MEDICAL ROBOTS

SRT-H: A hierarchical framework for autonomous surgery via language-conditioned imitation learning

Ji Woong (Brian) Kim[1]*†, Juo-Tung Chen[1], Pascal Hansen[1]‡, Lucy Xiaoyang Shi[2], Antony Goldenberg[1], Samuel Schmidgall[1], Paul Maria Scheikl[1]§, Anton Deguet[1], Brandon M. White[3], De Ru Tsai[4], Richard Jaepyeong Cha[4,5,6], Jeffrey Jopling[3], Chelsea Finn[2], Axel Krieger[1]*

Research on autonomous surgery has largely focused on simple task automation in controlled environments. However, real-world surgical applications demand dexterous manipulation over extended durations and robust generalization to the inherent variability of human tissue. These challenges remain difficult to address using existing logic-based or conventional end-to-end learning strategies. To address this gap, we propose a hierarchical framework for performing dexterous, long-horizon surgical steps. Our approach uses a high-level policy for task planning and a low-level policy for generating low-level trajectories. The high-level planner plans in language space, generating task-level or corrective instructions that guide the robot through the long-horizon steps and help recover from errors made by the low-level policy. We validated our framework through ex vivo experiments on cholecystectomy, a commonly practiced minimally invasive procedure, and conducted ablation studies to evaluate key components of the system. Our method achieves a 100% success rate across eight different ex vivo gallbladders, operating fully autonomously without human intervention. The hierarchical approach improved the policy's ability to recover from suboptimal states that are inevitable in the highly dynamic environment of realistic surgical applications. This work demonstrates step-level autonomy in a surgical procedure, marking a milestone toward clinical deployment of autonomous surgical systems.

INTRODUCTION

Autonomous surgical systems offer the potential to improve outcomes, reduce costs, and expand access to high-quality care. However, most surgical robots today remain teleoperated because of fundamental challenges. From a vision perspective, surgical scenes are highly complex, involving morphological variation between patients, constant environmental changes during interventions, and visual occlusions such as blood and smoke from cautery tools. Motion planning in this setting is difficult because of the partial observability of organ tissue and their unpredictable dynamics. Additionally, surgical tasks must be performed with high precision and safety, making the development of these systems very challenging.

Prior works have addressed surgical autonomy through various strategies in simulation (*1*–*3*) and real-world settings (*4*–*8*). Various studies explored tabletop tasks, such as peg transfer, needle pickup, and deformable object manipulation, using model-based strategies (*5*, *6*, *9*–*11*), reinforcement learning (*3*, *7*, *12*–*15*), and imitation learning (*16*–*20*). In particular, learning-based methods showed promise in tackling challenging contact-rich manipulation tasks (*21*), such as suture knot tying (*22*), that are otherwise difficult to solve with model-based strategies. Although promising, most learning-based works have been demonstrated in controlled environments and have not been extended to realistic in vivo or ex vivo settings. Therefore, whether these strategies will succeed in the complex and diverse environment of surgery remains uncertain.

On the other hand, there have been notable in vivo autonomous demonstrations, such as needle steering (*8*) and anastomosis tasks (*4*). Although promising, these studies primarily tackled the navigation steps of the procedure, which is much simpler than manipulation, and relied on hand-crafted strategies that were specifically optimized for a single application. In vivo studies demonstrate the promise of robotics being deployed in clinically relevant environments; however, the applied strategies are unlikely to generalize, scale, or address complex manipulation problems that are very common in surgery.

In this work, we aimed to move beyond the scope of prior approaches by addressing several critical and previously unaddressed dimensions of surgical autonomy. First, we focused on contact-rich manipulation tasks that require diverse tool use, including grabbing, clipping, and cutting, which are skills common in real surgical procedures. Second, we conducted this work in a realistic ex vivo setting with substantial variability in tissue appearance, anatomy, and morphology across organs, mirroring the diversity encountered in human surgeries. Third, rather than tackling individual skills, we tackled entire surgical steps that unfold over several minutes and require persistent coordination, decision-making, and adaptation. The combination of these challenges, including contact-rich manipulation, anatomical variation, and long-horizon execution, has been unexplored in prior work and is nontrivial to solve using conventional approaches. Our goal was to show that these challenges can be overcome with a unified design using data-driven methods. Solving this challenge in such a generalizable way is essential for

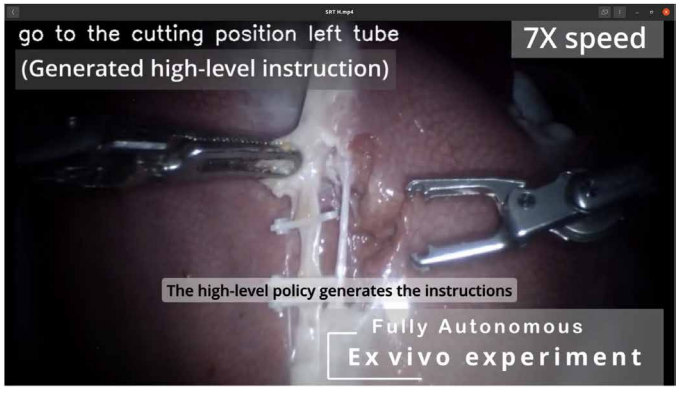
[1]Laboratory for Computational Sensing and Robotics, Johns Hopkins University, Baltimore, MD 21218, USA. [2]Department of Computer Science, Stanford University, Stanford, CA 94305, USA. [3]Department of Surgery, Johns Hopkins University, Baltimore, MD 21218, USA. [4]Optosurgical, Columbia, MD 21046, USA. [5]Sheikh Zayed Institute for Pediatric Surgical Innovation, Children's National Hospital, Washington, DC 20010, USA. [6]Department of Pediatrics, George Washington University School of Medicine and Health Sciences, Washington, DC 20052, USA.
*Corresponding author. Email: jwbkim@stanford.edu (J.W.K.); axel@jhu.edu (A.K.)
†Present address: Department of Computer Science, Stanford University, Stanford, CA 94305, USA.
‡Present address: Division of Intelligent Medical Systems, German Cancer Research Center, Heidelberg, Germany.
§Present address: Amazon Robotics, Berlin, Germany.

progressing toward clinically viable and general-purpose autonomous systems.

Toward this end, we present Hierarchical Surgical Robot Transformer (SRT-H), a framework for autonomous, step-level autonomy in surgery (Movie 1). SRT-H uses a hierarchical architecture composed of a high-level (HL) policy that issues natural language instructions, including task and corrective instructions, and a low-level (LL) policy that executes LL trajectories. This structure allowed us to decompose complex procedures into shorter tasks and enabled the HL policy to correct mistakes made by the LL policy, which naturally arose during long-horizon steps. Furthermore, using language enabled an intuitive interface for intermittent user intervention and fine-tuning. Specifically, users can temporarily override HL decisions with natural language instructions, and these interventions are stored and used for continual learning via a DAgger-style (*23*) loop.

SRT-H was built on a transformer-based architecture and trained end-to-end via imitation learning, using only red, green, and blue (RGB) images paired with language annotations. It avoids reliance on depth sensors, segmentation modules, or specialized hardware. We evaluated SRT-H on the clipping-and-cutting step of cholecystectomy, a common laparoscopic procedure performed more than 700,000 times annually in the United States (*24*). This step involves identifying the cystic duct and artery, placing clips, and severing them. By disabling the clip-latching mechanism, we enabled collection of hundreds of demonstrations from a single porcine tissue, making large-scale data collection feasible. In contrast, other steps like dissection are destructive and yield only one demonstration per specimen, motivating our focus on clipping and cutting steps of cholecystectomy.

To train and evaluate our system, we collected 16,000 trajectories (~17 hours of data) across 34 ex vivo porcine gallbladders. We then tested SRT-H on eight unseen gallbladders, and, in each case, the system successfully completed all 17 required tasks autonomously, generalizing across anatomies and self-correcting its mistakes midprocedure. Ablation studies highlight the critical role of both the hierarchical structure and the corrective language interface in enabling timely and effective corrective behaviors. Compared to an expert surgeon, our framework showed comparable performance, although with longer execution time. In summary, SRT-H provides a scalable and adaptable framework for autonomous surgery, with potential to advance toward generalizable autonomy in real-world surgical settings.



Movie 1. A language-guided imitation learning framework for autonomous robotic cholecystectomy surgeries. Using cholecystectomy as a case study, our framework automated key steps in gallbladder removal, focusing on the complex process of clipping and cutting the cystic duct and artery. The system performed 17 tasks fully autonomously, achieving successful results in all eight ex vivo studies without human intervention. Robustness was demonstrated through challenging scenarios and appearance variations, where the model adapted and executed tasks confidently, highlighting its potential for generalizing across surgical settings.

RESULTS

In the following sections, we describe the design and workflow of our autonomous surgery system and then present the experiment results. We first evaluated our system's ability to complete the cholecystectomy procedures using eight unseen ex vivo porcine tissues. The framework's performance was evaluated on the basis of the success rate, total time, and number of self-corrections made (see the "Core experiment results" section for details). We further evaluated SRT-H against ablative variants to show the effect of different design choices on the performance of the framework. We evaluated these variants on the basis of their success rate, total time, and ability to recover from failure states (see the "Comparison with variants" section). The success rate of failure recoveries was evaluated by placing the instruments into failure states and observing whether each variant could recover to complete the procedure successfully. We also independently performed ablative comparisons for the HL policy and quantified each design choice's effect on its performance (see the "HL policy ablation studies" section). Last, we evaluated our framework against an expert surgeon on the basis of the success rate, time to completion, and the smoothness of the trajectories (see the "Comparison with expert surgeon" section).

Experiment design

Figure 1A shows the hardware configuration of our system, which consists of a da Vinci Research Kit (dVRK) Si with wrist cameras mounted near the instrument tips. The stereo endoscope of the dVRK provides a global view of the surgical scene, and the wrist cameras provide a close-up view of interactions between instruments and tissue. Prior works (*22*, *25*) demonstrated that wrist cameras can help with generalizing to different workspace heights and out-of-distribution scenarios because of the more consistent view provided by the wrist cameras. Although the sizes of the wrist cameras used in this study are quite large and perhaps not clinically practical for minimally invasive surgery, their design can be further downsized. In the following, we describe the general workflow of the procedure within cholecystectomy that is automated, its challenges, and the steps for deploying SRT-H. The steps for clipping and cutting the duct and artery are shown in Fig. 1B. The objective of this step is as follows: Three clips were added to the left tubular structure (typically, the duct), and then three clips were added to the right tubular structure (typically the artery). For each tube, the first two clips were placed proximally near the bottom and the third clip distally at the top. Note that the clips prevent any leakage of biological fluids after the gallbladder is removed; in particular, the two clips placed at the base remain in the patient and must, therefore, provide a secure, long-lasting seal. Then, the tube was transected between the second and third clip of each tube, where there is the most gap for the scissors to enter. In general, the duct and artery are in close proximity; therefore, the left grasper must apply tension at the neck of the gallbladder to stretch the tubes apart and make room for the

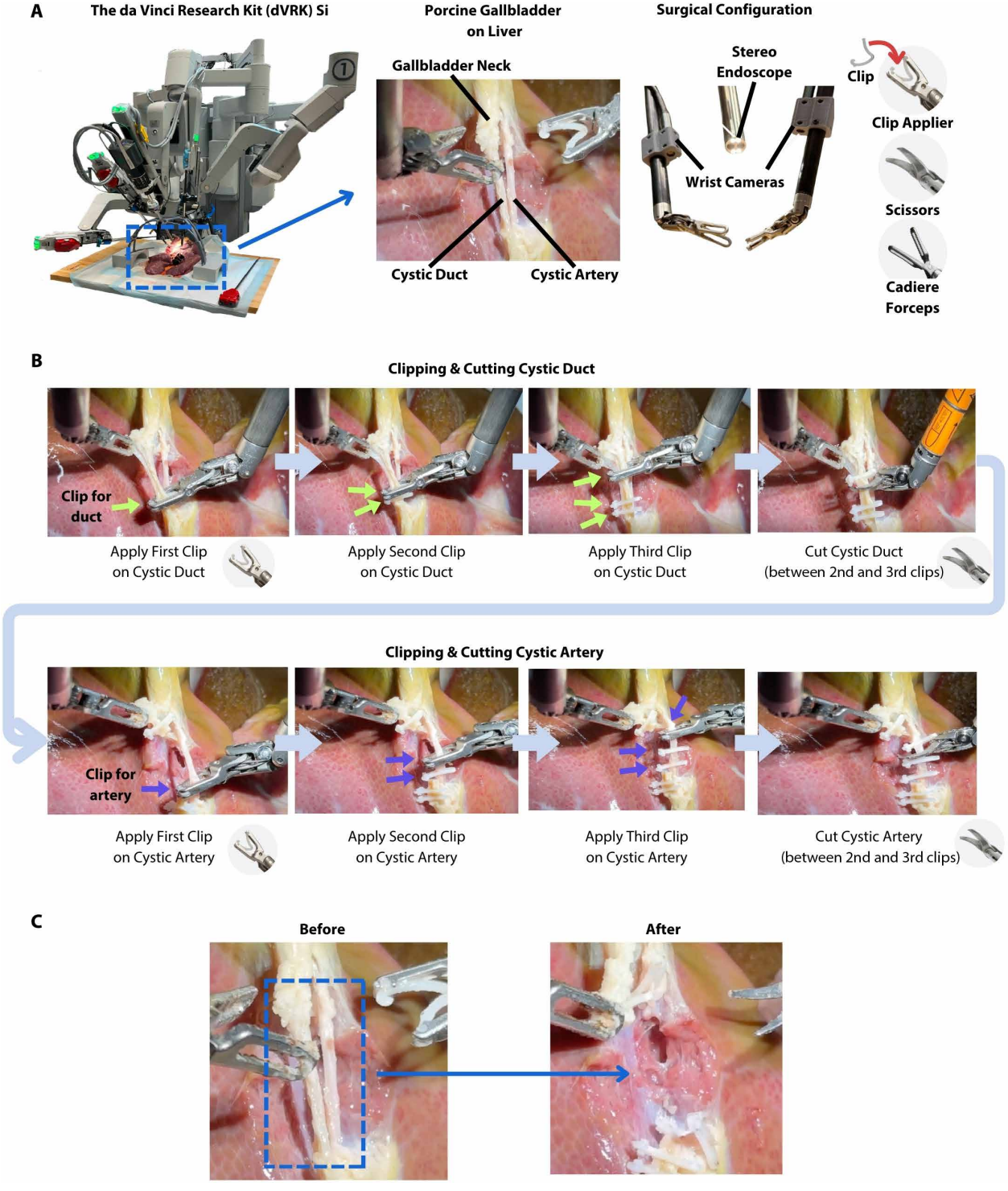


Fig. 1. System and task overview. (**A**) We used the dVRK Si to deploy our policy, which includes an endoscope and two additional wrist cameras mounted for a better view of the interactions between instruments and tissue. (**B**) The autonomous surgical steps include clipping and cutting the gallbladder's artery and duct. (**C**) The before and after pictures illustrate the objective of this procedure; the duct and artery are completely severed, without spilling any of their internal fluids because of the use of clips.

clip applier or the scissors to enter the gap. After each clip was applied, an assistant on standby near the dVRK loaded another clip and also performed tool changes between clip applier and scissors after completion of the relevant steps (filling the role of a surgical nurse).

There are several challenging elements to this procedure. From a visual and anatomical point of view, the appearance of the ducts and arteries varies greatly among patients in terms of their diameter, length, proximity, angle from each other, and amount of connective tissue left on the surface of the tubes, which can make perception challenging (*26*). From a manipulation point of view, precise bimanual coordination of the arms is necessary. In particular, when adding clips to the left tube, the left gripper must grab the neck of the gallbladder head and stretch it to make sufficient space between the duct and the artery, and the clip applier must pry in between the tight space between the tubes to successfully apply the clip (*27*). During this step, the clip applier can overshoot and miss the duct entirely, mistakenly clip the right tube (artery), or apply the clips at a suboptimal location, e.g., applying the third clip too close to the second clip so as to leave no space for the scissors to perform the cut. Overall, to succeed in these steps, the policy must perceive and track the location of the deformable duct and artery, keep an internal count of how many clips have been applied so far, detect whether sufficient stretch has been applied to make room for prying in the clip applier tool, and apply the clips at an optimal location without damaging the surrounding tissues.

During the autonomous trials when SRT-H was deployed, the operator clicked a button on the graphical user interface (GUI) to initiate the system. After the system autonomously applied each clip, the system automatically paused on its own and waited for the operator to load another clip. The operator then loaded another clip, and the procedure resumed. This interaction was repeated for all six clips that were applied to the duct and artery. Between the clip-applying steps, when a scissor was required, similar steps were carried out; the robot autonomously requested for a tool change, and the operator resumed the procedure after making the tool change.

The architectural details of SRT-H are shown in Fig. 2. Briefly, SRT-H was implemented as two transformer decoders; one is part of the HL policy and the other of the LL policy. The HL policy took in a history of endoscope images as input and generated three outputs: the task instruction, corrective instructions, and correction flag (Boolean). Either the task or corrective instruction was provided as input to the LL policy, with the correction flag serving as a binary switch that determined which instruction was sent to the LL policy. The LL policy then took the given instruction, along with the current observations of the surgical scene, to generate a hybrid-relative trajectory (*22*), the action representation optimized for training on dVRK robots.

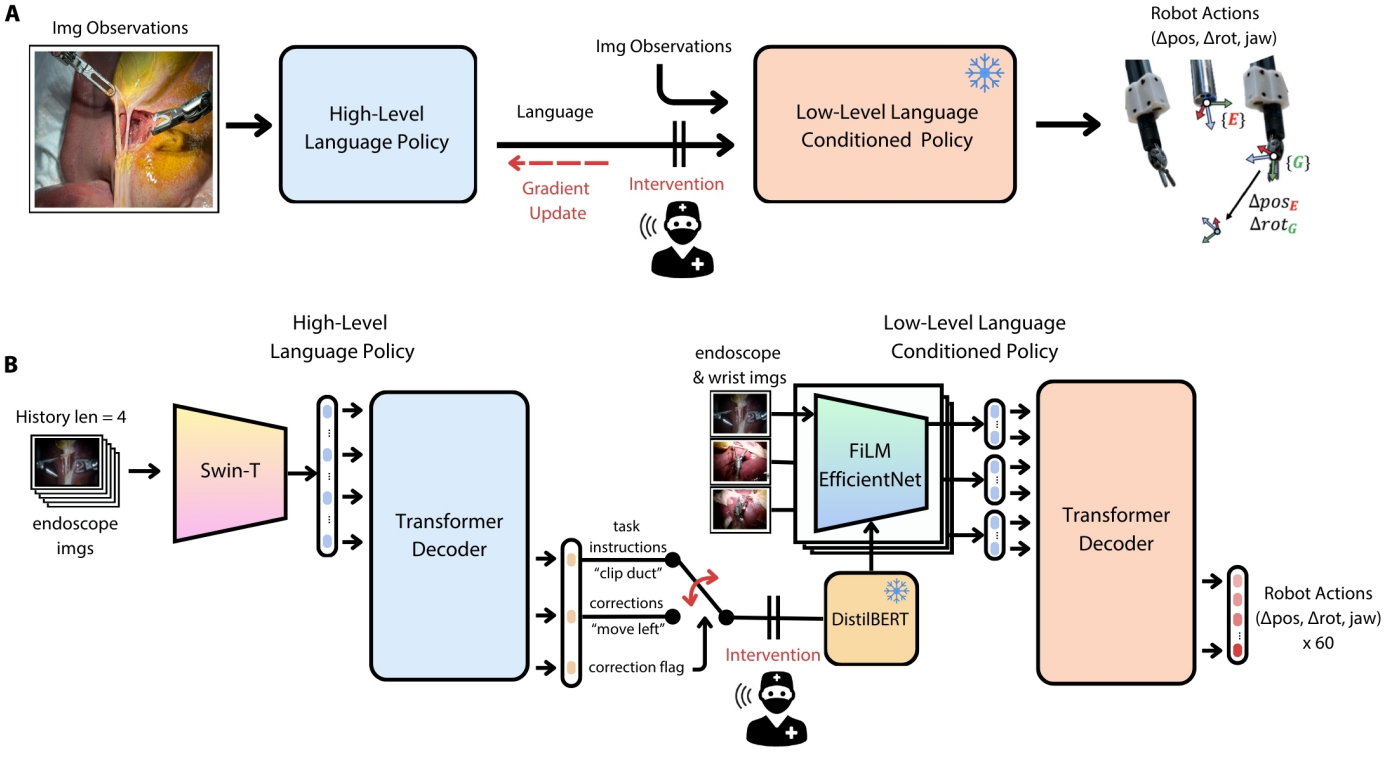


Fig. 2. Model overview and architecture. (**A**) The architecture of our framework consists of an HL policy that generates language instructions given the image observations and an LL policy that conditions on the language instructions and image observations to generate robot motions in Cartesian space. (**B**) On a more granular level, the HL policy consists of a Swin-T model to encode the visual observations into tokens that are processed by a transformer decoder to generate language instructions. The language instructions are processed by a pretrained and frozen DistilBERT model to generate language embeddings. The image observations are passed to an EfficientNet that conditions on the language embeddings through FiLM layers. The combined embeddings are passed to a transformer decoder to generate a sequence of actions that are encoded in delta position and orientation values. Img/imgs, images; History len, history length.

Core experiment results

For the core experiments, SRT-H was evaluated on eight different unseen gallbladders. Table 1 shows the result of each experiment including the success rate, total duration, and number of self-corrections made. We observed that SRT-H was able to complete all of the procedures successfully without any human interventions and, on average, completed the procedure within 317 s or 5 min and 17 s. This duration excludes the time of reloading the clips and making tool changes performed by the operator. Furthermore, when failure states were encountered, SRT-H was able to correct its own mistakes and complete the procedure successfully. On average, self-corrections were made approximately six times throughout the entire procedure. We provide additional information about the individual self-corrections in fig. S7. Figure 3 shows the placement of each clip before the artery and duct were cut in more detail. The clips fully encompassed the ducts and arteries, maintained close spacing between the bottom two clips on each tube, and left sufficient spacing between the second and third clip on each tube for easy access for the scissors to make the cut. Overall, across diverse tissues, SRT-H demonstrated consistent capability in recognizing the relevant tissue structures, maintaining a reasonable pace, and recovering itself from its own failures to complete all cases successfully. In general, the upper-most clips were placed close to the gallbladder infundibulum, but at times, they may not have been positioned at the highest point. To alleviate such placement issues in the future, we may collect additional data where clips are positioned as far up as possible, allowing SRT-H to more accurately replicate ideal placement. Similarly, in some cases, the clips were placed quite low in the surgical field because of suboptimal demonstrations, and these issues could similarly be improved by collecting better demonstrations.

Additionally, we encountered a non–safety-critical robot failure in one of the eight experiments that was not related to SRT-H, when the scissors broke and had to be replaced before continuing. In addition, the dVRK system had to be reinitialized three times during manual tool changes, also unrelated to SRT-H. These issues arose because we were using the very first dVRK Si still undergoing development, and the hardware system was not yet perfected. These hardware-related issues have since been resolved.

Comparison with variants

We further evaluated SRT-H against several variants, including SRT-H trained with task instructions only (no corrective instructions), SRT-H trained without wrist cameras, SRT-H's HL policy trained without additional DAgger data (collected using expert language corrections during prior policy rollouts), and end-to-end architecture with only the LL policy. For all of the tests, each variant was evaluated on the basis of its success rate and total duration. To ensure a fair comparison, all variants were evaluated using the same gallbladders and starting positions, with a 90-s maximum time limit set for completing each task.

The full results of these evaluations are shown in Fig. 4. In terms of success rates (Fig. 4A), SRT-H scored the highest (100%) in both normal and recovery scenarios (Fig. 5). SRT-H using task instructions was a close second, given that it also scored highest under normal scenarios (100%); however, because of the lack of corrective vocabulary, its performance in recovery scenarios was lower (66.7%). Omitting wrist cameras also reduced the success rates in both scenarios (77.8 and 50%, respectively), highlighting their importance in highly diverse ex vivo scenarios beyond table-top settings. SRT-H without HL fine-tuning resulted in diminished performance (77.8 and 75%, respectively), demonstrating the importance of using a competent HL policy and the efficacy of fine-tuning the HL policy. The end-to-end policy variant scored the lowest in both scenarios (33.3%).

In terms of total duration (Fig. 4C), SRT-H performed the fastest on average for both normal and recovery scenarios. The other variants required more time because they made mistakes, which they could not recover from, or fell into repeating loops of retry behaviors. In general, however, the rate of motion for all variants was similar, and their differences were dictated by how competent the policy was in recovery behaviors.

We also evaluated how the amount of data affected policy performance. As shown in Fig. 4B, we evaluated SRT-H with 33.3, 66.6, and 100% of the entire dataset as training data. These variants scored success rates of 66.7, 77.8, and 100%, respectively. This evaluation indicates that, beyond the design of the architecture, the amount of data plays a critical role in policy performance.

HL policy ablation studies

For the HL policy, several design choices were made to address perception challenges arising from differences in gallbladder color, texture, and anatomy. First, in addition to the full view, we incorporated a center-cropped version of the most critical operating area as input. The center-crop size is 432 pixels by 480 pixels, and the cropping

Table 1. Core experiment metrics. Procedures were performed on $n = 8$ ex vivo porcine gallbladder (Gallbl.) tissues, showing the success rates, total duration, and number of self-corrections over all tasks of the procedure.

	Success rate (%)	Duration (s)	# Self-corrections
Gallbl. 1	100	290	2
Gallbl. 2	100	315	8
Gallbl. 3	100	304	14
Gallbl. 4	100	300	3
Gallbl. 5	100	396	6
Gallbl. 6	100	318	12
Gallbl. 7	100	274	1
Gallbl. 8	100	337	5
Average	100	317	6

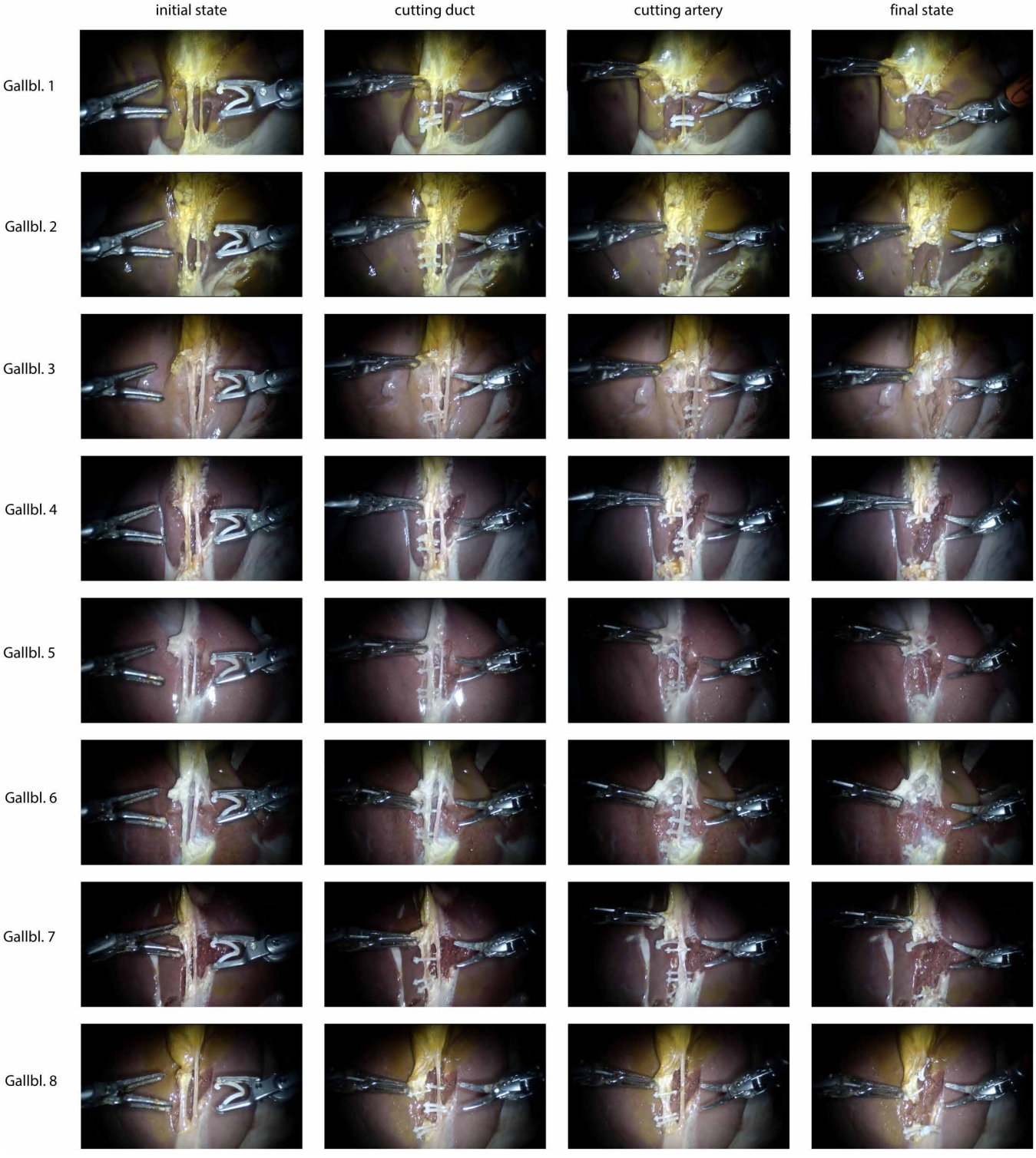


Fig. 3. Core experiment sequences. Images of the initial and final states, as well as observations of the clip positions for the duct and artery before the cut was made for all eight gallbladders. The clips were sufficiently secured around the ducts and arteries, and sufficient space between the second and third clips of each tube was left for the scissors to make the cuts. The individual gallbladders (Gallbl.) varied noticeably in color, texture, and anatomy.

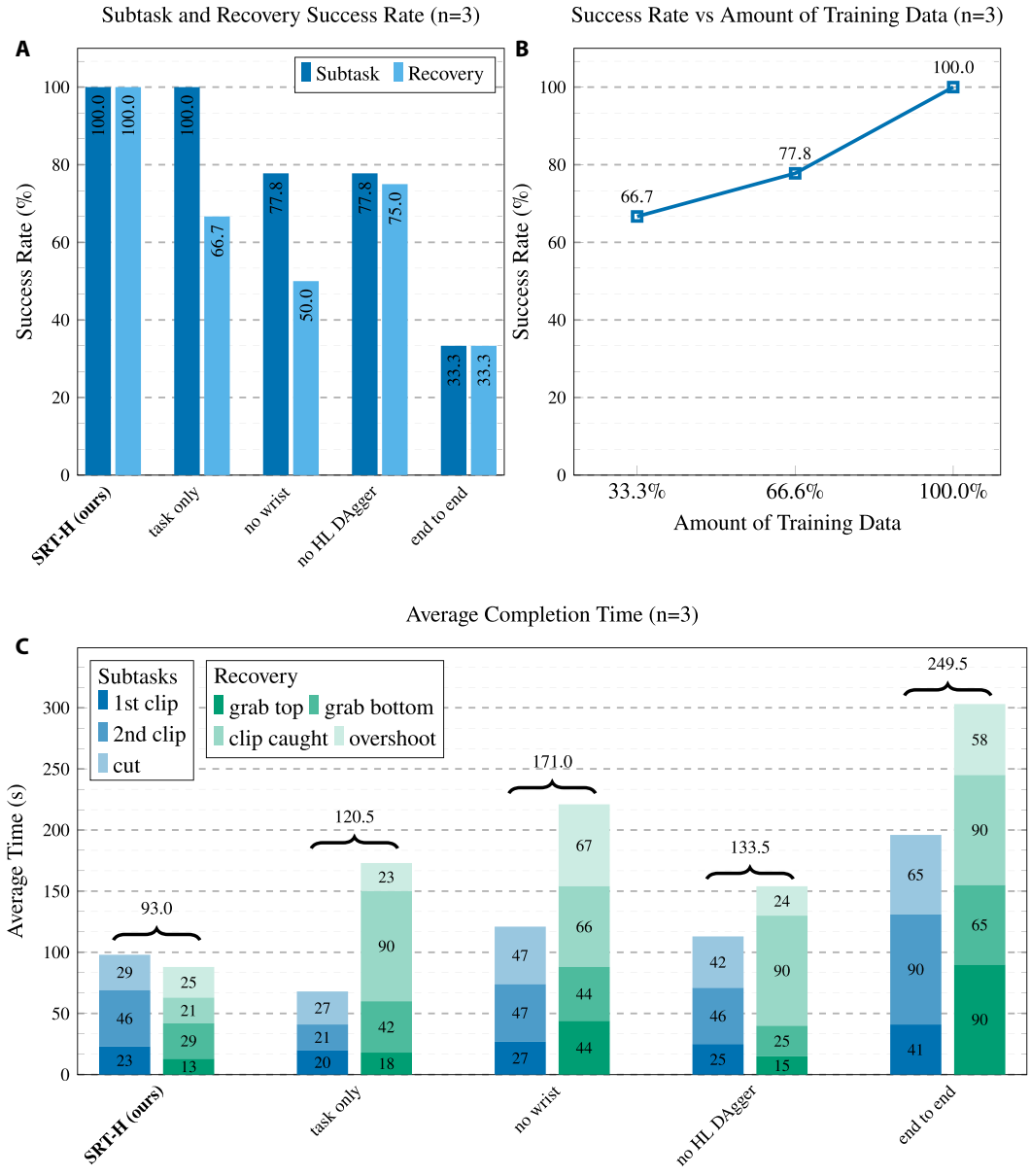


Fig. 4. Comparisons against variants. (**A**) We compared the success rate of our method, SRT-H, against various variants on subtasks and recovery scenarios for $n = 3$ gallbladders. These three gallbladders were independent of the eight gallbladders used in the experiment. (**B**) The success rates of SRT-H for $n = 3$ gallbladders with respect to the amount of training data used. (**C**) The average completion times over $n = 3$ gallbladders for SRT-H and ablative variants.

location is always fixed on the original image. This allows the model to focus on the most relevant information in the surgical field by providing this area at a higher resolution compared with the full view. Second, we modified the cross-entropy (CE)–based loss function by scaling it with the L_1 distance between the predicted and reference task instructions. This adjustment was intended to improve the policy's ability to distinguish between tasks that are temporally distant but visually similar. Third, to mitigate the effect of occlusions during surgery, we included a history of four past image frames, each spaced 1 s apart, along with the current frame. This temporal context allowed the HL policy to retain crucial temporal information, ensuring robust performance even when important details were temporarily obscured. We conducted an ablation study to determine the contribution of each design choice by systematically omitting each one during model training. Performance was evaluated on the basis of both accuracy and F1 score for three classification tasks: predicting task instructions, corrective instructions, and identifying recovery modes.

Results show that our HL policy achieved an accuracy and F1 score of ~97% for task instruction predictions. Removing the center crop input or using only the CE loss for task instructions resulted in a decrease in accuracy and F1 score of around 2 to 2.5%. Omitting the observation history led to an even more substantial drop in performance, exceeding 10% for the task instruction predictions and a similar decline for the corrective instruction and recovery mode prediction. In the other two prediction tasks, our model also

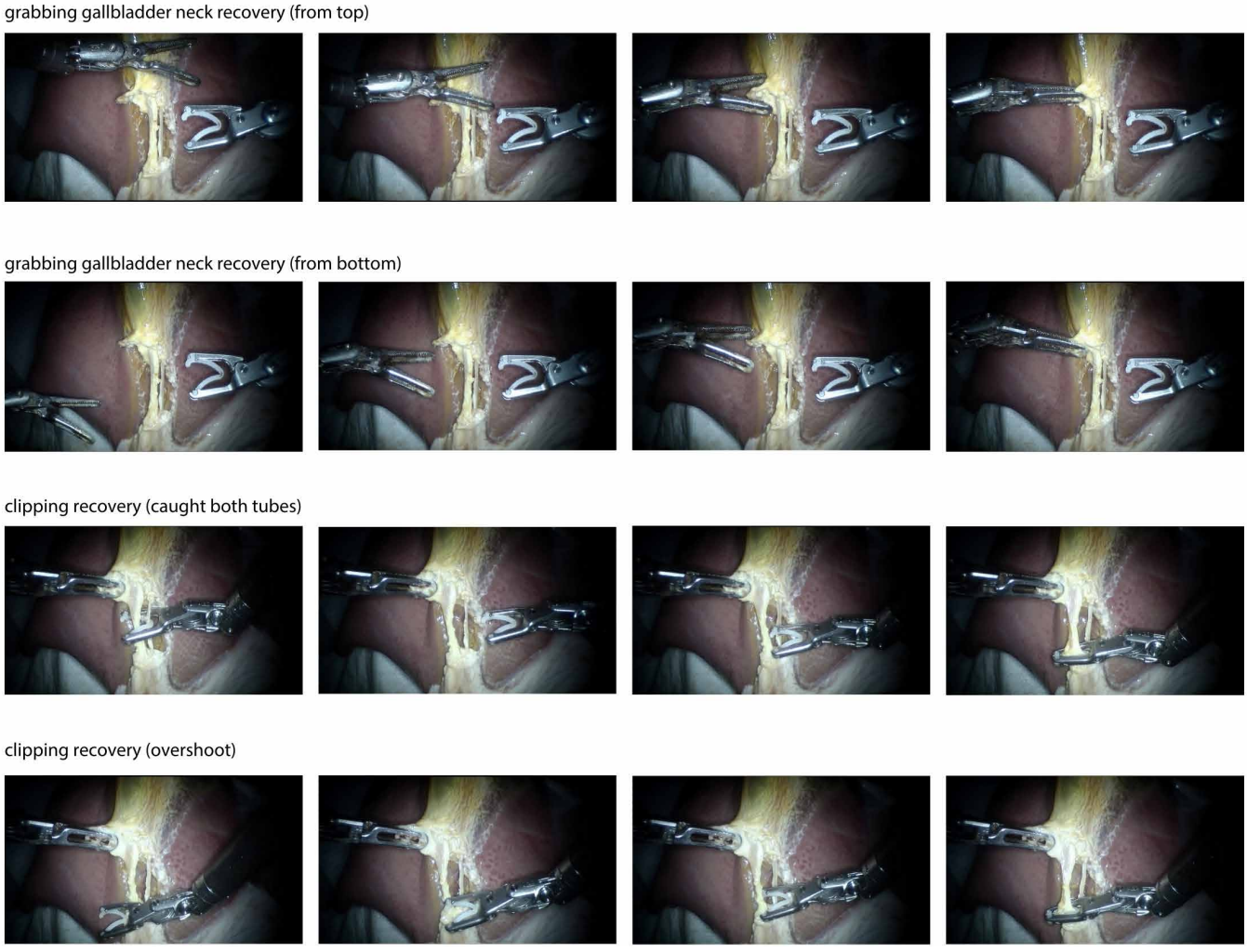


Fig. 5. Recovering from failure states. We manually placed the instruments into failure states to evaluate SRT-H's ability to recover from disadvantageous states of the environment. Each row illustrates a specific failure state and a sequence of images that show how SRT-H recovers from it.

outperformed the variation that excluded the center crop input and the variant that only used the CE loss without scaling. Although the margin for recovery mode predictions was smaller, with an improvement of around 0.5 to 1%, the increase in corrective instruction predictions performance was more pronounced. This is particularly evident in the F1 score, highlighting the HL policy's ability to issue language corrections more consistently, achieving a 2 to 2.5% improvement. Overall, the HL policy achieved ~95% accuracy in identifying recovery modes and around 70% accuracy in predicting corrective instructions, of 18 possible motion classes (see the "Corrective language instructions" section in Supplementary Materials and Methods). We provide additional information on these evaluations in table S2.

As a further study, we applied GPT-4o, a state-of-the-art general-purpose vision-language model, as the HL policy for surgical task planning. GPT-4o was provided with the current endoscope image and all task instructions it could issue to guide the robot (see fig. S7). GPT-4o shows shortcomings in domain-specific understanding in issuing the correct task instruction. For example, it initially omitted the crucial step of "grabbing gallbladder" and prematurely initiated the action "clipping first clip left tube." Additionally, GPT-4o incorrectly prompted the go-back from clipping/cutting instructions before completing the task. Thus, GPT-4o would not be able to guide the LL policy through a full cholecystectomy procedure, because it was unable to issue the correct task instructions.

Comparison with expert surgeon

We performed a preliminary comparison between SRT-H and an expert surgeon. Given the same gallbladder, both performed several tasks, including adding the first and third clip to the artery and cutting it. Each round, SRT-H was deployed first, and the surgeon was asked to repeat the same task. For adding the clips, modified clips with a disabled latching mechanism were used. For cutting, right before the policy attempted to close its grippers to complete the cut, the robot was stopped to avoid permanent damage to the tissue. The surgeon had experience in performing both robotic and manual cholecystectomy. The surgeon did not have prior experience with the dVRK system but was given sufficient time to become familiar with using the system. Note that the participating surgeon study did not contribute to the training data.

The results are shown in Fig. 6, which shows qualitative comparisons of the trajectories from the endoscope view and also in Cartesian space. We quantitatively report the mean jerk, trajectory length, and total duration during the tasks for both the surgeon and SRT-H. In general, we regard the better performer as the one that performed with the least mean jerk, trajectory length, and total duration.

Our results show that the surgeon completed all tasks faster than SRT-H. However, we observed that SRT-H navigated with shorter trajectory length and less mean jerk compared with the surgeon; therefore, SRT-H generates smoother and shorter trajectories. However, the surgeon was much faster in executing all of the steps. As a qualitative comparison, the two-dimensional (2D) projections of the trajectories show that SRT-H and the surgeon performed the procedure in a similar manner, based on the overall shape and appearance of the trajectories. In general, despite these promising findings, we avoid making strong claims that SRT-H outperforms the surgeon. We also lacked a sufficient number of gallbladders for a more in-depth comparison. A more detailed analysis may be addressed in further extension of this work. Our goal is to give an initial intuition of how our framework's performance compares with that of an experienced surgeon.

DISCUSSION

In this work, we introduce SRT-H, a scalable framework for achieving step-level autonomy in robotic surgery. In comparison with prior work, which primarily focused on assistive tools (*28*, *29*) and task-level autonomy (*4*, *8*, *30*), our research takes a step forward by moving toward autonomy at the step level. The results of our study demonstrate the effectiveness of SRT-H in automating the clipping and cutting procedure of a cholecystectomy intervention. Ablative studies show the effectiveness of our hierarchical design, which incorporates HL and LL policies. This design also demonstrates the ability to generalize across unseen ex vivo tissues and self-correct errors in real time. We demonstrated our approach across eight

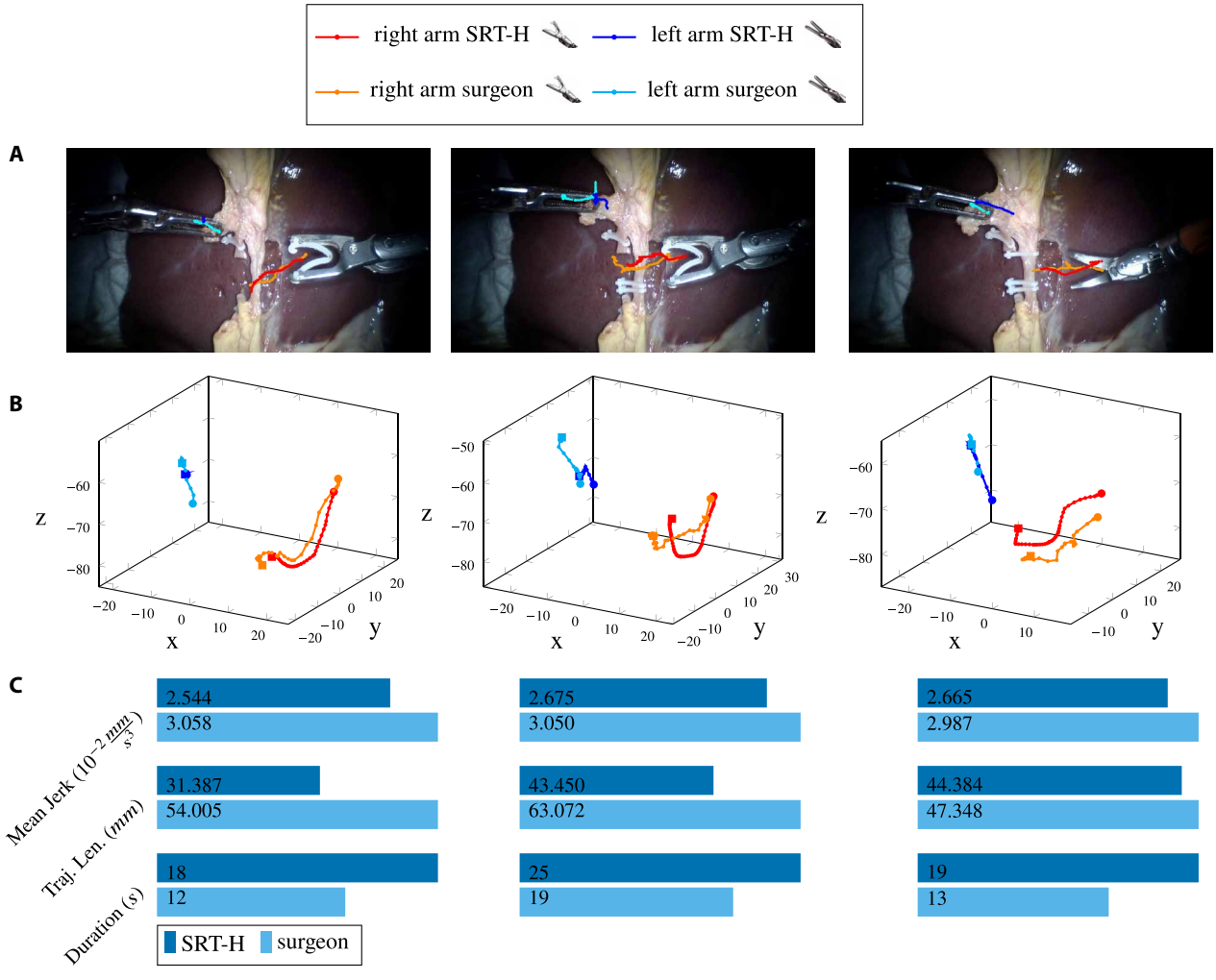


Fig. 6. Qualitative motion comparison between SRT-H and a surgeon. We evaluated SRT-H against a human surgeon on the same gallbladder for the subtasks of applying the first and third clip to the artery, as well as cutting the artery. (**A**) 2D projection and (**B**) 3D plot of instrument paths for SRT-H (dark blue and red) and human surgeon (light blue and orange) as absolute positions in millimeters. (**C**) Quantitative comparisons between SRT-H and human surgeon based on the total duration of task execution (duration, in seconds) and trajectory length (Traj. Len.) of the instruments (in millimeters), as well as the mean jerk (in 10^{-2} mm s^{-3}) calculated over the instrument paths.

gallbladders, achieving a 100% operation success rate for the ligation and division step.

Prior work

Levels of autonomy

The level of autonomy (LoA) in medical robots is categorized across distinct levels (*31*), ranging from pure teleoperation to full autonomy. LoA 0 represents no autonomy, where the robot functions purely as a tool controlled by a human operator. LoA I is defined by robot assistance, where the robot provides continuous control support, such as mechanical guidance or virtual constraints, but the human remains in full control. LoA II refers to task autonomy, where robots autonomously perform specific tasks, like running sutures, initiated by human input via discrete control commands. LoA III, conditional autonomy, allows the system to generate task strategies autonomously but requires the human operator to select among them or approve an autonomously selected strategy. Systems at LoA IV, classified as high autonomy, can make medical decisions independently but still require supervision by a qualified doctor. Last, LoA V represents full autonomy, where the robot is capable of performing an entire procedure without any human intervention.

Examples of high LoAs

Higher LoAs (IV) have been achieved by a few systems. One such system is the CyberKnife (*32*), which autonomously performs radiosurgery for brain and spine tumors under human supervision. This system operates in highly structured environments, using noninvasive techniques where tissues are rigid and stable, reducing the complexity of automation. Another LoA IV system is the Veebot (*33*), which autonomously performs blood sampling by identifying and selecting suitable veins. These systems demonstrate progress in autonomous surgery; however, they operate under controlled conditions, and the gap between these systems toward achieving full autonomy in dynamic, soft-tissue environments remains considerable.

Our present SRT-H work falls in LoA IV, given that it is capable of reliable and autonomous execution while self-correcting its mistakes; these self-corrective instructions are generated by itself and not issued by the user of the system. However, our system is not failure-proof to out-of-distribution scenarios; therefore, the surgeon should always oversee its operation.

Additionally, we briefly mention further evolved definitions of LoA, which include level of environmental complexity (LoEC) and level of task complexity (LoTC) (*34*). According to these metrics, our work falls in LoEC IV and LoTC IV. Our work can be categorized as LoEC IV because soft and realistic tissues are involved, although without topological motion (e.g., breathing), which is the further requirement needed to reach LoEC V. In terms of LoTC, our work falls into category IV because we consider advanced surgical tasks that require spatial understanding of the scene, but the model lacks clinical and anatomical knowledge, which is the further requirement to reach LoTC V.

We also draw a direct comparison to a highly relevant prior work involving autonomous bowel anastomosis (*4*). Although anastomosis may seem like a more technically demanding task, our work demonstrates a greater step forward in comparison. More specifically, in this earlier work, the procedure took place under highly controlled conditions: The bowels were scaffolded on a fixture, fluorescent markers were used for tracking, and a specialized needle-throwing device simplified suturing to a basic reach task. Even with these advantages, the system occasionally made errors that required manual surgeon intervention. Moreover, the prior approach relied on a hand-crafted state machine with model-based planning, which lacks expressivity. By contrast, our present work requires no special fixtures, tracking markers, or specialized surgical devices. Instead, it uses imitation learning to acquire more sophisticated and adaptable manipulation skills, which are difficult to capture with purely hand-crafted methods. For example, our system can delicately maneuver through the narrow space between the duct and artery, place clips at appropriate locations, and execute precise cuts without harming nearby tissue, all of which would be challenging to program explicitly. Crucially, the model can self-correct during the procedure, reducing the need for human intervention at test time. Furthermore, our method is expressive and scalable: By gathering demonstration data from additional procedures, we can potentially apply the same approach to a wide variety of surgical tasks, including anastomosis.

Robot transformers

Outside of surgery, advancements in robotics have led to the development of general-purpose task-solving models (*35*–*39*). These models are trained by imitation on extensive real-world robotics datasets, processing images from robot cameras, and following natural language task descriptions to generate robotic actions. The resulting controllers exhibit the ability to adapt to novel situations and demonstrate task-solving capabilities that extend well beyond the scope of their training data (*37*). These models interpret commands that were not part of the training data and exhibit the ability to reason on the basis of user instructions, such as which object to use as an improvised hammer (a rock) or finding a drink that is best for someone who is tired (an energy drink).

Limitations

From ex vivo to in vivo

One important area for further research is translating our system from ex vivo experiments to in vivo clinical environments. Translating from ex vivo to in vivo brings several challenges, such as operating in the surgical site, addressing bleeding and tissue motion, and fitting the wrist cameras through laparoscopic ports. Given that our approach is robot agnostic and only depends on the relative position of the robot end effectors, surgical access and operation do not present many challenges. Because our approach operates through visual guidance (instead of a model-based approach) and has the ability to self-correct, we believe that it can adapt to motion and blood if it is incorporated as part of the training data or potentially zero shot (see fig. S5 for reference). However, further studies are required to confirm this. Additionally, although the current wrist-camera configuration in our work would likely not fit into laparoscopic ports, modern cameras provide strong imaging quality with sub-millimeter form factors (*40*, *41*) and can be easily integrated into surgical tools with minimal size increase of ports. Another concern with the use of wrist cameras may be potential occlusions because of fog and blood on the camera lenses. A potential solution to deal with these issues is to translate the strategies used for endoscopic cameras to wrist cameras. For instance, antifogging solutions like Fred (*42*) may be used for fogging scenarios. For blood wiping, there are commercial solutions like ClickClean (*43*) or ClearCam (*44*), which physically remove any occlusions on the lens without removing the surgical tools. Furthermore, normalizing the usage of wrist cameras in the operating room may take time, considering that they are devices not widely available in the market.

Making SRT-H safer

A further extension of this work may focus on expanding the system's capabilities to cover a broader range of surgical procedures. The presented SRT-H framework supports the ability to learn across multiple surgical procedures using the same model parameters, to which diverse learning is believed to improve performance on individual tasks (*21*, *36*, *37*). Risk management remains a crucial aspect of surgical robotics. Further research could incorporate conservative Q-learning (CQL) (*45*) and conformal prediction (*46*, *47*) into the SRT-H system to address uncertainty during surgery. CQL would help prevent overestimation of the SRT-H's actions in unfamiliar situations, and conformal prediction would provide real-time feedback on the system's confidence levels. Safety switching with robotic systems can be performed with on-site surgeons or through teleoperation, much like the proposed safety protocols used in autonomous driving systems (*48*, *49*). Additionally, with enhanced perception, it may be possible to simulate robot behaviors in simulation and refine plans before executing in the real world for greater safety (*50*). Last, although we demonstrate this approach primarily through full autonomy without supervision, our approach also supports real-time language interventions from expert surgeons, making it practical for potential integration into hospitals as a tool for surgeons to reduce fatigue on simple procedures or for areas with no access to trained surgeons. Intervention could be requested by the system on the basis of uncertainty calculations and could be performed by a remote operator (*46*).

MATERIALS AND METHODS

Data collection

Training data were collected by two experienced human demonstrators on the dVRK system. Dataset D_1, collected by the first demonstrator, contains data from 31 different gallbladders. The second demonstrator collected data for three additional gallbladders, which are denoted as dataset D_2. All gallbladder organs were sourced from Animal Technologies Inc. (Tyler, TX, USA). Note that both data collectors were nonclinical research assistants, trained by a surgical resident with extensive experience performing cholecystectomies. The first assistant was the primary data collector and contributed the most to the dataset. By the time the second data collector joined the project, most of the necessary data had been collected; therefore, the contributed dataset was much smaller. We define $D = D_1 \cup D_2$ as the union of both datasets. The visual data include video streams from the dVRK stereo endoscope, which has a resolution of 960 pixels by 540 pixels, and two wrist cameras, each with a resolution of 640 pixels by 480 pixels, mounted on the instruments of the surgical robot's left and right arms. Both video and kinematic streams were recorded at 30 frames per second.

Before collecting data, a demonstrator performed blunt dissection with Maryland forceps on a given gallbladder to reach the critical view of safety, where the cystic duct and artery are clearly identifiable. Certain gallbladders with abnormal tissue structures were not used, including the ones where the artery crossed over the duct and where the artery branched (see fig. S6 for reference). About 10% of gallbladders were excluded because of these anatomical anomalies. Although the model can handle such variations if sufficient demonstration data are available, their rarity made it difficult to collect data at scale. Addressing these edge cases is beyond the scope of this work and is left for future investigation. To simulate an accurate setup for the surgery, an expert surgeon recommended cholecystectomy port locations using a plastic abdominal dome. These ports were then isolated and modeled in computer-aided design to create an open structure that held the port locations for each arm of the surgical robot, as shown in Fig. 2A. This way, the dissection area remained open rather than concealed, which is ideal for frequent wrist camera mounting, clip reloading, and tool switching. This open setup may raise concerns that ambient lighting may affect the lighting conditions. However, we found that its effect on the endoscopic and wrist cameras' image quality was negligible.

The clipping and cutting portions of cholecystectomy include 17 tasks in total. These include grabbing the gallbladder (1), adding six clips ($2 \times 6 = 12$), and cutting twice for the duct and artery ($2 \times 2 = 4$), summing to 17 ($1 + 12 + 4 = 17$). Note that the tasks for adding the clips and cutting involve two tasks: the motion for adding the clip or cutting and the retraction.

To acquire multiple trials from a single gallbladder, we used a few tricks. For clipping motions, we used clips with the latching mechanism disabled. This allowed us to perform clipping motions repetitively without actually locking it to the duct or the artery. For the cutting motions, we performed the motion of placing the scissors, but we did not close the scissors at the last step. During postprocessing, we extended the kinematics data to simulate a cutting motion. Using this strategy, it was possible to acquire multiple demonstration data using a single gallbladder with minimal damage. This may raise concerns that simply closing the grippers might not guarantee cutting. In practice, if the cut was not successful, which was very rare in our experiments, then the policy often tried to cut again because it observed that the duct/artery was not cut and remained intact in the image observation. Also, multiple cuts were generally not necessary because the scissors were very sharp. These strategies served to aid with data collection without harming the tissues and did not take away from the generality of the methods.

We used the above logistics to collect many expert demonstrations. Additionally, we further collected samples that show recovery from suboptimal states to augment the dataset. These recovery demonstrations helped the learned policies to recover from their own mistakes.

After training the policies on the base dataset, we additionally collected a DAgger (*51*) dataset D_{corr} as described in (*52*) to improve the base model performance by learning from verbal corrections of common mistakes during policy rollout. The DAgger algorithm iteratively collects data from the policy's own actions and corrects them using expert feedback to refine the policy. Within our DAgger dataset, only the language predictions were corrected; therefore, it is denoted as HL DAgger for the rest of the paper. The language corrections were either issued during the experiment or added during postprocessing. The dataset is summarized in Table 2, providing relevant statistics such as the number of demonstrations, images, and duration for both optimal and recovery demonstrations. These numbers represent the total number of trajectories collected across all gallbladders, encompassing all tasks involved in the clipping and cutting steps of the cholecystectomy procedure.

HL policy

Problem definition

The HL policy, denoted $\pi_{\text{HL}}(p_t, c_t, m_t | o_{t-k:t})$, took as input the current image observation o at time step t, along with k preceding observations from the left camera stream of the dVRK Si endoscope.

Table 2. Dataset summary. Statistics are for the data collected by the two main data collectors (D_1 and D_2) and in the HL DAgger experiments (D_{corr}).

	Data collector 1	Data collector 2	HL DAgger
Number of gallbladders	31	3	15
Optimal demonstrations			
Number	12,304	885	264
Images	1,472,551	127,325	54,638
Time (s)	49,085	4,211	1,821
Recovery demonstrations			
Number	4,904	263	352
Images	704,797	40,297	75,017
Time (s)	23,493	1,343	2,500

As output, the HL policy generated three predictions: the next task p_t (i.e., surgical phase) to be executed by the LL policy; a correction flag c_t indicating whether the robot is in a recovery mode; and a corrective (motion) instruction m_t that specified cardinal actions, such as "move right arm to the right" or "move left arm toward me," that should be executed instead if the robot was in recovery mode. A CE loss was used for all three predicted outputs (see Eq. 1). For the task instruction component, the CE loss was scaled by the L_1 distance between the predicted and reference label to improve the policy's ability to distinguish between tasks that were temporally distant but visually similar. The individual loss components were weighted on the basis of their relative importance to the task. The task instruction had the highest priority, so its weight, $w_p = 0.4$, was set higher than the weight for the correction flag and corrective instruction predictions, which were $w_c = w_m = 0.3$. The resulting objective function minimizes the expected weighted sum of the task, correction, and motion losses and is given as the following. We use a hat symbol to denote the outputs predicted by the HL policy, whereas the corresponding ground-truth values from the dataset are written without the hat.

$$\min_{\pi_{HL}} \mathbb{E}_{(o_{t-k:t}, p_t, c_t, m_t)\sim D} \Bigg[w_p \cdot \underbrace{L_{CE}\left(\pi_{HL}\left(\hat{p}_t \mid o_{t-k:t}\right), p_t\right)}_{\text{Task CE loss}} \cdot \underbrace{\|\hat{p}_t - p_t\|_1}_{\text{Task } L_1 \text{ distance}} + w_c \cdot \underbrace{L_{CE}\left(\pi_{HL}\left(\hat{c}_t \mid o_{t-k:t}\right), c_t\right)}_{\text{Correction CE loss}} + w_m \cdot \underbrace{L_{CE}\left(\pi_{HL}\left(\hat{m}_t \mid o_{t-k:t}\right), m_t\right)}_{\text{Motion CE loss}} \Bigg] \quad (1)$$

Model architecture

The HL policy architecture, illustrated in Fig. 2B, consists of a vision encoder, a transformer decoder (*53*), and separate multilayer perceptron heads to generate the three classification outputs. Each image underwent preprocessing, including standardization based on the mean and SD of the color channels calculated over the entire dataset, ensuring zero mean and unit SD. The image was resized to 224 pixels by 224 pixels to match the resolution used for pretraining the vision encoder. Alongside this global view, a centered crop that captured the most task-critical region was extracted and resized to 224 pixels by 224 pixels. The centered crop covered the inner 50% of the width and captured the lower 80% of the height, starting from the bottom. This approach was inspired by LLaVA's AnyRes technique (*54*), which divides images into multiple patches while preserving the global scene context. However, instead of generating multiple patches, we focused on extracting only the most task-relevant patch, emphasizing the center of the surgical area. The vision encoder is the tiny variant of the Swin Transformer (*55*) pretrained on ImageNet (*56*). The Swin Transformer was selected because of its high performance on limited data and its ability to produce a compact output token size of 768, which makes it suitable for temporal modeling with a downstream Transformer architecture. During surgery, important details are often occluded. For instance, a clip could easily be occluded by an instrument. To retain information crucial for classification, we included a history of $k = 4$ past image frames, each spaced 1 s apart, along with the current frame as input to the HL policy, following the approach of Shi *et al.* (*52*). The embeddings from the vision encoder were used as inputs to the transformer decoder, configured with eight heads and six layers. To preserve temporal information, sinusoidal position embeddings were added to the input sequence. The vision encoder outputs were passed to the Transformer directly without pooling to preserve spatial information, similar to the approach by Zhao *et al.* (*57*). By assigning unique learnable embeddings as task-specific queries (*57*), the transformer decoder can effectively attend to relevant spatial and temporal details, optimizing the alignment of each output with the most appropriate image frames and their features.

Training

The HL policy base model was trained on dataset D with the AdamW (*58*) optimizer, a learning rate of 1×10^{-5}, and a weight decay of 5×10^{-2}. To improve both convergence and generalization, an annealing cosine weight schedule with a linear warmup of five epochs was applied. We also incorporated data augmentation techniques, including RandAugment (*59*) and coarse dropout from Albumentations (*60*), to boost visual robustness. Because of our specific dataset design, which consisted of multiple individual recordings per task rather than continuous procedure recordings, two randomly sampled continuous task recordings were concatenated to artificially generate task transitions during HL policy training. To encourage the policy to learn a wider range of task semantics, 60% of the input sequences began within a recovery mode demonstration, exposing the policy to varying task executions and recovery

scenarios. During training, we applied a prediction offset, where the policy was trained to predict the task instruction 0.5 s into the future rather than predicting the current surgical state. This encouraged the policy to anticipate upcoming actions and better handle task transitions (*52*). After the HL DAgger dataset D_{corr} was collected, the HL policy was fine-tuned on the merged dataset $D \cup D_{\text{corr}}$.

Inference

Every 3 s, the HL policy predicted a new task instruction p_t, correction flag c_t, and corrective instruction m_t. On the basis of the corrective flag c_t, the language instruction provided to the LL policy was then either the task instruction p_t or the corrective motion m_t, as defined by Eq. 2

$$l_t = \begin{cases} p_t, \text{if } c_t = 0 \\ m_t, \text{if } c_t = 1 \end{cases} \tag{2}$$

During inference, a human supervisor can override the HL policy's outputs via voice command or by selecting a task instruction or correction from a drop-down menu in our application GUI. If a manual correction was made, then the HL policy outputs were overridden for the next 3 s.

LL policy

Problem definition

The LL policy was formulated as a language-conditioned policy $\pi_{\text{LL}}(a_{t:t+k} \mid o_t, l_t)$ to predict a sequence of robot actions $a_{t:t+k}$ based on the current image observation o_t and language instruction l_t. l_t can either be p_t or m_t depending on the correction flag c_t. The input observations include the stereo endoscope's left image, along with images from the left and right wrist cameras. For the action representation, we adopted the hybrid-relative action representation from (*22*), which models relative Cartesian translations with respect to the endoscope tip and rotations relative to the end effector. This formulation compensates for the dVRK's kinematic inconsistencies (*61*), leading to more consistent multistep motion predictions. The policy was trained using behavior cloning, where the objective is to minimize the L_1 loss between the predicted action sequence and reference actions. The objective function is expressed in Eq. 3

$$\min_{\pi_{\text{LL}}} \mathbb{E}_{(o_t, l_t, a_{t:t+k}) \sim D} \left[\left\| \pi_{\text{LL}}(\hat{a}_{t:t+k} \mid o_t, l_t) - a_{t:t+k} \right\|_1 \right] \tag{3}$$

Model architecture

The LL policy was built on a decoder-only, BERT-like Transformer (*62*) that maps visual inputs to robot actions, as shown in Fig. 2B. The visual inputs consisted of images from the endoscope and wrist cameras, and they were encoded via a pretrained EfficientNet-B3 (*63*). The encodings were then fused with language instruction embeddings from the HL policy using feature-wise linear modulation (FiLM) layers (*64*). Language instructions were encoded using distilled bidirectional encoder representations from transformers (DistilBERT) (*65*). The fused visual and language embeddings, along with positional embeddings, were passed into the transformer decoder. The action space was a 20D vector representing the relative actions for both robot arms (three for translation, six for rotation, and one for jaw angle per arm). Note that, for the rotation, we used the 6D rotation formulation (*22*, *66*), where the rotation was represented by the first two columns of the rotation matrix. Its third column can be extrapolated by multiplying the first two columns, thus recovering the full rotation matrix. With action chunking, the decoder outputs a $k \times 20$ tensor given the current observation. To optimize performance (*57*), we predicted robot actions for a 2-s horizon, resulting in a chunk size of $k = 60$.

Training

During training, the input images were resized to 224 pixels by 224 pixels. To prevent overfitting, we applied several data augmentation techniques, including random cropping, rotation, shifting, color jittering, and coarse dropout using Albumentations (*60*). Additionally, we applied a 7% random dropout to one of the three input images, preventing the policy from overrelying on any single image observation. To generate corrective language labels from the demonstration data, we examined a future chunk of actions and computed the motion trend along each axis. By comparing the magnitudes of motion across axes, we could determine the dominant direction of movement within that action segment and thus assign directional motion labels such as "move the left arm to the right" or "move the right arm toward me." The chunk size here was set to 10 because we wanted to capture the unit of motion in the collected trajectories. If we set the chunk size too small, then the generated instructions would be too noisy, and, if the chunk size was too large, then the more delicate motions would be ignored. During training, task instructions (e.g., "grabbing gallbladder") were used when sampling from the base dataset, and corrective instructions (e.g., "move left arm toward me") were used when sampling from recovery demonstrations. This enabled the LL policy to execute appropriate actions when given task instructions and recover from failure modes when given corrective instructions. The policy contains ~72 million parameters and is trained on a single RTX 4090 GPU (24 GB). Each epoch takes around 4 min with a batch size of 10, and training runs for 1500 epochs (100 hours) before evaluation.

Inference

Inference time to produce a single action was ~20 ms on the same hardware. To optimize performance, we set different execution horizons (the number of actions executed before resampling the LL policy) for various phases of the procedure. For the "grabbing gallbladder" phase, we found through preliminary experiments that a shorter horizon caused the robot to change strategies too frequently, leading to hesitation and continuous pose adjustments without fully committing to a successful strategy. Setting the execution horizon to 30 time steps ensured that the robot committed to a single strategy. In contrast, for the other phases, we set a shorter execution horizon of 20 time steps to enable more frequent replanning. These phases required high precision, particularly when maneuvering the right arm between the duct and artery. Additionally, manual tool switching and clip loading between tasks were necessary during experiments. To manage these transitions, we implemented a logic-based state machine to automatically pause both the HL and LL policies during phase transitions. For instance, the pauses were triggered when shifting from "going back from the first clip left tube" to "clipping second clip left tube" or from "going back from third clip right tube" to "going to the cutting position on the right tube."

Statistical analysis

The means computed in Fig. 4 with sample size N and data point x were computed using the following equation:

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Supplementary Materials

The PDF file includes:
Materials and Methods
Results
Tables S1 and S2
Figs. S1 to S7

Other Supplementary Material for this manuscript includes the following:
MDAR Reproducibility Checklist

REFERENCES AND NOTES

1. P. M. Scheikl, B. Gyenes, R. Younis, C. Haas, G. Neumann, F. Mathis-Ullrich, M. Wagner, Lapgym - An open source framework for reinforcement learning in robot-assisted laparoscopic surgery. *J. Mach. Learn. Res.* **24**, 1–42 (2023).
2. Q. Yu, M. Moghani, K. Dharmarajan, V. Schorp, William Chung-Ho Panitch, J. Liu, K. Hari, H. Huang, M. Mittal, K. Goldberg, A. Garg, Orbitsurgical: An open-simulation framework for learning surgical augmented dexterity. arXiv:2404.16027 [cs.RO] (2024).
3. J. Xu, B. Li, B. Lu, Y.-H. Liu, Q. Dou, P.-A. Heng, "Surrol: An open-source reinforcement learning centered and dVRK compatible platform for surgical robot learning" in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (IEEE, 2021), pp. 1821–1828.
4. H. Saeidi, J. D. Opfermann, M. Kam, S. Wei, S. Leonard, M. H. Hsieh, J. U. Kang, A. Krieger, Autonomous robotic laparoscopic surgery for intestinal anastomosis. *Sci. Robot.* **7**, eabj2908 (2022).
5. X. Liang, C.-P. Wang, N. U. Shinde, F. Liu, F. Richter, M. Yip, Medic: Autonomous surgical robotic assistance to maximizing exposure for dissection and cautery. arXiv:2409.14287 [cs.RO] (2024).
6. G. Fagogenis, M. Mencattelli, Z. Machaidze, B. Rosa, K. Price, F. Wu, V. Weixler, M. Saeed, J. E. Mayer, P. E. Dupont, Autonomous robotic intracardiac catheter navigation using haptic vision. *Sci. Robot.* **4**, eaaw1977 (2019).
7. B. Thananjeyan, A. Garg, S. Krishnan, C. Chen, L. Miller, K. Goldberg, "Multilateral surgical pattern cutting in 2D orthotropic gauze with deep reinforcement learning policies for tensioning" in *2017 IEEE International Conference on Robotics and Automation (ICRA)* (IEEE, 2017), pp. 2371–2378.
8. A. Kuntz, M. Emerson, T. E. Ertop, I. Fried, M. Fu, J. Hoelscher, M. Rox, J. Akulian, E. A. Gillaspie, Y. Z. Lee, F. Maldonado, R. J. Webster III, R. Alterovitz, Autonomous medical needle steering in vivo. *Sci. Robot.* **8**, eadf7614 (2023).
9. M. Hwang, J. Ichnowski, B. Thananjeyan, D. Seita, S. Paradis, D. Fer, T. Low, K. Goldberg, Automating surgical peg transfer: Calibration with deep learning can exceed speed, accuracy, and consistency of humans. *IEEE Trans. Autom. Sci. Eng.* **20**, 909–922 (2023).
10. M. Afshar, J. Carriere, T. Meyer, R. S. Sloboda, S. Husain, N. Usmani, M. Tavakoli, A model-based multi-point tissue manipulation for enhancing breast brachytherapy. *IEEE Trans. Med. Robot. Bionics* **4**, 1046–1056 (2022).
11. J. Hu, D. Jones, M. R. Dogar, P. Valdastri, Occlusion-robust autonomous robotic manipulation of human soft tissues with 3-D surface feedback. *IEEE Trans. Robot.* **40**, 624–638 (2024).
12. Y. Ou, M. Tavakoli, Sim-to-real surgical robot learning and autonomous planning for internal tissue points manipulation using reinforcement learning. *IEEE Robot. Autom. Lett.* **8**, 2502–2509 (2023).
13. P. M. Scheikl, E. Tagliabue, B. Gyenes, M. Wagner, D. Dall'Alba, P. Fiorini, F. Mathis-Ullrich, Sim-to-real transfer for visual reinforcement learning of deformable object manipulation for robot-assisted surgery. *IEEE Robot. Autom. Lett.* **8**, 560–567 (2023).
14. Z.-Y. Chiu, F. Richter, E. K. Funk, R. K. Orosco, M. C. Yip, "Bimanual regrasping for suture needles using reinforcement learning for rapid motion planning" in *2021 IEEE International Conference on Robotics and Automation (ICRA)* (IEEE, 2021), pp. 7737–7743.
15. M. Haiderbhai, R. Gondokaryono, A. Wu, L. A. Kahrs, Sim2real rope cutting with a surgical robot using vision-based reinforcement learning. *IEEE Trans. Autom. Sci. Eng.* **22**, 4354–4365 (2024).
16. P. M. Scheikl, N. Schreiber, C. Haas, N. Freymuth, G. Neumann, R. Lioutikov, F. Mathis-Ullrich, Movement primitive diffusion: Learning gentle robotic manipulation of deformable objects. *IEEE Robot. Autom. Lett.* **9**, 5338–5345 (2024).
17. C. Shin, P. W. Ferguson, S. A. Pedram, J. Ma, E. P. Dutson, J. Rosen, "Autonomous tissue manipulation via surgical robot using learning based model predictive control" in *2019 International Conference on Robotics and Automation (ICRA)* (IEEE, 2019), pp. 3875–3881.
18. A. K. Tanwani, A. Yan, J. Lee, S. Calinon, K. Goldberg, Sequential robot imitation learning from observations. *Int. J. Robot. Res.* **40**, 1306–1325 (2021).
19. H. Su, A. Mariani, S. E. Ovur, A. Menciassi, G. Ferrigno, E. De Momi, Toward teaching by demonstration for robot-assisted minimally invasive surgery. *IEEE Trans. Autom. Sci. Eng.* **18**, 484–494 (2021).
20. A. Pore, E. Tagliabue, M. Piccinelli, D. Dall'Alba, A. Casals, P. Fiorini, "Learning from demonstrations for autonomous soft-tissue retraction" in *2021 International Symposium on Medical Robotics (ISMR)* (IEEE, 2021), pp. 1–7.
21. S. Schmidgall, J. W. Kim, A. Kuntz, A. E. Ghazi, A. Krieger, General-purpose foundation models for increased autonomy in robot-assisted surgery. arXiv:2401.00678 [cs.RO] (2024).
22. J. W. Kim, T. Z. Zhao, S. Schmidgall, A. Deguet, M. Kobilarov, C. Finn, A. Krieger, "Surgical robot transformer (SRT): Imitation learning for surgical subtasks" in *Proceedings of the 8th Annual Conference on Robot Learning (CoRL 2024)*, vol. 270 of *Proceedings of Machine Learning Research*, P. Agrawal, O. Kroemer, W. Burgard, Eds. (MLResearchPress, 2024), pp. 130–144.
23. S. Ross, G. Gordon, D. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning" in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, vol. 15 of *Proceedings of Machine Learning Research*, G. Gordon, D. Dunson, M. Dudík, Eds. (MLResearch Press, 2011), pp. 627–635.
24. M. Acalovschi, F. Lammert, The growing global burden of gallstone disease. *World Gastroenterol. News* **17**, 6–9 (2012).
25. K. Hsu, M. J. Kim, R. Rafailov, J. Wu, C. Finn, "Vision-based manipulators need to also see from their hands" in *The Tenth International Conference on Learning Representations* (ICLR, 2022), pp. 1–30.
26. R. Gupta, A. Kumar, C. P. Hariprasad, M. Kumar, Anatomical variations of cystic artery, cystic duct, and gall bladder and their associated intraoperative and postoperative complications: An observational study. *Ann. Med. Surg.* **85**, 3880–3886 (2023).
27. A. Majumder, M. S. Altieri, L. M. Brunt, How do I do it: Laparoscopic cholecystectomy. *Ann. Laparosc. Endosc. Surg.* **5**, 10.21037/ales.2020.02.06 (2020).
28. J. Marescaux, F. Rubino, The ZEUS robotic system: Experimental and clinical applications. *Surg. Clin. North Am.* **83**, 1305–1315 (2003).
29. K. Price, J. Peine, M. Mencattelli, Y. Chitalia, D. Pu, T. Looi, S. Stone, J. Drake, P. E. Dupont, Using robotics to move a neurosurgeon's hands to the tip of their endoscope. *Sci. Robot.* **8**, eadg6042 (2023).
30. A. Shademan, R. S. Decker, J. D. Opfermann, S. Leonard, A. Krieger, P. C. W. Kim, Supervised autonomous robotic soft tissue surgery. *Sci. Transl. Med.* **8**, 337ra64 (2016).
31. T. Haidegger, Autonomy for surgical robots: Concepts and paradigms. *IEEE Trans. Med. Robot. Bionics* **1**, 65–76 (2019).
32. G. Kurup, Cyberknife: A new paradigm in radiotherapy. *J. Med. Phys.* **35**, 63–64 (2010).
33. T. S. Perry, Profile: Veebot. Making a robot that can draw blood faster and more safely than a human can, *IEEE Spectrum*, 26 July 2013; https://spectrum.ieee.org/profile-veebot.
34. T. D. Nagy, T. Haidegger, Performance and capability assessment in surgical subtask automation. *Sensors* **22**, 2501 (2022).
35. S. Reed, K. Zolna, E. Parisotto, S. G. Colmenarejo, A. Novikov, G. Barth-Maron, M. Gimenez, Y. Sulsky, J. Kay, J. T. Springenberg, T. Eccles, J. Bruce, A. Razavi, A. Edwards, N. Heess, Y. Chen, R. Hadsell, O. Vinyals, M. Bordbar, N. de Freitas, A generalist agent. arXiv:2205.06175 [cs.AI] (2022).
36. A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, J. Ibarz, B. Ichter, A. Irpan, T. Jackson, S. Jesmonth, N. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, I. Leal, K.-H. Lee, S. Levine, Y. Lu, U. Malla, D. Manjunath, I. Mordatch, O. Nachum, C. Parada, J. Peralta, E. Perez, K. Pertsch, J. Quiambao, K. Rao, M. S. Ryoo, G. Salazar, P. R. Sanketi, K. Sayed, J. Singh, S. Sontakke, A. Stone, C. Tan, H. Tran, V. Vanhoucke, S. Vega, Q. H. Vuong, F. Xia, T. Xiao, P. Xu, S. Xu, T. Yu, B. Zitkovich, "Rt-1: Robotics transformer for real-world control at scale" in *Proceedings of Robotics: Science and Systems*, K. Bekris, K. Hauser, S. Herbert, J. Yu, Eds. (RSS Foundation, 2023), 10.15607/RSS.2023.XIX.025.
37. A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn, P. Florence, C. Fu, M. G. Arenas, K. Gopalakrishnan, K. Han, K. Hausman, A. Herzog, J. Hsu, B. Ichter, A. Irpan, N. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, I. Leal, L. Lee, T.-W. E. Lee, S. Levine, Y. Lu, H. Michalewski, I. Mordatch, K. Pertsch, K. Rao, K. Reymann, M. Ryoo, G. Salazar, P. Sanketi, P. Sermanet, J. Singh, A. Singh, R. Soricut, H. Tran, V. Vanhoucke, Q. Vuong, A. Wahid, S. Welker, P. Wohlhart, J. Wu, F. Xia, T. Xiao, P. Xu, S. Xu, T. Yu, B. Zitkovich, "Rt-2: Vision-language-action models transfer web knowledge to robotic control" in *Proceedings of the 7th Conference on Robot Learning*, vol. 229 of *Proceedings of Machine Learning Research*, J. Tan, M. Toussaint, K. Darvish, Eds. (MLResearchPress, 2023), pp. 2165–2183.
38. A. O'Neill, A. Rehman, A. Gupta, A. Maddukuri, A. Gupta, A. Padalkar, A. Lee, A. Pooley, A. Gupta, A. Mandlekar, A. Jain, A. Tung, A. Bewley, A. Herzog, A. Irpan, A. Khazatsky, A. Rai, A. Gupta, A. Wang, A. Kolobov, A. Singh, A. Garg, A. Kembhavi, A. Xie, A. Brohan, A. Raffin, A. Sharma, A. Yavary, A. Jain, A. Balakrishna, A. Wahid, B. Burgess-Limerick, B. Kim, B. Schölkopf, B. Wulfe, B. Ichter, C. Lu, C. Xu, C. Le, C. Finn, C. Wang, C. Xu, C. Chi, C. Huang, C. Chan, C. Agia, C. Pan, C. Fu, C. Devin, D. Xu, D. Morton, D. Driess, D. Chen, D. Pathak, D. Shah, D. Büchler, D. Jayaraman, D. Kalashnikov, D. Sadigh, E. Johns, E. Foster, F. Liu, F. Ceola, F. Xia, F. Zhao, F. V. Frujeri, F. Stulp, G. Zhou, G. S. Sukhatme, G. Salhotra, G. Yan, G. Feng, G. Schiavi, G. Berseth, G. Kahn, G. Yang, G. Wang, H. Su, H.-S. Fang, H. Shi,

H. Bao, H. B. Amor, H. I. Christensen, H. Furuta, H. Bharadhwaj, H. Walke, H. Fang, H. Ha, I. Mordatch, I. Radosavovic, I. Leal, J. Liang, J. Abou-Chakra, J. Kim, J. Drake, J. Peters, J. Schneider, J. Hsu, J. Vakil, J. Bohg, J. Bingham, J. Wu, J. Gao, J. Hu, J. Wu, J. Wu, J. Sun, J. Luo, J. Gu, J. Tan, J. Oh, J. Wu, J. Lu, J. Yang, J. Malik, J. Silvério, J. Hejna, J. Booher, J. Tompson, J. Yang, J. Salvador, J. J. Lim, J. Han, K. Wang, K. Rao, K. Pertsch, K. Hausman, K. Go, K. Gopalakrishnan, K. Goldberg, K. Byrne, K. Oslund, K. Kawaharazuka, K. Black, K. Lin, K. Zhang, K. Ehsani, K. Lekkala, K. Ellis, K. Rana, K. Srinivasan, K. Fang, K. P. Singh, K.-H. Zeng, K. Hatch, K. Hsu, L. Itti, L. Y. Chen, L. Pinto, L. Fei-Fei, L. Tan, L. Fan, L. Ott, L. Lee, L. Weihs, M. Chen, M. Lepert, M. Memmel, M. Tomizuka, M. Itkina, M. G. Castro, M. Spero, M. Du, M. Ahn, M. C. Yip, M. Zhang, M. Ding, M. Heo, M. K. Srirama, M. Sharma, M. J. Kim, M. Z. Irshad, N. Kanazawa, N. Hansen, N. Heess, N. J. Joshi, N. Suenderhauf, N. Liu, N. D. Palo, N. M. M. Shafiullah, O. Mees, O. Kroemer, O. Bastani, P. R. Sanketi, P. Miller, P. Yin, P. Wohlhart, P. Xu, P. D. Fagan, P. Mitrano, P. Sermanet, P. Abbeel, P. Sundaresan, Q. Chen, Q. Vuong, R. Rafailov, R. Tian, R. Doshi, R. Martín-Martín, R. Baijal, R. Scalise, R. Hendrix, R. Lin, R. Qian, R. Zhang, R. Mendonca, R. Shah, R. Hoque, R. Julian, S. Bustamante, S. Kirmani, S. Levine, S. Lin, S. Moore, S. Bahl, S. Dass, S. Sonawani, S. Tulsiani, S. Song, S. Xu, S. Haldar, S. Karamcheti, S. Adebola, S. Guist, S. Nasiriany, S. Schaal, S. Welker, S. Tian, S. Ramamoorthy, S. Dasari, S. Belkhale, S. Park, S. Nair, S. Mirchandani, T. Osa, T. Gupta, T. Harada, T. Matsushima, T. Xiao, T. Kollar, T. Yu, T. Ding, T. Davchev, T. Z. Zhao, T. Armstrong, T. Darrell, T. Chung, V. Jain, V. Kumar, V. Vanhoucke, V. Guizilini, W. Zhan, W. Zhou, W. Burgard, X. Chen, X. Chen, X. Wang, X. Zhu, X. Geng, X. Liu, X. Liangwei, X. Li, Y. Pang, Y. Lu, Y. J. Ma, Y. Kim, Y. Chebotar, Y. Zhou, Y. Zhu, Y. Wu, Y. Xu, Y. Wang, Y. Bisk, Y. Dou, Y. Cho, Y. Lee, Y. Cui, Y. Cao, Y.-H. Wu, Y. Tang, Y. Zhu, Y. Zhang, Y. Jiang, Y. Li, Y. Li, Y. Iwasawa, Y. Matsuo, Z. Ma, Z. Xu, Z. J. Cui, Z. Zhang, Z. Fu, Z. Lin, Open X-Embodiment: Robotic learning datasets and RT-X models (2023); https://robotics-transformer-x.github.io.

39. Y. Hu, Q. Xie, V. Jain, J. Francis, J. Patrikar, N. Keetha, S. Kim, Y. Xie, T. Zhang, H.-S. Fang, S. Zhao, S. Omidshafiei, D.-K. Kim, Ali-akbar Agha-mohammadi, K. Sycara, M. Johnson-Roberson, D. Batra, X. Wang, S. Scherer, C. Wang, Z. Kira, F. Xia, Y. Bisk, Toward general-purpose robots via foundation models: A survey and metaanalysis. arXiv:2312.08782 [cs.RO] (2023).
40. M. Ballester, H. Wang, J. Li, O. Cossairt, F. Willomitzer, Single-shot synthetic wavelength imaging: Sub-mm precision ToF sensing with conventional CMOS sensors. *Opt. Lasers Eng.* **178**, 108165 (2024).
41. J. Sayers, N. G. Czakon, P. K. Day, T. P. Downes, R. P. Duan, J. Gao, J. Glenn, S. R. Golwala, M. I. Hollister, H. G. LeDuc, B. A. Mazin, P. R. Maloney, O. Noroozian, H. T. Nguyen, J. A. Schlaerth, S. Siegel, J. E. Vaillancourt, A. Vayonakis, P. R. Wilson, J. Zmuidzinas, "Optics for music: a new (sub) millimeter camera for the Caltech submillimeter observatory" in *Millimeter, Submillimeter, and Far-Infrared Detectors and Instrumentation for Astronomy V* (SPIE, 2010), vol. 7741, pp. 255–266.
42. C. Nezhat, V. Morozov, A simple solution to lens fogging during robotic and laparoscopic surgery. *J. Soc. Laparoendosc. Surg.* **12**, 431 (2008).
43. ClickClean, ClickClean laparoscope lens shield device; https://clickclean-medeon.com/.
44. ClearCam, Clearcam—Laparoscope lens cleaning; www.clearcam-med.com/.
45. Y. Chebotar, Q. Vuong, A. Irpan, K. Hausman, F. Xia, Y. Lu, A. Kumar, T. Yu, A. Herzog, K. Pertsch, K. Gopalakrishnan, J. Ibarz, O. Nachum, S. Sontakke, G. Salazar, H. T. Tran, J. Peralta, C. Tan, D. Manjunath, J. Singht, B. Zitkovich, T. Jackson, K. Rao, C. Finn, S. Levine, "Q-transformer: Scalable offline reinforcement learning via autoregressive q-functions" in *Proceedings of the 7th Conference on Robot Learning*, vol. 229 of *Proceedings of Machine Learning Research*, J. Tan, M. Toussaint, K. Darvish, Eds. (MLResearchPress, 2023), pp. 3909–3928.
46. A. Z. Ren, A. Dixit, A. Bodrova, S. Singh, S. Tu, N. Brown, P. Xu, L. Takayama, F. Xia, J. Varley, Z. Xu, D. Sadigh, A. Zeng, A. Majumdar, "Robots that ask for help: Uncertainty alignment for large language model planners" in *Proceedings of the 7th Conference on Robot Learning*, J. Tan, M. Toussaint, K. Darvish, Eds. (MLResearchPress, 2023), vol. 229, pp. 661–682.
47. A. N. Angelopoulos, S. Bates, Conformal prediction: A gentle introduction. *Found. Trends Mach. Learn.* **16**, 494–591 (2023).
48. T. Zhang, Toward automated vehicle teleoperation: Vision, opportunities, and challenges. *IEEE Internet Things J.* **7**, 11347–11354 (2020).
49. T. Lim, M. Hwang, E. Kim, H. Cha, Authority transfer according to a driver intervention intention considering coexistence of communication delay. *Computers* **12**, 228 (2023).
50. S. Gupta, K. Yao, L. Niederhauser, A. Billard, Action contextualization: Adaptive task planning and action tuning using large language models. *IEEE Robot. Autom. Lett.* **9**, 9407–9414 (2024).
51. M. Kelly, C. Sidrane, K. Driggs-Campbell, M. J. Kochenderfer, "Hgdagger: Interactive imitation learning with human experts" in *2019 International Conference on Robotics and Automation (ICRA)* (IEEE, 2019), pp. 8077–8083.
52. L. X. Shi, Z. Hu, T. Z. Zhao, A. Sharma, K. Pertsch, J. Luo, S. Levine, C. Finn, "Yell at your robot: Improving on-the-fly from language corrections" in *Proceedings of Robotics: Science and Systems*, D. Kulic, G. Venture, K. Bekris, E. Coronado, Eds. (2024), 10.15607/RSS.2024.XX.025.
53. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, "Attention is all you need" in *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett, Eds. (Curran Associates, 2017), vol. 30, pp. 5998–6008.
54. H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, Y. J. Lee, LLaVA-NeXT: Improved reasoning, OCR, and world knowledge, LLaVa (2024); https://llava-vl.github.io/blog/2024-01-30-llava-next/.
55. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, St. Lin, B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows" in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (IEEE, 2021), pp. 10012–10022.
56. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, "ImageNet: A large-scale hierarchical image database" in *2009 IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2009), pp. 248–255.
57. T. Z. Zhao, V. Kumar, S. Levine, C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware" in *Proceedings of Robotics: Science and Systems*, K. Bekris, K. Hauser, S. Herbert, J. Yu, Eds. (RSS Foundation, 2023), 10.15607/RSS.2023.XIX.016.
58. I. Loshchilov, F. Hutter, "Decoupled weight decay regularization" in *ICLR 2019: The Seventh International Conference on Learning Representations* (ICLR, 2019).
59. E. D. Cubuk, B. Zoph, J. Shlens, Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space" in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (CVPRW) (IEEE, 2020), pp. 3008–3017.
60. A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, A. A. Kalinin, Albumentations: Fast and flexible image augmentations. *Information* **11**, 125 (2020).
61. M. Hwang, B. Thananjeyan, S. Paradis, D. Seita, J. Ichnowski, D. Fer, T. Low, K. Goldberg, Efficiently calibrating cable-driven surgical robots with RGBD fiducial sensing and recurrent neural networks. *IEEE Robot. Autom. Lett.* **5**, 5937–5944 (2020).
62. J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding" in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, T. Soorio, Eds. (Association for Computational Linguistics, 2019), pp. 4171–4186.
63. M. Tan, Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks" in *Proceedings of the 36th International Conference on Machine Learning*, K. Chaudhuri, R. Salakhutdinov, Eds., vol. 97 of *Proceedings of Machine Learning Research* (MLResearchPress, 2019), pp. 6105–6114.
64. E. Perez, F. Strub, H. De Vries, V. Dumoulin, A. Courville, "FiLM: Visual reasoning with a general conditioning layer" in *Proceedings of the AAAI Conference on Artificial Intelligence* (AAAI, 2018), vol. 32, pp. 3942–3951.
65. V. Sanh, L. Debut, J. Chaumond, T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter" in *5th Workshop on Energy Efficient Machine Learning and Cognitive Computing @ NeurIPS 2019* (NeurIPS, 2019), pp. 1–5.
66. Y. Zhou, C. Barnes, L. Jingwan, Y. Jimei, L. Hao, "On the continuity of rotation representations in neural networks" in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2019), pp. 5745–5753.

Acknowledgments

Funding: Research reported in this publication was supported by the Advanced Research Projects Agency for Health (ARPA-H) under award number 75N91023C00048 and NSF/FRR 2144348, NIH R56EB033807, and NSF DGE 2139757. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the US government. **Author contributions:** Conceptualization: J.W.K., A.K., S.S., D.R.T., R.J.C., and C.F. Methodology: J.W.K., C.F., and A.K. Software: J.W.K., L.X.S., P.H., J.-T.C., A.D., and P.M.S. Visualization: J.W.K., S.S., P.H., J.-T.C., and P.M.S. Data curation: A.G., J.W.K., P.H., J.J., and B.M.W. Formal analysis: J.W.K., P.H., and J.-T.C. Funding acquisition: A.K. and R.J.C. Supervision: A.K., J.W.K., and C.F. Writing—original draft: J.W.K., S.S., P.H., J.-T.C., P.M.S., and A.G. Writing—review and editing: J.W.K., S.S., P.H., J.-T.C., P.M.S., A.G., A.K., C.F., L.X.S., D.R.T., and R.J.C. **Competing interests:** A.K., C.F., J.W.K., D.R.T., J.-T.C., P.H., S.S., and P.M.S. have a provisional patent pending: "Imitation learning for surgical robots with kinematics errors using self-corrections." R.J.C. has ownership interests in and serves as a scientific advisor for Optosurgical LLC. All other authors declare that they have no competing interests. **Data and materials availability:** All data needed to support the conclusions of this manuscript are included in the main text or Supplementary Materials. The datasets and code used to generate Figs. 4 and 6 and fig. S7 are available at Zenodo: https://zenodo.org/records/15637074.

Submitted 30 September 2024
Accepted 11 June 2025
Published 9 July 2025
10.1126/scirobotics.adt5254

Science Robotics

SRT-H: A hierarchical framework for autonomous surgery via language-conditioned imitation learning

Ji Woong (Brian) Kim, Juo-Tung Chen, Pascal Hansen, Lucy Xiaoyang Shi, Antony Goldenberg, Samuel Schmidgall, Paul Maria Scheikl, Anton Deguet, Brandon M. White, De Ru Tsai, Richard Jaepyeong Cha, Jeffrey Jopling, Chelsea Finn, and Axel Krieger

Sci. Robot. **10** (104), eadt5254. DOI: 10.1126/scirobotics.adt5254

Science Robotics (ISSN 2470-9476) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title *Science Robotics* is a registered trademark of AAAS.