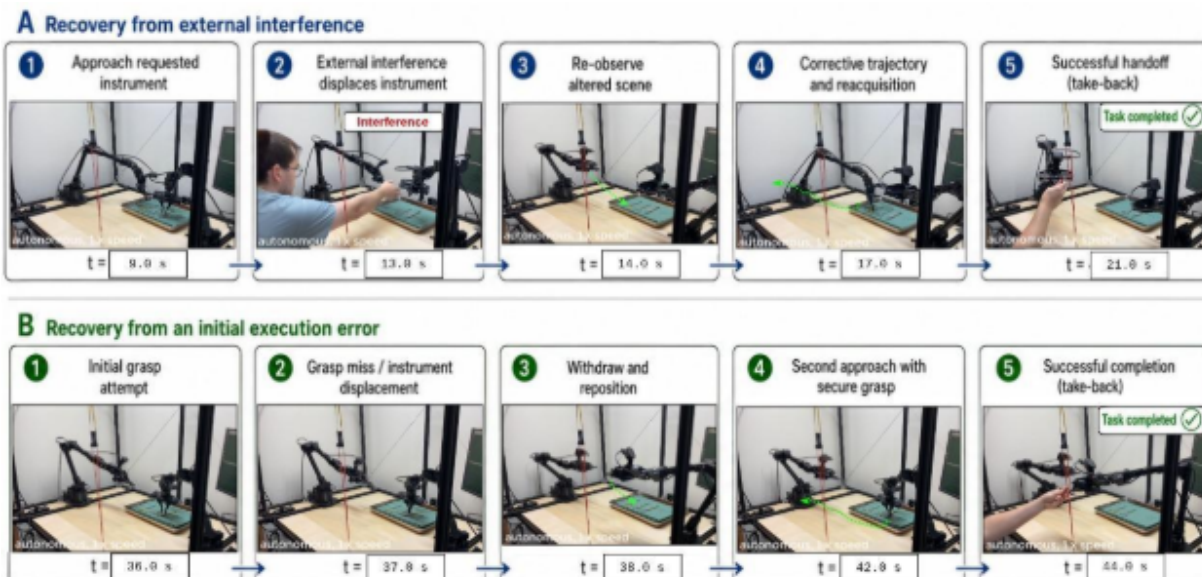


Figure 4

Closed-loop recovery from execution error and external interference. Representative autonomous sequences demonstrate the reactive behavior of the vision-language-action policy. In the first sequence, the physical scene is altered during task execution, invalidating the robot's initial trajectory. The policy incorporates updated visual observations, generates corrective actions, reacquires the target, and completes the requested instrument exchange. In the second sequence, an unsuccessful initial approach is followed by autonomous repositioning, renewed grasp acquisition, and task completion. Both examples were completed without manual task reinitialization or explicit trajectory reprogramming. These qualitative demonstrations illustrate the ability of the policy to remain goal-directed after the physical state deviates from the originally predicted action sequence. Time stamps (t) indicate elapsed time from the start of each corresponding autonomous sequence, as shown in Supplemental Video 3 at <https://surgicalintelligencelabs.com/asia>, Recoveries.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Supplementarymaterial.docx](#)