MEDICAL ROBOTS

Surgical embodied intelligence for generalized task autonomy in laparoscopic robot-assisted surgery

Yonghao Long[1], Anran Lin[1], Derek Hang Chun Kwok[2], Lin Zhang[2], Zhenya Yang[1], Kejian Shi[1], Lei Song[1], Jiawei Fu[1], Hongbin Lin[3], Wang Wei[1], Kai Chen[1], Xiangyu Chu[3], Yang Hu[2], Hon Chi Yip[4]*, Philip Wai Yan Chiu[4], Peter Kazanzides[5], Russell H. Taylor[5], Yunhui Liu[3], Zihan Chen[2], Zerui Wang[2]*, Samuel Kwok Wai Au[2,3], Qi Dou[1]*

Surgical robots capable of autonomously performing various tasks could enhance efficiency and augment human productivity in addressing clinical needs. Although current solutions have automated specific actions within defined contexts, they are challenging to generalize across diverse environments in general surgery. Embodied intelligence enables general-purpose robot learning with applications for daily tasks, yet its application in the medical domain remains limited. We introduced an open-source surgical embodied intelligence simulator for an interactive environment to develop reinforcement learning methods for minimally invasive surgical robots. Using such embodied artificial intelligence, this study further addresses surgical task automation, enabling zero-shot transfer of simulation-trained policies to real-world scenarios. The proposed method encompasses visual parsing, a perceptual regressor, policy learning, and a visual servoing controller, forming a paradigm that combines the advantages of data-driven policy and classic controller. The visual parsing uses stereo depth estimation and image segmentation with a visual foundation model to handle complex scenes. Experiments demonstrated autonomy in seven game-based skill training tasks on the da Vinci Research Kit, with a proof-of-concept study on haptic-assisted skill training as a practical application. Moreover, we conducted automation of five surgical assistive tasks with the Sentire surgical system on ex vivo animal tissues with various scenes, object sizes, instrument types, and illuminations. The learned policies were also validated in a live-animal trial for three tasks in dynamic in vivo surgical environments. We hope this open-source infrastructure, coupled with a general-purpose learning paradigm, will inspire and facilitate future research on embodied intelligence toward autonomous surgical robots.

INTRODUCTION

Surgical robots have performed millions of minimally invasive procedures worldwide (*1*), with an annual growth rate exceeding ~18% in the past 3 years (*2*). The trend of an aging population further brings growing clinical demand, placing strain on surgeons who must manage more patients. Autonomy is envisaged for the next-generation surgical robots to enhance operational efficiency, consistency, and human productivity (*3*, *4*).

Traditional methods with task-specific engineering often rely on predefined models or rules tailored to specific contexts. This approach limits their generalizability and applicability to a wide range of surgical tasks intended for automation in general surgery. Artificial intelligence (AI)–powered approaches offer the potential for general-purpose solutions, enabling robots to learn and perform various tasks in diverse environments. A unified data-driven paradigm can allow the framework to be replicated across different surgical tasks with minimal task-specific engineering. For instance, if a framework works for needle manipulation, then it should work similarly for soft tissue retraction, despite their differences, provided that corresponding training data for the new task are available.

Embodied intelligence has demonstrated success in achieving such a goal across various daily tasks (*5*, *6*). However, applying this concept to surgical robotics is challenging. Unlike the general robot learning community, which has a number of popular simulators such as Isaac Gym (*7*) and Habitat (*8*) for robot embodiment, open-source infrastructure for surgical embodied intelligence is still in its early stages. Current surgical simulators include dVRL (*9*), UnityFlexML (*10*), LapGym (*11*), AMBF-RL (*12*), Surgical Gym (*13*), ORBIT-Surgical (*14*), and our previously developed SurRoL (*15*). Although all of these simulators facilitate reinforcement learning (RL) studies to investigate task generalizability in surgical robotics, none have yet demonstrated zero-shot sim-to-real transfer based on robotic visual inputs. The learned policies usually rely on state representations that are not bridged to visual perception, rendering them not ready for in vivo testing in minimally invasive procedures.

This study introduces a vision-based learning paradigm incorporating surgical embodied intelligence for achieving zero-shot sim-to-real transfer for various tasks (Fig. 1). Our proposed paradigm, called VPPV, consists of four sequential components: visual parsing, a perceptual regressor, policy learning, and a visual servoing controller. The perceptual regressor takes image segmentation and depth estimation maps as inputs (either read from the simulator or computed from real-world images) and produces physically meaningful state vectors for policy learning. This allows our framework to overcome the sim-to-real gap by disentangling the simulation-trained policy from raw image perception. It enables direct deployment to real-world scenes as long as the abstracted visual inputs to the perceptual regressor are reliable. This is assured by visual parsing, which leverages foundation models to perceive complex surgical environments. The RL-based policy generates a high-level trajectory to drive the surgical instrument to the target object. A traditional

[1]Department of Computer Science and Engineering, Chinese University of Hong Kong, HKSAR, China. [2]Cornerstone Robotics Ltd., HKSAR, China. [3]Department of Mechanical and Automation Engineering, Chinese University of Hong Kong, HKSAR, China. [4]Department of Surgery, Chinese University of Hong Kong, HKSAR, China. [5]Department of Computer Science, Johns Hopkins University, Baltimore, USA.
*Corresponding author. Email: qidou@cuhk.edu.hk (Q.D.); jerry.wang@csrbtx.com (Z.W.); hcyip@surgery.cuhk.edu.hk (H.C.Y.)

visual servoing controller then executes the final low-level manipulation such as grasping.

Our results demonstrate that the generalized AI framework learned in the simulator can be applied to surgical skill training tasks on the da Vinci Research Kit (dVRK) and surgical assistive tasks in both an ex vivo setting and in vivo animal trials using a commercialized robotic surgical platform (Sentire surgical system, Cornerstone Robotics). Movie 1 summarizes the method and representative results of this paper.

RESULTS

Surgical embodied intelligence simulation environment

We developed an open-source surgical robot learning simulator, called SurRoL (*15*), as an infrastructural platform for surgical embodied intelligence. It has so far been adopted by researchers from international institutions for studying topics such as RL-based task automation (*16*), data-efficient learning (*17*), and human-robot shared control (*18*, *19*). Table 1 summarizes the characteristics of SurRoL compared with other simulators available in the community.

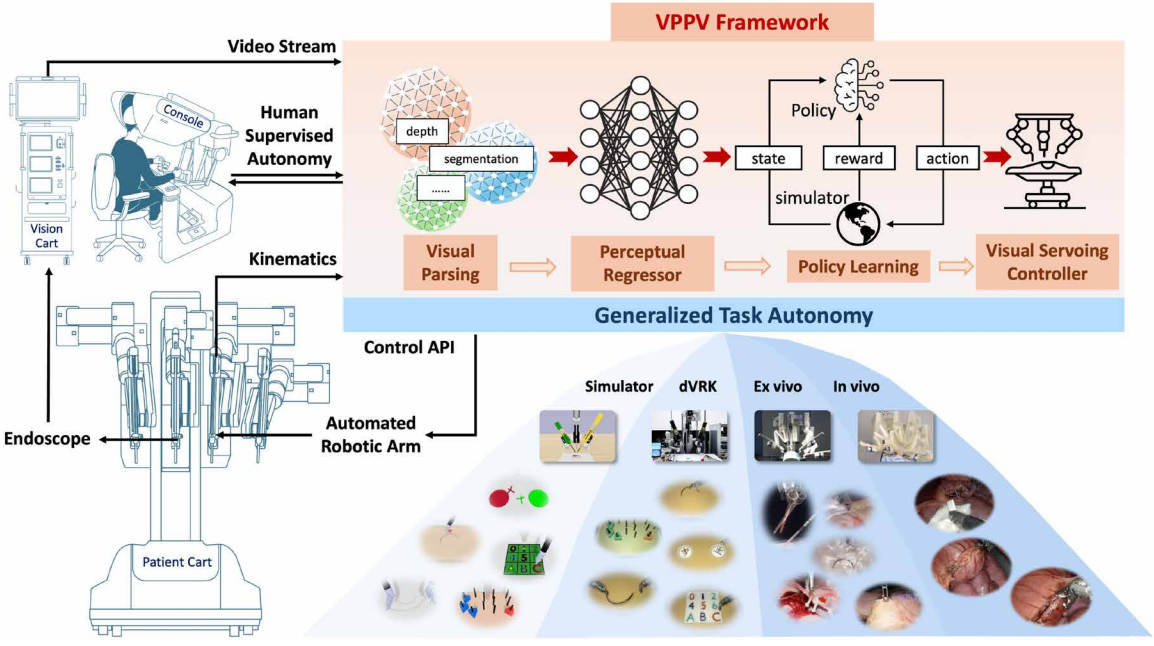
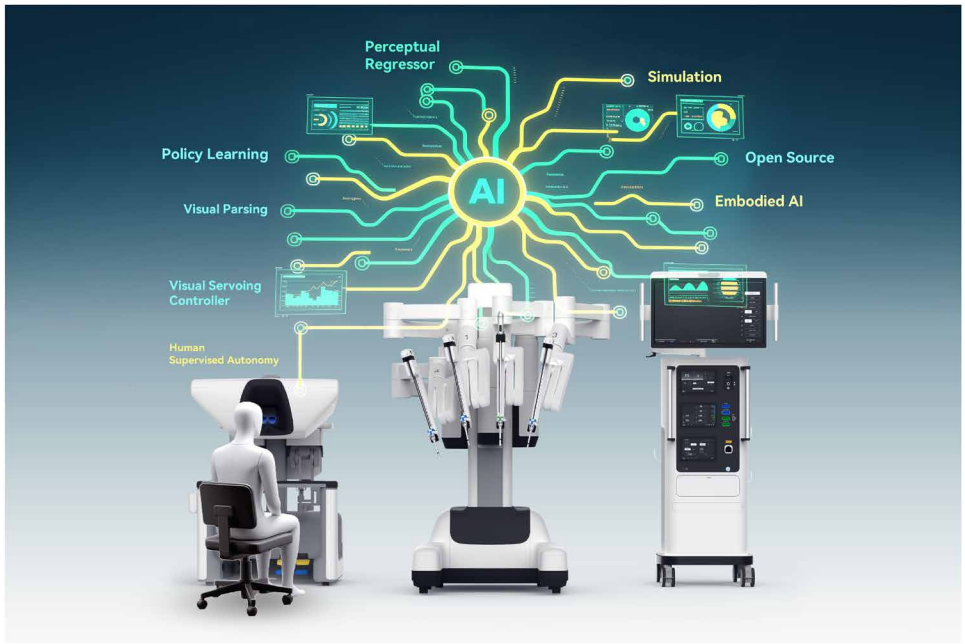


Fig. 1. Concept overview. A vision-based learning paradigm trained through embodied intelligence enables generalized task automation for surgical robots.



Movie 1. Summary of the method and representative results in this paper.

Table 1. Comparison of existing open-source simulators for surgical robotics on multifaceted characteristics. N.A., not applicable; FOV, field of view; PBD, position-based dynamics; FEM, finite element method; PPO, proximal policy optimization; SOTA, state-of-the-art; IL, imitation learning.

Name	Year of first release	Physics engine and renderer	Soft-body simulation	Data-driven scene simulation	Human interaction	Supported tasks	Policy learning library	Sim-to-real deployment
dVRL (*9*)	2019	V-REP, OpenGL	N.A.	N.A.	N.A.	Fundamental action	DDPG	State-based (PSM reach, pick)
UnityFlexML (*10*)	2020	Flex, Unity3D	PBD	N.A.	Teleoperation	Fundamental action	PPO	Vision-based (tissue retraction)
LapGym (*11*)	2023	SOFA, OpenGL	FEM	N.A.	N.A.	ECM FOV control, fundamental action, basic skill training	PPO	N.A.
AMBF-RL (*12*)	2022	Bullet, OpenGL	PBD	N.A.	Teleoperation	Fundamental action	DDPG	N.A.
Surgical Gym (*13*)	2024	PhysX, ray tracing	N.A.	N.A.	N.A.	ECM FOV control, fundamental action	PPO	N.A.
ORBIT-Surgical (*14*)	2024	PhysX, ray tracing	FEM	N.A.	Teleoperation	Fundamental action, basic skill training	PPO, BC	Trajectory replay (2 tasks)
SurRoL (ours) (*15*)	2021	Bullet with custom-developed soft-body engine, OpenGL	MPM	3D Gaussian splatting	Teleoperation, active haptic assistance	ECM FOV control, fundamental action, basic skill training	DDPG, BC, DT SOTA RL/IL	Vision-based (7 tasks; zero-shot)

SurRoL was one of the earliest RL simulators provided to the field, and it can now comprehensively support the physics engine, a visualization renderer, soft-body simulation, human interaction, assets for robotic tasks, a policy learning library, and sim-to-real deployment. More details are provided in Materials and Methods.

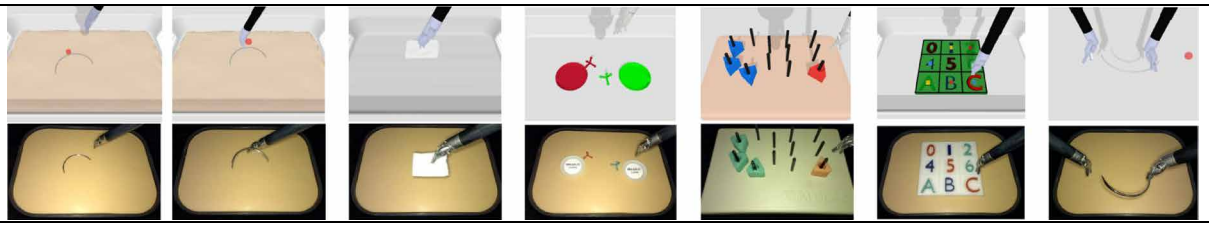
Policy learning in the embodied environment and sim-to-real transfer for task autonomy

Infrastructure for robot learning

SurRoL, similar to any embodied intelligence simulator, offers a Gym environment for robot learning, specifically designed for surgical robotics. We created domain-specific assets and tasks that enable researchers to focus on proposing learning algorithms without the need to reinvent the wheel. The simulator includes a range of game-based tasks aligned with the da Vinci skill training curriculum (*20*), including NeedleReach, NeedlePick, GauzeRetrieve, PickAndPlace, PegTransfer, MatchBoard, and NeedleRegrasp, each varying in difficulty and representing fundamental surgical skills. The embodiment robot can interact with virtual objects for different skills and receive state or visual feedback from the simulator, which can read environmental data, including object positions, RGB images, semantic masks, depth maps, and robotic kinematics. Our learning library supported the implementation for RL methods, featuring built-in functionalities such as a replay buffer (i.e., storing explored trajectories from agents), samplers (i.e., sampling trajectories from buffers to optimize parameters of agents), and parallelized data collection to accelerate model training. A data collection application programming interface (API) was provided to collect script-based or human data, and the agent allowed both value- and policy-based RL methods with either deterministic or stochastic optimizers.

Benchmark of state-based policy learning methods in SurRoL

We established a benchmark for policy learning methods within our simulator, providing access to data, code, and trained models. This served as a reference for the performance of existing approaches and aided in the development of methods for surgical task automation. In addition, it facilitated researchers in the broader robot learning community to easily validate their out-of-the-box algorithms for surgical robotics applications. For the benchmark, we compared a number of end-to-end learning methods that combined visual understanding, trajectory prediction, and action execution in one unified network. Most existing methods belong to state-based policy learning, explicitly read the environment states from the simulator (e.g., object position, rotation, and robot pose), and directly use them to predict actions. Fig. 2 shows the results of different methods for SurRoL tasks. We included well-known RL methods such as deep deterministic policy gradient (DDPG) (*21*), classic imitation learning methods such as behavior cloning (BC) (*22*), state-of-the-art offline learning methods such as decision transformer (DT) (*23*), and several other representative RL and imitation learning methods (*24*–*30*). We also developed an end-to-end method, MoE-GCDT, which uses a goal-conditioned DT for multitask pretraining and adopts the insight of mixture of experts (MoE) from a large language model. It achieved a success rate of more than 95% for the one-step tasks NeedleReach, NeedlePick, and GauzeRetrieve. For multistep tasks, it maintained stable performance with a success rate of 68% for PickAndPlace, 88% for PegTransfer, 37% for MatchBoard, and 80% for NeedleRegrasp, setting a record for state-based methods.

	Tasks	NeedleReach	NeedlePick	GauzeRetrieve	PickAndPlace	PegTransfer	MatchBoard	NeedleRegrasp
Snapshots	Simulator							
	Real world							
State-based*	DDPG (*21*)	100%	79%	59%	14%	44%	9%	6%
	BC (*22*)	100%	23%	12%	17%	56%	0%	12%
	DT (*23*)	88%	92%	70%	44%	79%	19%	70%
	CoL (*24*)	100%	92%	80%	20%	61%	2%	10%
	AWAC (*25*)	90%	34%	38%	24%	19%	12%	20%
	IQL (*26*)	100%	9%	42%	24%	78%	0%	6%
	BET (*27*)	96%	78%	81%	56%	73%	18%	63%
	DEX (*28*)	100%	98%	79%	38%	77%	6%	56%
	RLPD (29)	100%	100%	94%	45%	86%	14%	62%
	RLIF (*30*)	99%	100%	93%	49%	85%	16%	58%
	MoE-GCDT	100%	100%	98%	68%	88%	37%	80%
Vision-based*	ALOHA (*31*)	100%	95%	81%	19%	14%	7%	11%
	Diffusion (*32*)	79%	21%	10%	5%	2%	4%	2%
VPPV (tested in simulator)*		100%	100%	100%	90%	98%	83%	96%
VPPV (tested on dVRK)†		100%	84%	96%	82%	86%	80%	84%

*Tested in SurRoL, and the success rate is reported for 100 independent trials. †Trained in SurRoL and tested on dVRK real-world settings. Success rate results are reported for 50 independent trials.

Fig. 2. Benchmark of different policy learning methods in simulator and comparison with our method on a real-world robot. CoL, cycle of learning; AWAC, advantage weighted actor critic; IQL, implicit Q-learning; BeT, behavior transformer; DEX, demonstration-guided exploration; RLPD, reinforcement learning with prior data; RLIF, reinforcement learning via intervention feedback; GCDT, goal-conditioned decision transformer.

Vision-based policy learning and zero-shot sim-to-real transfer

Recent literature on embodied AI has increasingly explored vision-based policies, aiming to predict robot actions from raw image inputs in an end-to-end manner. This approach is more challenging than state-based policies because it does not rely on perfect state inputs; instead, the learning system must extract robust features from image observations for accurate motion planning. A key challenge for a vision-based policy is overcoming the sim-to-real gap because differences in visual appearance between the simulator and the real world can lead to deviations in latent features, ultimately affecting the accuracy of action prediction. We implemented the latest vision-based policy learning methods of ALOHA (*31*) and Diffusion Policy (*32*) in benchmark implementations. Although these methods could complete skill learning in the simulator (see results in Fig. 2), the policies trained in simulation could not effectively generalize to the dVRK in varied real-world settings.

To address this problem, our proposed vision-based paradigm, VPPV, can achieve zero-shot sim-to-real transfer for different tasks. Details are described in Materials and Methods. The key idea is to robustly bridge high-dimensional visual observations to low-dimensional state representations for RL-based motion planning. Our regressed latent state was designed to be compact and explainable, thus robust to disturbance in visual inputs. For each task, we trained and tested the VPPV paradigm in SurRoL and deployed the learned regression and policy models to real-world dVRK for evaluation (Fig. 2). In simulation, the success rate of VPPV achieved 100% for NeedleReach, NeedlePick, and GauzeRetrieve; 90% for PickAndPlace; 98% for PegTransfer; 83% for MatchBoard; and 96% for NeedleRegrasp. In the real-world setting with dVRK, we tested 50 trials across various scenarios to validate its generalizability (movie S1). The success rate of VPPV achieved 100% for NeedleReach, 84% for NeedlePick, and 96% for GauzeRetrieve for different settings of needle sizes and gauze appearances. For relatively longer-horizon manipulation tasks, we achieved 82% for PickAndPlace (i.e., pick a small stick sized 11.6 mm and place it inside a plate sized 28.6 mm) and 86% for PegTransfer (i.e., pick a small peg with a thickness of 2.0 mm and transfer it to a target). For the more difficult task of MatchBoard, which required the policy to simultaneously support grasping multiple objects, our success rate was 80%. For the

bimanual task of NeedleRegrasp, we achieved a success rate of 84%. Overall, the average performance drop of the sim-to-real transfer was 8% across all seven tasks. The performance of all simulation-trained policies was robust in the real world and stable when varying the positions and orientations of the objects.

Human interface and application to haptic-assisted skill training with AI simulator

We connected the simulator with human interface devices to enable manual teleoperation of the virtual robot in SurRoL, allowing this embodied AI infrastructure to support future needs related to human-robot collaboration research. We had initially developed human-in-the-loop interaction (*33*) using the Touch device (3D Systems Corporation). Here, we further strengthened this feature by adding a previously unreported user interface with the master tool manipulators (MTMs) of the dVRK hardware (Fig. 3A). As a result, users could use dVRK's MTMs to control instruments attached to patient-side manipulators (PSMs) in the simulator for seven–degree of freedom manipulation, including relative displacements, absolute rotations, and gripper opening angles. A communication protocol between MTMs and SurRoL was designed to achieve network connections with low latency at a speed of 25 frames per second (fps) during human interaction.

To demonstrate that the human interface is capable of facilitating downstream research, we conducted a proof-of-concept study using our intelligent simulator for haptic-assisted skill training for surgical education. We converted the dVRK MTM from a passive device to an active guidance device, which can provide force feedback for trainees' path following in skill training (Fig. 3B). The path was automatically generated from the RL-predicted trajectory for an arbitrary scenario. This extended the task diversity and generalizability compared with previous work (*34*) that used predefined paths for haptic-assisted training. In addition, we developed an adaptive shared control mechanism that activated haptic assistance only when trainees deviated from the reference trajectory beyond a specified threshold (Fig. 3C). The force exerted by the MTM was calibrated to be both perceptible and comfortable for trainees. We conducted a user experiment for the new training scheme and observed that the haptic-assistance method can improve learning efficiency (Fig. 3D). We recruited eight users, including four medical and four engineering postgraduate students at CUHK (Chinese University of Hong Kong), all of whom had no prior experience using dVRK or any surgical robots. Participants were randomly divided into two groups (each with two medical and two engineering students), with one group using the haptic-assisted training, whereas the other engaged in self-play practice on the dVRK. After 15 min of training on a peg transfer task, we evaluated their learning outcome by measuring the time they needed to complete the task. Novices trained with haptic assistance completed the task significantly faster than the self-play group (average time of 10.5 ± 2.2 s versus 20.2 ± 1.8 s; $P < 0.01$), indicating that AI-assisted skill training helped the trainees reduce task completion time.

Soft-body simulation and data-driven generation of realistic surgical scenes

Soft-body simulation is an inseparable component of a surgical embodied intelligence environment. However, existing simulators still struggle with simulation speed and realism. We enhanced SurRoL with soft-object simulation using an efficient material point method (MPM). In this approach, a simulated soft object was represented by a collection of material particles (Fig. 4A), with each point movement computed through a hybrid Eulerian-Lagrangian method (*35*). We sped up simulation by adopting the moving least squares MPM (MLS-MPM) (*36*), Taichi parallel programming (*37*), and graphics processing unit (GPU) acceleration to reduce the computation cost. On a standard computer equipped with an NVIDIA GeForce RTX 3090 GPU, we maintained a computation time of less than 100 ms per

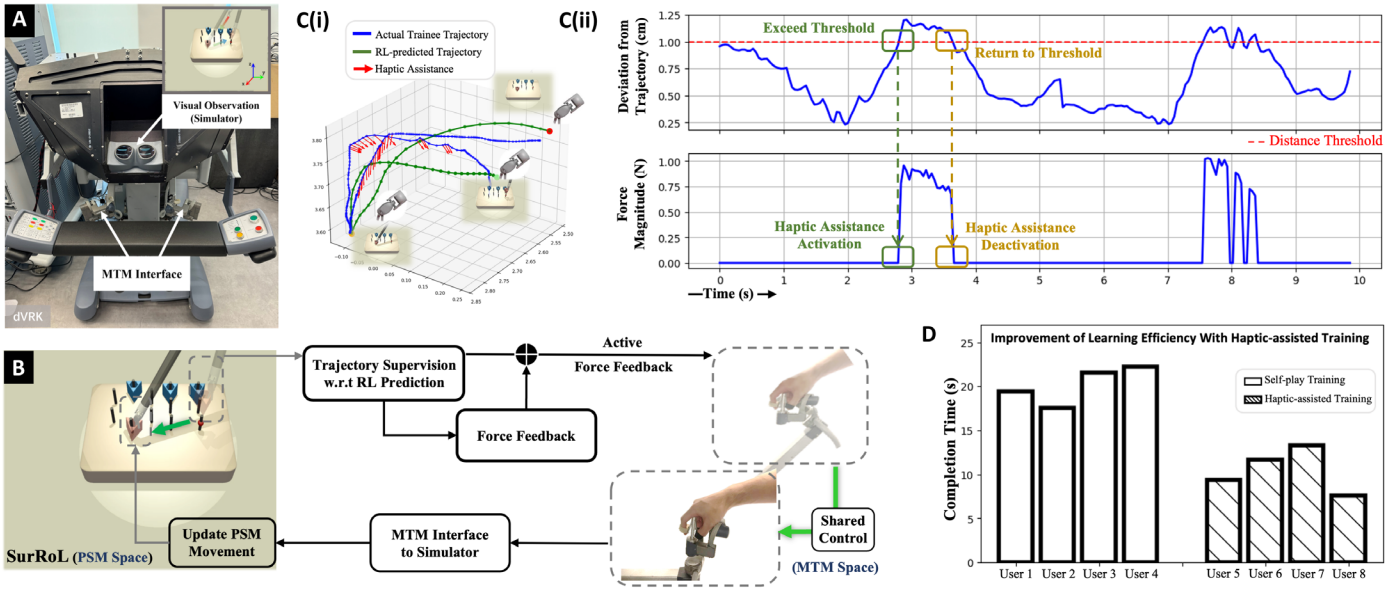


Fig. 3. Illustration and results of SurRoL's human interface and its application to haptic-assisted skill education. (**A**) Simulator connects to MTM of dVRK. (**B**) Our proposed haptic-assisted surgical skill training approach, which uses RL-predicted trajectory to offer intelligent guidance through human-robot shared control. (**C**) (i) Visualization of instrument motion trajectory (peg transfer task) where haptic assistance is activated upon detection of relatively large deviations. (ii) Visualization of the trajectory deviations and the force magnitude over time. (**D**) User study on the improvement of learning efficiency by measuring the task completion time after training.

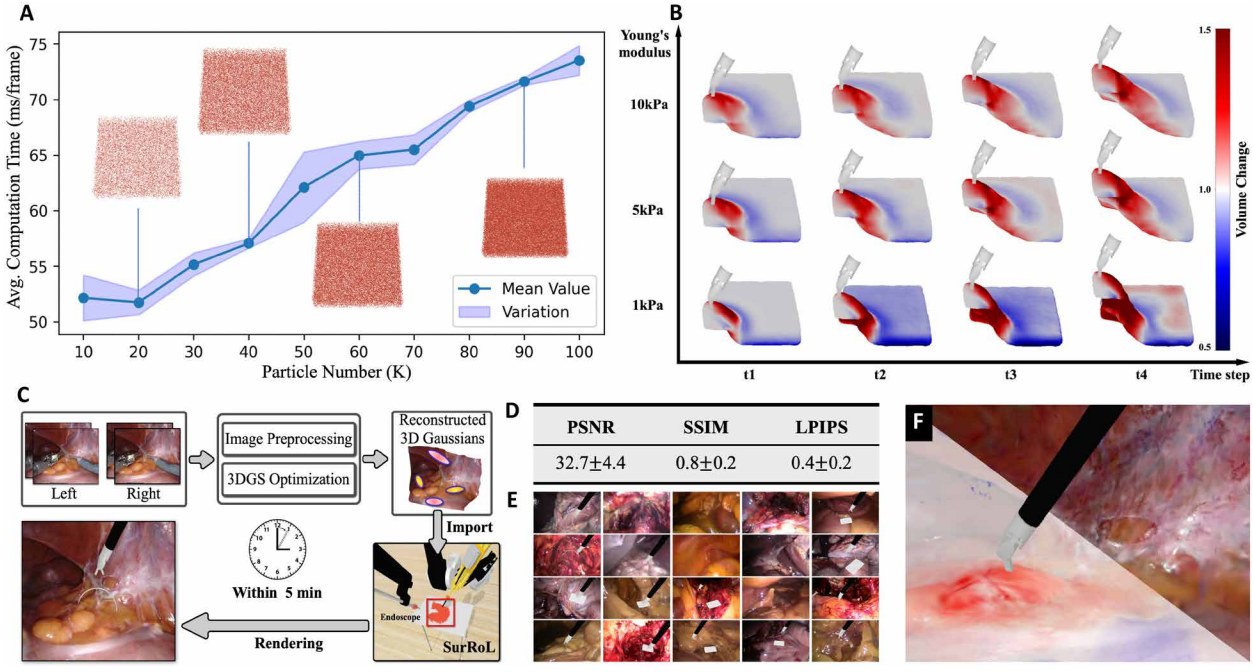


Fig. 4. Results of soft-body simulation and data-driven scene generation. (**A**) Computation time for MPM soft-body simulation with an increase in particle numbers. (**B**) Deformation field of simulated soft tissue with varying Young's modulus. (**C**) Data-driven surgical scene simulation by importing 3D reconstruction results into SurRoL. (**D**) Quantitative evaluation of reconstruction results on metrics of PSNR, SSIM, and LPIPS reported as mean ± SD. (**E**) Visualization of diverse scenes that can be generated from data. (**F**) Generated scene supports tool-tissue interaction with deformation computed from MPM in SurRoL.

frame when the particle count rose from 10,000 to 100,000. The physical parameters associated with particles included Young's modulus, Poisson's ratio, and density, which were used to compute additional properties such as velocity, strain-stress relationships, and deformation gradients during movement. Figure 3B illustrates the deformation field of soft tissue with different stiffnesses, represented by varying Young's modulus values of 1, 5, and 10 kPa. It was observed that lower stiffness resulted in more substantial deformation of the retracted tissue. Moreover, we implemented rigid–soft-body collision detection and interaction for supporting tissue manipulations such as pushing, pulling, and retraction with robot instruments.

To improve the photorealism of tissue appearances for soft-body simulation, we developed a data-driven solution. We reconstructed three-dimensional (3D) surgical scenes from recorded real stereo videos, which can be imported into SurRoL to create virtual scenes (Fig. 4C). This method automatically generated simulated textures from surgical images in less than 5 min, providing a cost-effective alternative to traditional handcrafted simulations. We used the method of 3D Gaussian splatting (3DGS) (*38*) for reconstruction and solved problems such as stereo depth estimation, surgical tool occlusion, and geometrical correction (*39*). Figure 4D reports the reconstruction performance on metrics of PSNR (peak signal-to-noise ratio), SSIM (structural similarity index measure), and LPIPS (learned perceptual image patch similarity) on average for 34 different scenes. Figure 4E shows various examples of the generated virtual scenes in SurRoL, with the flexibility to incorporate assets such as gripper, needle, and gauze. The 3D Gaussian scene representation was compatible with the MPM solver in SurRoL; therefore, the generated scene surface was deformable and allowed for interactions (Fig. 4F). This functionality potentially supported policy learning on various data for soft tissue manipulation tasks in embodied intelligence.

Ex vivo experiments on animal tissue for validation of diverse surgical assistive tasks

To validate the effectiveness of our generalized task automation and assess its feasibility for real surgical applications, we conducted ex vivo experiments for several surgical assistive tasks, including manipulation of the endoscopic camera, needle grasping, gauze picking, soft tissue retraction, and blood vessel clipping. These tasks are typically repetitive and fatiguing for surgeons in general surgery, making them suitable candidates for automation (*40*–*42*). We simulated corresponding environments in SurRoL and used our VPPV paradigm for all of the different tasks. Zero-shot sim-to-real transfer was performed for real-world validation.

To validate that our AI-based solution was independent of robotic hardware, we conducted ex vivo tissue experiments using a different platform, specifically a commercialized surgical robot—the Sentire surgical system developed by Cornerstone Robotics. The observations collected from these ex vivo experiments could help inform the subsequent in vivo animal trial using the same robotic platform. The robot was a teleoperated system for multiport minimally invasive surgery. Its endoscope provided a high-resolution 1080p stereo video stream. The system incorporated a state-of-the-art mechatronics platform, resulting in reduced system latency, faster control response, and more accurate kinematics compared with dVRK. Comprehensive testing had previously been conducted with hundreds of human trials by August 2024, proving the safety and effectiveness of the system. To deploy the AI algorithms, we developed a Python API for the Sentire surgical system based on the

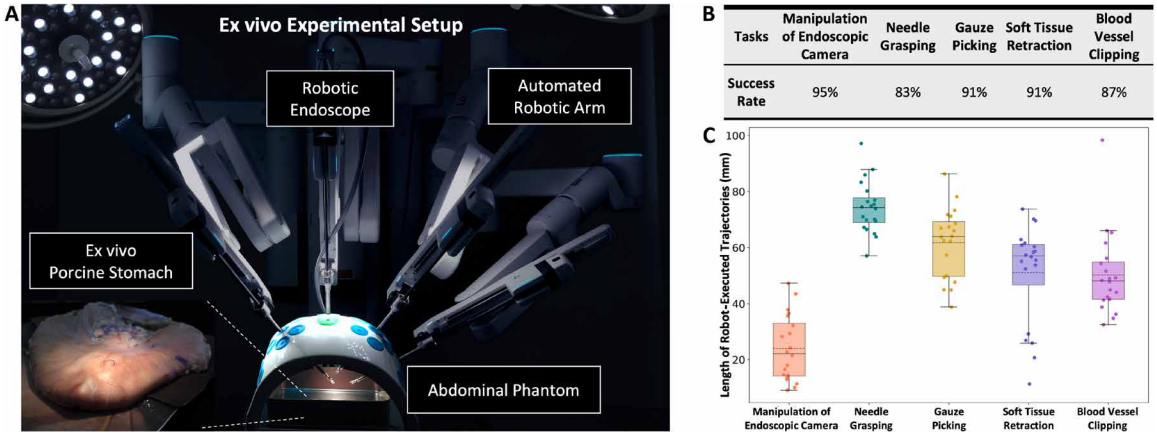


Fig. 5. Ex vivo validation on five surgical assistive tasks using the Sentire surgical system. (**A**) Illustration of the experimental setup using ex vivo porcine tissue. (**B**) Success rate of the five conducted surgical assistive tasks. (**C**) Results of the length of robot-executed trajectories for each task recorded from kinematics data.

Cornerstone research platform definitions, which operated over an Ethernet connection, enabling efficient kinematic data acquisition and transmission of AI-generated control commands.

Specifically, our ex vivo experimental setup (Fig. 5A) used the tissue of a porcine stomach with the left and right gastroepiploic arteries preserved. Three specimens were used in total, each weighing ~0.5 kg and measuring 15 cm in diameter and 3 cm in thickness when flattened. The stomach was rinsed, and a clinician exposed and stabilized the blood vessels. The endoscopic camera was placed 6 to 10 cm from the specimen to replicate a typical working distance in surgery. Each task was performed multiple times, with success rates measured by the number of consecutive attempts until 20 successful trials were achieved (Fig. 5B). In each trial, the orientation of the stomach was randomly set between −180° and 180°, whereas the center of the stomach was translated between −5 and 5 cm relative to the image center under endoscopic view adjustments. We also evaluated the method under environmental variations related to instrument types, object sizes, and scene appearances associated with each task, as well as illumination changes, simulated surgical smoke, and breathing in general. Movie S2 demonstrates results from the ex vivo experiments.

Manipulation of the endoscopic camera

The aim of this task was to automatically adjust the pose of the endoscope, such that the surgical instrument was stably kept in the middle of the field of view to maintain clear operational visibility. This task relies on the surgeon's manual adjustment by switching control between the camera and instruments, which is tedious and interrupts operations. Our automation relied on visual parsing and policy learning with the goal of aligning the instrument tip and scene centroid. Visual servoing was not involved in this task. The overall success rate was 95%, with 20 successful trials of 21 attempts. To demonstrate the robustness (Fig. 6A), we tested different instrument types (i.e., large needle driver, fenestrated bipolar forceps, monopolar curved scissors, and medium-large clip applier) and different instrument positions in the scene, varying brightness and vapor.

Needle grasping

The aim of this task was to automatically grasp a needle that was placed on the surface of the tissue. This action is commonly performed in surgical steps such as anastomosis. Automation of this task used segmentation and depth estimation of the needle, and the regressed state vector represented its 6D pose and a graspable point on the needle. After executing the RL-predicted trajectory to reach the needle, the grasping action adopted a classic controller relying on an accurate hand-eye calibration. Overall, our method achieved a success rate of 83%, with 20 successful trials of 24 attempts. We evaluated its robustness (Fig. 6B) to different needle poses, needle sizes (35, 20, and 15 mm), illumination changes, and smoky scenes.

Gauze picking

The aim of this task was to automatically grasp a piece of gauze within the surgical scene and place it over a blood area. It is often performed by a surgeon's assistant. Autonomous picking required segmentation of the gauze (using a point prompt for the foundation model) and estimation of its distance to the gripper. Given that many positions on the gauze can be picked, the target was set as the gauze center within the context of this task. The overall success rate was 91%, with 20 successful trials of 22 attempts. We validated its robustness (Fig. 6C) with realistic settings, including different shapes of gauze, varying distances between the gauze and bleeding area, blood on or not on the gauze, changes in brightness, and vapor.

Soft tissue retraction

The aim of this task was to automatically grasp and retract the target tissue. This assistive action is typically conducted by the "third robotic hand" to tension the tissue or hold on to the tissue to expose space for operation. Automating it helps to reduce the surgeon's alternating hands to control three robotic arms through the pedal. Our solution started by image segmentation and depth estimation for the tissue area. The graspable position was reached by RL-based motion planning. The policy was trained with the physically based soft-body simulation. A classic controller was adopted to grasp and lift the tissue with a predefined retracting distance. The success rate of this task was 91%, with 20 successful trials of 22 attempts. We tested diverse settings (Fig. 6D), including different retraction points, instrument types, illumination, and vapor. Moreover, we practiced human-robot collaborative tissue manipulation, with one automated arm for tissue retraction and the other two arms performing dissection under human control.

Blood vessel clipping

The aim of this task was to automatically place a clip across a blood vessel to occlude the blood flow. It is commonly performed in surgery to minimize blood loss or facilitate subsequent arterial cutting.

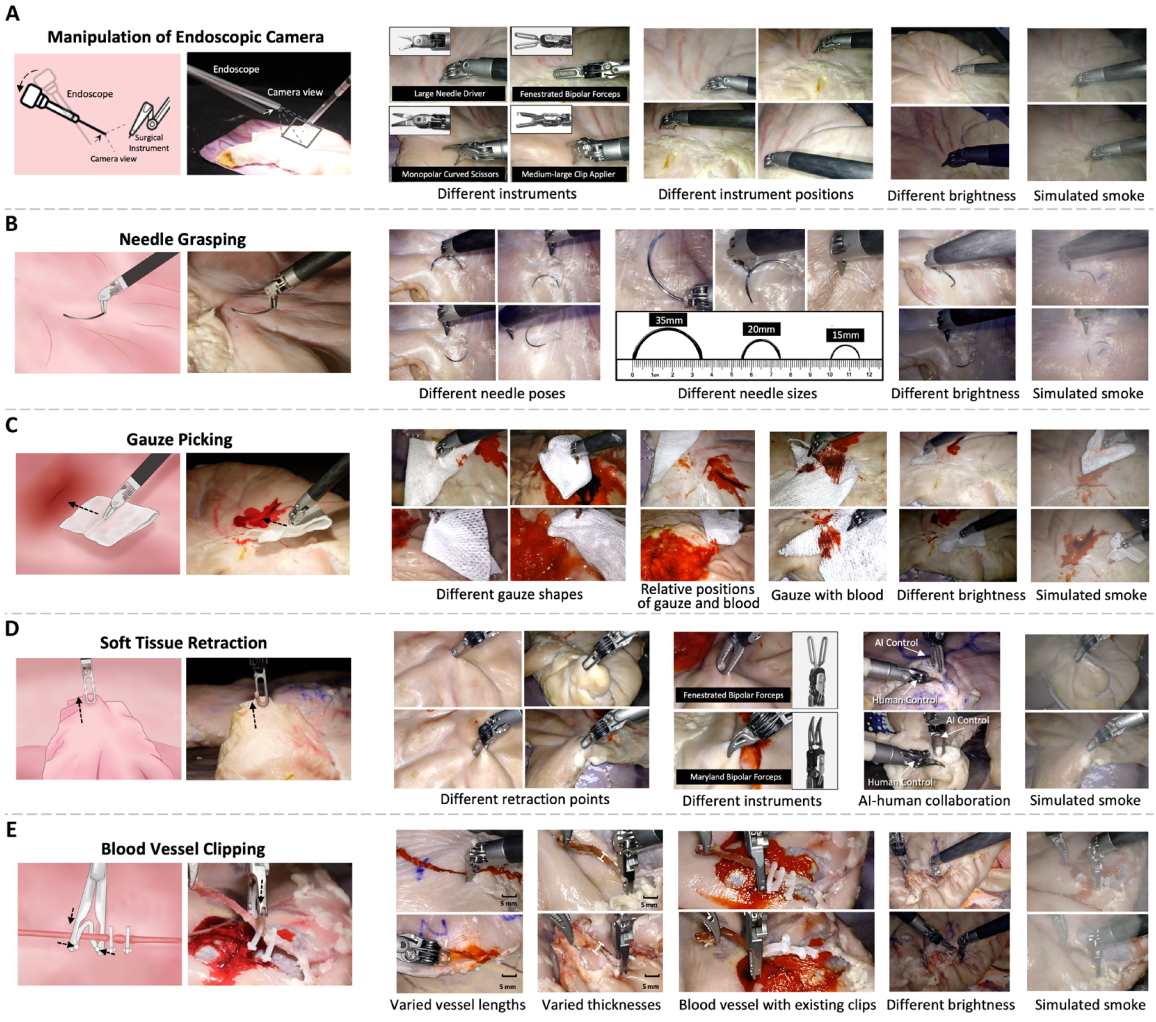


Fig. 6. Diverse settings of real-world scenes on the five tasks in ex vivo experiments. (**A**) Manipulation of endoscopic camera. (**B**) Needle grasping. (**C**) Gauze picking. (**D**) Soft tissue retraction. (**E**) Blood vessel clipping.

Typically, the task relies on a surgical assistant to manually apply the clip using a nonrobotic clip applier at the surgeon's request. Automation of this task involved visual parsing and regression of the blood vessel's state, enabling effective motion planning for the robotic clip applier. When approaching the clipping position, a visual servoing controller was used to precisely execute the final-step clipping. The overall success rate of this task was 87%, with 20 successful trials of 23 attempts. We tested robustness (Fig. 6E) regarding variations in real-world settings and showed that the method can handle vessels with different diameters and consecutively put clips on the same vessel.

For the five surgical assistive tasks, we recorded instrument kinematics data from the robot to evaluate the efficiency of the RL-learned policy by analyzing the trajectory lengths (Fig. 5C). For the manipulation of the endoscopic camera task, the endoscope moved within a small range of 24.0 ± 11.4 mm, which was sufficient for adjusting the surgical field of view. The executed trajectory lengths for the other tasks varied on the basis of the distance between the instrument and target object: 74.4 ± 9.1 mm for needle grasping, 61.8 ± 12.1 mm for gauze picking, 51.1 ± 17.4 mm for soft tissue retraction, and 50.3 ± 14.4 mm for blood vessel clipping. The average deviations of the executed trajectories from the expected optimal trajectories (i.e., the shortest paths between starting and ending positions) were 3.3, 9.1, 8.3, 6.1, and 6.0 mm, respectively. These results indicated that the RL-predicted trajectories effectively reached the target objects with minimal redundant movement relying on the accurate scene understanding.

In vivo study of a live porcine model under supervised autonomy

We conducted an in vivo study using a live porcine model to evaluate the performance of various automated tasks in laparoscopic robot-assisted surgery (Fig. 7). Our solution only used visual feedback from the robotic stereo camera without any additional sensors. Three of the five assistive tasks were selected for live-animal trials (i.e., gauze picking, soft tissue retraction, and blood vessel clipping) because they are relatively more complex and involve major scene differences between in vivo and ex vivo, which cannot be fully mimicked outside the body. Movie S3 demonstrates the results of the in vivo study.

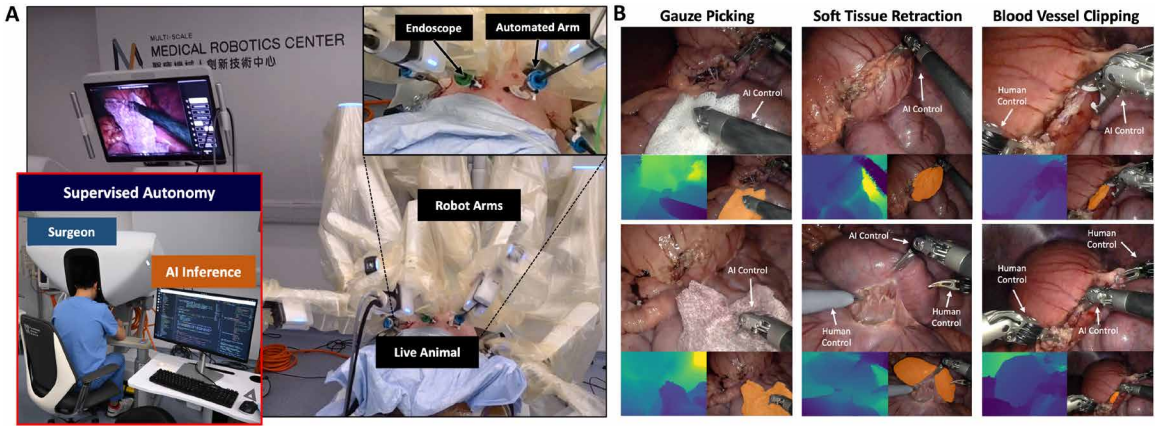


Fig. 7. In vivo validation with a live-animal trial under supervised autonomy. (**A**) In vivo experimental setup in the hybrid operating room at Multi-Scale Medical Robotics Center in Hong Kong. (**B**) Illustration of the automated surgical tasks including gauze picking, soft tissue retraction, and blood vessel clipping. Each task shows two example scenes with endoscopic image (top), estimated depth map (bottom left), and target object segmentation (bottom right).

When implementing the in vivo validation, we adopted the concept of "supervised autonomy" (*43*), where the task autonomy was executed under close supervision of a surgeon who was ready to take over when needed. The trial was conducted with the same robot system, using the hybrid operating room in the Multi-Scale Medical Robotics Center of Hong Kong with ethics approval. We packaged our VPPV algorithms as end-to-end software, and its connection to the robotic API was the same as that of the ex vivo experiment. One healthy female pig of ~30 kg was used as the model for robot-assisted gastric mobilization performed by a certified gastrointestinal surgeon. Under general anesthesia in reverse Trendelenburg position, one 10-mm camera port was placed around the umbilicus with the introduction of capnoperitoneum. Subsequent working ports were placed in both the left and right upper abdomen, including three 8-mm ports for robotic instruments and one 12-mm port for the assistant surgeon. Each port was placed at least 8 cm apart to ensure adequate space for robotic instrument manipulation.

For the gauze picking scenario, we put the gauze near the pig's stomach and tested different gauze positions, sizes, and environmental illuminations. We completed a total of six trials until one failure was encountered, yielding a success rate of 83%, which was comparable to ex vivo performance. In vivo testing presented greater challenges, even for this seemingly simple task. When the gauze was stained with blood, its color closely resembled that of the surrounding soft tissue, and it often became partially stained once placed in the body. This necessitated that vision algorithms rely on high-level scene understanding rather than solely using color-based features to identify the gauze and its boundaries, underscoring the importance of using vision foundation models for robot perception. In addition, the respiratory movement of the animal could be distinctly observed in the visual input. We set the picking point at the center of the gauze to allow tolerance to such dynamics during the grasping action. In vivo tissue was soft, which increased the likelihood that the gripper may unintentionally grasp both the thin gauze and its underlying soft tissue, leading to safety risks. A potential solution could be setting a safety checkpoint to monitor the depth of the instrument's tip using robotic kinematics.

For the soft tissue retraction scenario, the console surgeon would first outline the target tissue that was deemed safe and feasible for retraction on the laparoscopic view to achieve the purpose of supervised autonomy. Then, the robot was allowed to automatically drive the instrument to access and retract the tissue by a distance specified by the human. To enhance the reliability of visual understanding in complex in vivo scenes, we used two positive point prompts to FastSAM (*44*) for the designated area, along with three negative prompts to robustly exclude irrelevant regions from the segmentation results. Given that different point prompts might generate the same segmentation results, the solution was not sensitive to variations in different surgeon inputs. We tested various soft tissue scenarios with differences in appearance, orientation, and light reflections on the tissue surface. The success rate was 77% with 10 of 13 trials. Failures occurred in cases where the gripper did not adequately press against the tissue surface, resulting in the tissue slipping from the opening gripper. We also explored human-robot collaboration, where the third robotic hand automatically retracted the small bowel to create space for mesenteric dissection by two other instruments controlled by the console surgeon. Through automated movement of the third robotic arm, the surgeon could focus on his two working arms, which could potentially increase the efficiency and reduce stress for the surgeon. In the animal trial, the quality of the view created by the automatic tissue retraction was deemed satisfactory for dissection.

For the blood vessel clipping scenario, the in vivo situation was distinct from the ex vivo situation. The unseen appearances of the blood vessels and the high precision required for this manipulation task made it the most difficult one in our animal study. In our setting of supervised autonomy, the surgeon first dissected the surrounding tissue and then stabilized the pig's right gastroepiploic artery using two instruments. A robotic arm then autonomously placed a clip across the vessel under close human monitoring. The robot successfully performed four clippings of six attempts, resulting in a success rate of 67%. On the basis of our observations, several key factors should be addressed to further optimize autonomy for this task. Respiratory movement complicated vessel localization and robot planning because of the dynamics of surrounding soft tissues. Maintaining the accuracy and stability of robotic hand-eye calibration was essential, which otherwise would cause failures, given that this task was not fault tolerant (i.e., the vessel had a diameter of ~5 mm, whereas

the clip measured only 10 mm). Furthermore, surgeons usually need to see both ends of the clip to confirm that it fully crosses the vessel before securely closing it. Our automated execution did not incorporate this verification step, which may pose a risk of inadequate clipping. This raised our awareness of the alignment between robotic manipulation and human behavior for safety considerations, which was as important as the last-step success of the task.

For all of the in vivo experiments, the implemented neural networks achieved a real-time inference speed. Specifically, the AI algorithms were deployed on a standard workstation equipped with an Intel i7-12700 central processing unit (CPU) and a single NVIDIA GeForce RTX 3090 GPU. For the image perception module, the time to process one frame was 40 ms for target object segmentation and 300 ms for stereo depth estimation. These two deep learning models could be computed in parallel with the same visual inputs from the endoscope. The policy prediction module was faster, and it took 7 ms to estimate the scene states and use them to predict actions for robotic end effectors. The final visual servoing low-level controller took 2 ms. Overall, our end-to-end deployment of the task automation paradigm can reach a running speed of 3 fps from the software aspect. The transmission of the AI-generated control commands to the robot platform's API was not set as fast as its raw speed. On average, the completion time was 19.2 s for the gauze picking task, 17.0 s for the soft tissue retraction task, and 8.3 s for the blood vessel clipping task. We slowed down the actual step size of robot movement for safety considerations and to allow close human monitoring under the implementation of supervised autonomy. With these analyses of time performance, the inference speed of deep neural networks (even for vision foundation models) was not as time consuming as expected. Therefore, running time efficiency should not be a bottleneck for future development of learning-based automation for surgical robots.

DISCUSSION

This paper introduces a surgical embodied intelligence platform and uses it to develop a learning-based paradigm for achieving generalized task autonomy on general surgical robots. The applicability of this approach was validated in real-world settings with dVRK and a commercialized robotic surgical system. Our learning-based solution demonstrated generalizability in two key aspects: task level and scene level. First, the VPPV is a unified learning paradigm that can be applied to various tasks, including seven game-based skill training tasks and five surgical assistive tasks, with minimal task-specific design. This versatility is enabled by embodied AI, which learns data-driven control policies through RL. Second, the simulation-trained policy can be directly transferred to real-world scenarios in a zero-shot manner, exhibiting robustness to environmental changes and complexities in in vivo surgical scenes. This capability stems from the incorporation of vision foundation models in the VPPV, which enables object-agnostic understanding of real-time image observations from the robotic endoscope.

Surgical embodied intelligence holds as much potential as embodied AI in the broader field of robot learning for daily tasks. However, its advancement has not progressed as rapidly as seen in general applications. A primary obstacle is the lack of ready and available open-source software infrastructure to support such research. Nevertheless, the community has a strong history of infrastructural joint efforts. The dVRK exemplifies an effective hardware infrastructure that has supported surgical robotics research and accelerated studies over the past 2 decades. Today's wave of embodied AI infrastructure is likely to inspire similar community efforts. A number of simulators (*9*–*15*) of this kind have already been developed and are actively maintained, including our own. Engaging more university groups and industry companies will further help build a long-term supporting ecosystem. The platforms will attract a larger community of researchers and expand the scope of research topics. Looking ahead, surgical robotics holds immense potential for advancement by building on embodied AI. Researchers can explore various avenues, including automating a richer diversity of surgical tasks, enhancing the efficiency of learning algorithms to work with limited medical data, and implementing safe RL techniques tailored for surgical contexts. Some low-hanging fruit, such as AI-assisted surgical skill training, as demonstrated in our proof-of-concept study, can be practical use cases for translational research.

Our experiments in this work concentrated on surgical assistive tasks consisting of only one or two steps. Future research will explore more complex, longer-horizon tasks through scalable control policy learning and chaining of learned basic skills. This aligns with the broader goal of enhancing surgical autonomy from level 2 to level 3 (*3*). However, this transition is challenging and unlikely to be achieved by relying solely on AI. Learning methods excel in image perception and high-level path planning, whereas classic control methods offer superior accuracy and reliability for low-level manipulation. The path to higher-level autonomy is likely to leverage the strengths of both. Safety and precision remain the ultimate criteria for evaluating an autonomous surgical robot, regardless of the techniques used. We incorporated this insight into our designed VPPV paradigm, ensuring its flexibility for future extension. The visual parsing module can be updated to incorporate the latest vision foundation models and integrate automatic prompt prediction for segmentation. The perceptual regressor and policy learning modules can be enhanced with advanced machine learning algorithms such as large language model–based approaches for more effective motion planning. The visual servoing module can integrate classic methods with model-based controllers to ensure predictability and stability for the final step of task execution. These modules will be independently or jointly refined, fostering a seamless integration of learning-based and classic methods to achieve higher levels of autonomy.

For the in vivo trial on a live animal, we adopted the concept of supervised autonomy. This approach allows the surgeon to delegate tasks to the AI controller while remaining ready to take over when necessary. In theory, this would reduce the stress and workload of the surgeon by adding an automated assistant, which could provide consistent focused camera control, tissue countertraction, and so on. Although this can be implemented, our observations revealed more intriguing scenarios in practice. For instance, when a third robotic arm is performing automated assistive tasks, such as tissue retraction, the surgeon's mind and hands could be focusing on manipulating the two other robotic arms during the surgery. This creates a collaborative context involving a human and an intelligent robot. In particular, in the event of an AI system failure, the surgeon may not have the cognitive bandwidth to detect a safety issue with the automated system and respond promptly. This opens up unexplored questions in the field about how to optimize ergonomics and maintain human cognitive performance in contexts of supervised autonomy after one or even more robotic arms are controlled by AI systems. To overcome such shortcomings, solutions would need to

be developed, such as the inclusion of a verbal command system that could allow instantaneous surgeon takeover or cessation of the ongoing automated task. The human-AI interface could also be deepened similarly, when minor modifications of the automated task such as retraction direction could be achieved through verbal commands.

We present failure cases in movie S3 for the in vivo tasks. Our analysis shows that although the predicted trajectory can guide the instrument in a reasonable direction toward the target object, the final-step manipulation lacks precision because of several practical factors. First, when transitioning from the RL policy to the visual servoing, the instrument is already positioned close to the target object. It may block the endoscope view of the intended grasping point, affecting the depth estimation. Second, the localization of the tentative grasping point for visual servoing is imperfect. Although we used an AI foundation model, achieving pixel-perfect precision in depth maps remains challenging, especially in unseen in vivo situations. This imprecision can lead to misalignment of the estimated grasping point, causing the gripper to approach the target at a suboptimal angle. The resulting contact can trigger unpredictable motion and deformation because of the elasticity of in vivo tissues. Such dynamics are not considered in our system, and how to address them remains an open question for future work.

The concept and fast-advancing techniques of embodied AI will revolutionize not only laparoscopic robot-assisted surgery but also other types of medical robotics as long as necessary developmental infrastructure is in place. A number of assistive tasks would be technically ready for consideration of in vivo trials at a pace faster than anticipated. These advanced "surgical copilots" could understand human needs, offer real-time help, or even take on complex tasks autonomously. Because such surgical copilots evolve, establishing comprehensive ethical guidelines and regulatory frameworks through collaborative efforts among clinicians, researchers, industry stakeholders, and regulatory bodies will be crucial. The roadmap will push the boundaries of AI-powered surgical autonomy while ensuring safe and responsible implementation, ultimately improving patient care.

MATERIALS AND METHODS

SurRoL

Our developed open-source simulator SurRoL (*15*) is designed for surgical embodied intelligence and is compatible with the dVRK platform. It has been listed as a machine learning tool on the GitHub of the dVRK software ecosystem (*45*). SurRoL consists of three main components. The first component is the digital twin of the dVRK robot, which involves simulated piecewise objects and robot articulation using Unified Robot Description Format (URDF), including PSMs, endoscopic camera manipulator (ECM), and end effectors that can open and close. It provides closed-form solutions for both forward and inverse kinematics, supporting precise robot manipulation. It enables human interface through devices including Touch (3D Systems Corporation) and dVRK's MTMs, facilitating the collection of demonstration data and haptic interactions. It also enables executable motion trajectories that can be transitioned from the simulator to the real-world robot. The second component is interactive simulation with rigid and soft bodies. The physics engine uses Bullet for collision detection and rigid-body dynamics, supporting environmental state interactions. It incorporates a Taichi-based plug-in that supports the MPM for soft-body simulation and interaction with rigid bodies. The third component is a library for RL and imitation learning methods. Its infrastructural support for robot learning and a benchmark of built-in learning-based methods are presented in Results. SurRoL is a highly dedicated simulator for surgical embodied intelligence among all available options in the field.

Vision-based learning for surgical task automation

Our VPPV learning framework is composed of a visual parsing module for endoscopic image perception, a perceptual regressor for image-based environmental state regression, a policy learning module for robot motion planning, and a visual servoing controller to lastly execute the manipulation task. Details of each module are as follows.

Visual parsing

Given the endoscopic image, we used the method of FastSAM (*44*) to obtain the object segmentation mask. The Segment Anything Model (SAM) (*46*) is a foundation model trained on large-scale datasets for zero-shot image segmentation. Despite its powerful segmentation capabilities, SAM's inference speed cannot be real time. Its improved version can be 50 times faster while maintaining comparable performance. FastSAM (*44*) accepts multiple types of prompts to generate segmentation results. For our task, we provided point prompts to obtain the object mask. In addition, the visual parsing module needed to extract object depth information in 3D. The endoscope in our surgical robot system was equipped with a stereo camera, allowing us to derive depth using stereo images and camera intrinsic. Specifically, we used the state-of-the-art depth estimation method IGEV (iterative geometry encoding volume) (*47*), which uses iterative geometry encoding to refine depth estimation, enhancing generalization and robustness in both textureless and intricate regions. This makes it suitable for dynamic surgical environments.

Perceptual regressor for zero-shot sim-to-real transfer

We designed a perceptual regressor to map the visual parsing results to corresponding environmental states to learn a vision-based policy. Instead of using a latent embedding from the raw endoscopic image, our perceptual layer leverages high-level abstracted representations, i.e., segmentation and depth maps. Let I be the raw endoscopic image. Its segmentation map I_s provides rich semantic information and instance-specific cues, whereas its depth map I_d offers essential 3D features, reducing distance ambiguity and enhancing contextual understanding. We used a pretrained ResNet-18 (*48*) to extract semantic and geometric embeddings from I_s and I_d. These embeddings were then fused using a fully connected layer to produce the final state representation. The output of the perceptual layer is a 9D state vector, which includes the 3D position and 3D orientation of the contact point on the object and the relative position between the target position and the instrument end effector in camera coordinates. To train the perceptual layer, we sampled 12,000 pairs of synthetic data in the SurRoL simulator, each consisting of a segmentation map, a depth map, and the corresponding ground-truth state parameters of the environment. The perceptual layer was optimized using a mean square error loss, which measured the difference between the predicted and ground-truth states. Domain randomization was used during model training. During real-world deployment, given an endoscopic image, we estimated its segmentation and depth maps using the visual parsing module. These estimated maps were then fed into the trained perceptual layer to recover the underlying environment state. Empirical results showed that abstracting the raw endoscopic image with semantic and depth information effectively handled the sim-to-real gap. Given the reliable perception performance of vision foundation

models, the obtained state parameters exhibited high robustness to various real-world environmental changes, such as variations in objects and lighting conditions.

Policy learning

On the basis of the environmental states estimated by the perceptual layer, we trained the vision-based control policy with RL. Specifically, for all game-based training tasks and the surgical assistive tasks of needle grasping, gauze picking, soft tissue retraction, and blood vessel clipping, the policy's input state s contained PSM pose information, regressed environmental state, and a 3D goal from the endoscopic image. For the manipulation of endoscopic camera task, the state s contained the position of the targeted PSM tip relative to the image center. We used the DDPG algorithm with an actor-critic framework (*49*). The actor network $\mu(s)$ was parameterized by θ^{μ}, which is responsible for generating action a on the basis of the current state of s. It uses a multilayer perceptron architecture, which takes the state vector as input and outputs a continuous action. The hidden layers of the actor consist of 256 neurons, allowing the network to capture complex mappings from states to actions. The critic network evaluates the quality of the predicted actions by estimating the action-value function $Q(s,a)$. This network produces a scalar output representing the expected return for the given state-action pair.

The critic network $Q(s,a)$ updates its weights θ^{Q} by minimizing the loss function, which is defined on the basis of the temporal difference error. The target value y_t for the critic network and the loss function L at time step t are defined as follows

$$y_t = r_t + \gamma Q\left(s_{t+1}, \mu\left(s_{t+1} \mid \theta^{\mu}\right) \mid \theta^{Q}\right) \tag{1}$$

$$L\left(\theta^{Q}\right) = E_{s_t \sim \eta(\theta^{\mu}), a_t \sim \theta^{\mu}}\left[\left(Q\left(s_t, a_t \mid \theta^{Q}\right) - y_t\right)^2\right] \tag{2}$$

where r_t is the reward received after taking an action a_t in state s_t, γ is the discount factor that quantifies the difference in importance between future and present rewards, E denotes the expectation operator, and $\eta(\theta^{\mu})$ is the discounted state visitation distribution for a policy specified by the actor. The actor network updates its weights θ^{μ} using the sampled policy gradient method (*50*), which uses the chain rule to backpropagate gradients from the critic network to the actor network, effectively training the policy to produce actions that maximize the expected reward.

Our method uses a sparse reward structure, evaluating whether the achieved goal matches the desired goal. It returns a float value of −1.0 for failure and 0.0 for success, thereby encouraging the agent to focus on goal attainment while discouraging suboptimal behaviors. This immediate reward provides distinct feedback and facilitates effective learning. During policy training, we froze the perceptual layer and concentrated on training the policy network. We introduced random noise to the depth and segmentation maps at each step of the agent's interaction with SurRoL to simulate real-world variations. For depth maps, we applied Gaussian noise with an SD randomly sampled from 0 to 3 mm to each pixel. For segmentation masks, each pixel in the mask was randomly excluded with a probability sampled from 0 to 0.3. Training the policy with noisy inputs enabled it to adapt to inaccuracies in state representation, thereby enhancing its generalization capability.

Visual servoing controller for last-step manipulation

We used a visual servoing controller to execute the final manipulation for surgical task automation. This algorithm relies on visual feedback of segmentation and depth, enabling stable tool positioning in dynamic environments. We used a position-based visual servoing control strategy. First, we identified the target goal from the endoscopic image and estimated its real-time 3D position in the camera coordinate system as P_t^{cam}. To control the PSM arm of the surgical robot, we transformed the target object's position from the camera coordinate system to the PSM coordinate system using a transformation matrix $T_{\text{cam}}^{\text{psm}}$ obtained via hand-eye calibration. The transformation is given by $P_t^{\text{psm}} = T_{\text{cam}}^{\text{psm}} P_t^{\text{cam}}$, where P_t^{psm} denotes the target object's position in the PSM base coordinate system. The robot's kinematics provide the current position of the PSM's end effector $P_{\text{cur}}^{\text{psm}}$ in real time. We computed the error between the target position and the current end-effector position as $e(t) = P_t^{\text{psm}} - P_{\text{cur}}^{\text{psm}}$. This error $e(t)$ is fed into the surgical robot arm controller, which generates a control signal to drive the PSM arm toward the target position. Because the error decreases, the PSM arm approaches the desired state. Because the error decreases and converges below the inherent accuracy limit of the control system, the PSM arm reaches the desired state and performs the manipulations to complete the task.

Soft-tissue simulation and data-driven scene realism

Using the dVRK-compatible embodiment environment and the built-in physics engine of SurRoL, we developed a particle-based method for soft-body simulation. Initially, the soft tissue was discretized into a set of Lagrangian particles within SurRoL. Each particle, referred to as a material point, was assigned a set of physical attributes for accurate simulation. We then modeled the tool-tissue contact point using a collision detection module and simulated tissue deformation by propagating particle motion from the contact point to the surrounding tissue particles, adhering to principles of continuum mechanics. Increasing the number of particles enhanced simulation precision but also extended computation time. To improve simulation efficiency, we used MLS-MPM (*36*). This method constructs a spatially regular grid around the soft tissue and accelerates computation through an effective stress divergence discretization technique.

Building on our efficient MPM-based soft-tissue simulation method, we further developed a data-driven approach for surgical scene simulation. We used Gaussian splatting (*38*) to reconstruct surgical scenes from endoscopic videos. The surgical scene was modeled as a set of 3D Gaussians, each parameterized by (μ, Γ, c, and σ), representing the Gaussian center, covariance matrix, color, and opacity, respectively. Using this Gaussian-based scene representation, we adopted a differentiable tile rasterizer to render endoscopic images. The scene reconstruction was achieved by optimizing all Gaussian parameters to minimize the difference between the rendered images and those sampled from the endoscopic video. To enhance the geometric accuracy of the reconstructed scene, we rendered depth maps and constrained their values with depth maps estimated from the endoscopic images. To address tool occlusion, we segmented the tool mask for each video frame and applied a mask-guided sampling technique, using only occlusion-free pixels to guide the optimization of the Gaussian-based scene representation.

After 3DGS-based surgical scene reconstruction, we then applied the MPM-based soft tissue simulation to the reconstructed scene. Specifically, we loaded the reconstructed 3D Gaussians into our SurRoL simulator and took each individual Gaussian as one material point. We used the covariance matrix Γ to control the spatial shape of 3D Gaussian. Given the tool-tissue contact point detected in SurRoL, the scene deformation map ϕ would be estimated using the following first-order approximation

$$\phi_p(X,t) = x_p + F\left(X - X_p\right) \quad (3)$$

where p denotes one reference point used to estimate the deformation, x_p denotes the position of this reference point in the world space, X_p denotes its corresponding position in the material space, and X refers to the position of a neighboring point relative to p in the material space. On the basis of this approximation, the update rule of the covariance matrix could be derived as $\Gamma' = F\Gamma F^T$. The deformation gradient F is updated in each MPM step by $F'_p = \left(1 + \Delta t C_p\right)F_p$, where C_p is the velocity gradient.

Integration of AI algorithms into the robotic platform

The developed AI algorithms are independent from robotic platforms but require some engineering work to connect the AI computer with the robot. There is no significant difference when implementing the VPPV paradigm in different settings (laboratory, ex vivo, and in vivo) and the robotic systems as long as the robot's API is available. We developed a general-purpose Python API with a series of functions to integrate AI algorithms with the Sentire surgical system. These included video_capture() for image data collection and the Collaborative Robotics Toolkit (*51*) commands measured_cp() and measured_jp() for reading robot proprioceptive data and move_cp() and move_jp() for executing predicted robot actions. These functions ensure seamless operation by connecting the output of one module to the input of the next. We used an ultrafast, high-definition stereo endoscope developed by Cornerstone Robotics to capture 1080p, 60-fps stereo RGB video of the surgical field. The integrated light source provided adjustable brightness for optimal visibility. The video stream was extracted via high-definition multimedia interface (HDMI) and converted using a universal serial bus (USB) adapter (U3SDH, ACASIS, China) before being introduced into the VPPV framework. Endoscopic images were processed by the visual parsing module to describe the surgical environment. The scene understanding data were then used to predict the current environmental state. The perceptual outputs, combined with the robot's proprioceptive data, formed the inputs for the policy model that predicted the desired robot actions. After executing the predicted actions, the endoscope captured new images to initiate the next cycle. This integrated system of the camera, AI algorithms for perceptual and policy models, and robotic arms ensures effective interaction and smooth operation in a dynamic surgical environment.

Statistical analysis

We evaluated the success rate of all game-based, ex vivo, and in vivo experiments by calculating the proportion of successful trials relative to the total number of trials. Tables and results for other metrics are reported as mean (±SD) unless specified differently. For the evaluation of AI-assisted surgical skill training, a *t* test was conducted for the two groups of users, with a *P* value smaller than 0.01 considered statistically significant.

Supplementary Materials

The PDF file includes:
Materials and Methods
Figs. S1 to S5
References (*52*–*54*)

Other Supplementary Material for this manuscript includes the following:
Movies S1 to S3
MDAR Reproducibility Checklist

REFERENCES AND NOTES

1. P. E. Dupont, B. J. Nelson, M. Goldfarb, B. Hannaford, A. Menciassi, M. K. O'Malley, N. Simaan, P. Valdastri, G.-Z. Yang, A decade retrospective of medical robotics research from 2010 to 2020. *Sci. Robot.* **6**, eabi8017 (2021).
2. Intuitive Surgical Inc., Annual report 2024 (2024); https://isrg.intuitive.com/static-files/500ff989-ad91-4b32-a59e-f94a34d75997.
3. P. Fiorini, K. Y. Goldberg, Y. Liu, R. H. Taylor, Concepts and trends in autonomy for robot-assisted surgery. *Proc. IEEE* **110**, 993–1011 (2022).
4. H. Saeidi, J. D. Opfermann, M. Kam, S. Wei, S. Léonard, M. H. Hsieh, J. U. Kang, A. Krieger, Autonomous robotic laparoscopic surgery for intestinal anastomosis. *Sci. Robot.* **7**, eabj2908 (2022).
5. J. Gu, F. Xiang, X. Li, Z. Ling, X. Liu, T. Mu, Y. Tang, S. Tao, X. Wei, Y. Yao, X. Yuan, P. Xie, Z. Huang, R. Chen, H. Su, "ManiSkill2: A unified benchmark for generalizable manipulation skills," in *Proceedings of the International Conference on Learning Representations (ICLR)* (ICLR, 2023), pp. 1–30.
6. S. Luo, M. Jiang, S. Zhang, J. Zhu, S. Yu, I. Dominguez Silva, T. Wang, E. Rouse, B. Zhou, H. Yuk, X. Zhou, H. Su, Experiment-free exoskeleton assistance via learning in simulation. *Nature* **630**, 353–359 (2024).
7. V. Makoviychuk, L. Wawrzyniak, Y. Guo, M. Lu, K. Storey, M. Macklin, D. Hoeller, N. Rudin, A. Allshire, A. Handa, G. State, "Isaac Gym: High performance GPU-based physics simulation for robot learning," in *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)* (NeurIPS, 2021), pp. 1–12.
8. A. Szot, A. Clegg, E. Undersander, E. Wijmans, Y. Zhao, J. Turner, N. Maestre, M. Mukadam, D. Chaplot, O. Maksymets, A. Gokaslan, V. Vondrus, S. Dharur, F. Meier, W. Galuba, A. Chang, Z. Kira, V. Koltun, J. Malik, M. Savva, D. Batra, "Habitat 2.0: Training home assistants to rearrange their habitat," in *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)* (NeurIPS, 2021), pp. 251–266.
9. F. Richter, R. K. Orosco, M. C. Yip, Open-sourced reinforcement learning environments for surgical robotics. arXiv:1903.02090 [cs.RO] (2019).
10. E. Tagliabue, A. Pore, D. Dall'Alba, E. Magnabosco, M. Piccinelli, P. Fiorini, "Soft tissue simulation environment to learn manipulation tasks in autonomous robotic surgery," in *Proceedings of the International Conference on Intelligent Robots and Systems (IROS)* (IEEE, 2020), pp. 3261–3266.
11. P. M. Scheikl, B. Gyenes, R. Younis, C. Haas, G. Neumann, M. Wagner, F. Mathis-Ullrich, LapGym - An open source framework for reinforcement learning in robot-assisted laparoscopic surgery. *J. Mach. Learn. Res.* **24**, 1–42 (2023).
12. V. M. Varier, D. K. Rajamani, F. Tavakkolmoghaddam, A. Munawar, G. S. Fischer, "AMBF-RL: A real-time simulation based reinforcement learning toolkit for medical robotics," in *Proceedings of the International Symposium on Medical Robotics (ISMR)* (IEEE, 2022), pp. 1–8.
13. S. Schmidgall, A. Krieger, J. Eshraghian, "Surgical Gym: A high-performance GPU-based platform for reinforcement learning with surgical robots," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)* (IEEE, 2024), pp. 13354–13361.
14. Q. Yu, M. Moghani, K. Dharmarajan, V. Schorp, W. C. H. Panitch, J. Liu, K. Hari, H. Huang, M. Mittal, K. Goldberg, A. Garg, "ORBIT-Surgical: An open-simulation framework for learning surgical augmented dexterity," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)* (IEEE, 2024), pp. 15509–15516.
15. 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 *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (IEEE, 2021), pp. 1821–1828.
16. Y. Ou, S. Zargarzadeh, M. Tavakoli, Robot learning incorporating human interventions in the real world for autonomous surgical endoscopic camera control. *J. Med. Robot. Res.* **8**, 2340004 (2023).
17. R. Bendikas, V. Modugno, D. Kanoulas, F. Vasconcelos, D. Stoyanov, Learning needle pick-and-place without expert demonstrations. *IEEE Robot. Autom. Lett.* **8**, 3326–3333 (2023).
18. Z. J. Hu, Z. Wang, Y. Huang, A. Sena, F. R. y Baena, E. Burdet, Towards human-robot collaborative surgery: Trajectory and strategy learning in bimanual peg transfer. *IEEE Robot. Autom. Lett.* **8**, 4553–4560 (2023).
19. H. Zheng, Z. J. Hu, Y. Huang, X. Cheng, Z. Wang, E. Burdet, "A user-centered shared control scheme with learning from demonstration for robotic surgery," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)* (IEEE, 2024), pp. 15195–15201.
20. P. P. Gomez, R. E. Willis, K. R. Van Sickle, Development of a virtual reality robotic surgical curriculum using the da Vinci Si surgical system. *Surg. Endosc.* **29**, 2171–2179 (2015).
21. A. Nair, B. McGrew, M. Andrychowicz, W. Zaremba, P. Abbeel, "Overcoming exploration in reinforcement learning with demonstrations," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)* (IEEE, 2018), pp. 6292–6299.
22. M. Bain, C. Sammut, "A framework for behavioural cloning" in *Machine Intelligence 15*, *Intelligent Agents* (Oxford Univ., 2000), pp. 103–129.

23. L. Chen, K. Lu, A. Rajeswaran, K. Lee, A. Grover, M. Laskin, P. Abbeel, A. Srinivas, I. Mordatch, "Decision transformer: Reinforcement learning via sequence modeling," in *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)* (NeurIPS, 2021), pp. 15084–15097.
24. V. G. Goecks, G. M. Gremillion, V. J. Lawhern, J. Valasek, N. R. Waytowich, "Integrating behavior cloning and reinforcement learning for improved performance in dense and sparse reward environments," in *Proceedings of the International Conference on Autonomous Agents and MultiAgent Systems* (AAMAS, 2020), pp. 465–473.
25. A. Nair, A. Gupta, M. Dalal, S. Levine, AWAC: Accelerating online reinforcement learning with offline datasets. arXiv:2006.09359 [cs.LG] (2020).
26. I. Kostrikov, A. Nair, S. Levine, "Offline reinforcement learning with implicit Q-learning," in *Proceedings of the International Conference on Learning Representations (ICLR)* (ICLR, 2022), pp. 1–11.
27. N. M. Shafiullah, Z. Cui, A. A. Altanzaya, L. Pinto, "Behavior transformers: Cloning *k* modes with one stone," in *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)* (NeurIPS, 2022), pp. 22955–22968.
28. T. Huang, K. Chen, B. Li, Y. H. Liu, Q. Dou, "Demonstration-guided reinforcement learning with efficient exploration for task automation of surgical robot," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)* (IEEE, 2023), pp. 4640–4647.
29. P. J. Ball, L. Smith, I. Kostrikov, S. Levine, "Efficient online reinforcement learning with offline data," in *Proceedings of the International Conference on Machine Learning (ICML)* (ICML, 2023), pp. 1577–1594.
30. J. Luo, P. Dong, Y. Zhai, Y. Ma, S. Levine, "RLIF: Interactive imitation learning as reinforcement learning," in *Proceedings of the International Conference on Learning Representations (ICLR)* (ICLR, 2024), pp. 1–23.
31. Z. Fu, T. Z. Zhao, C. Finn, "Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation," in *Proceedings of the Annual Conference on Robot Learning (CoRL)* (CoRL, 2024), pp. 1–18.
32. C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, S. Song, Diffusion policy: Visuomotor policy learning via action diffusion. *Int. J. Robot. Res.*, 10.1177/02783649241273668 (2024).
33. Y. Long, W. Wei, T. Huang, Y. Wang, Q. Dou, Human-in-the-loop embodied intelligence with interactive simulation environment for surgical robot learning. *IEEE Robot. Autom. Lett.* **8**, 4441–4448 (2023).
34. L. Xiong, C. B. Chng, C. K. Chui, P. Yu, Y. Li, Shared control of a medical robot with haptic guidance. *Int. J. Comput. Assist. Radiol. Surg.* **12**, 137–147 (2017).
35. D. Sulsky, S.-J. Zhou, H. L. Schreyer, Application of a particle-in-cell method to solid mechanics. *Comput. Phys. Commun.* **87**, 236–252 (1995).
36. Y. Hu, Y. Fang, Z. Ge, Z. Qu, Y. Zhu, A. Pradhana, C. Jiang, A moving least squares material point method with displacement discontinuity and two-way rigid body coupling. *ACM Trans. Graph.* **37**, 1–14 (2018).
37. Y. Hu, T.-M. Li, L. Anderson, J. Ragan-Kelley, F. Durand, Taichi: A language for high-performance computation on spatially sparse data structures. *ACM Trans. Graph.* **38**, 1–16 (2019).
38. B. Kerbl, G. Kopanas, T. Leimkühler, G. Drettakis, 3D Gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.* **42**, 1–14 (2023).
39. Z. Yang, K. Chen, Y. Long, Q. Dou, SimEndoGS: Efficient data-driven scene simulation using robotic surgery videos via physics-embedded 3D Gaussians. arXiv:2405.00956 [cs.RO] (2024).
40. T. Osa, C. Staub, A. Knoll, "Framework of automatic robot surgery system using visual servoing," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (IEEE, 2010), pp. 1837–1842.
41. B. W. King, L. A. Reisner, A. K. Pandya, A. M. Composto, R. D. Ellis, M. D. Klein, Towards an autonomous robot for camera control during laparoscopic surgery. *J. Laparoendosc. Adv. Surg. Tech.* **23**, 1027–1030 (2013).
42. A. Attanasio, B. Scaglioni, M. Leonetti, A. F. Frangi, W. Cross, C. S. Biyani, P. Valdastri, Autonomous tissue retraction in robotic assisted minimally invasive surgery – A feasibility study. *IEEE Robot. Autom. Lett.* **5**, 6528–6535 (2020).
43. M. Yip, S. Salcudean, K. Goldberg, K. Althoefer, A. Menciassi, J. D. Opfermann, A. Krieger, K. Swaminathan, C. J. Walsh, H. H. Huang, I.-C. Lee, Artificial intelligence meets medical robotics. *Science* **381**, 141–146 (2023).
44. X. Zhao, W. Ding, Y. An, Y. Du, T. Yu, M. Li, M. Tang, J. Wang, Fast segment anything. arXiv:2306.12156 [cs.CV] (2023).
45. P. Kazanzides, Z. Chen, A. Deguet, G. S. Fischer, R. H. Taylor, S. P. DiMaio, "An open-source research kit for the da Vinci surgical system," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)* (IEEE, 2014), pp. 5434–6439.
46. A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W. Y. Lo, P. Dollar, R. Girshick, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)* (IEEE, 2023), pp. 4015–4026.
47. G. Xu, X. Wang, X. Ding, X. Yang, "Iterative geometry encoding volume for stereo matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2023), pp. 21919–21928.
48. K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2015), pp. 770–778.
49. Y. Jia, X. Y. Zhou, Policy gradient and actor-critic learning in continuous time and space: Theory and algorithms. *J. Mach. Learn. Res.* **23**, 1–50 (2022).
50. D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, M. Riedmiller, "Deterministic policy gradient algorithms," in *Proceedings of the International Conference on Machine Learning (ICML)* (ICML, 2014), pp. 387–395.
51. Y. H. Su, A. Munawar, A. Deguet, A. Lewis, K. Lindgren, Y. Li, R. H. Taylor, G. S. Fischer, B. Hannaford, P. Kazanzides, "Collaborative Robotics toolkit (CRTK): Open software framework for surgical robotics research," in *Proceedings of the 2020 Fourth IEEE International Conference on Robotic Computing (IRC)* (IEEE, 2020), pp. 48–55.
52. G. Claudio, F. Spindler, F. Chaumette, "Vision-based manipulation with the humanoid robot romeo" in *IEEE-RAS International Conference on Humanoid Robots (Humanoids)* (IEEE, 2016), pp. 286–293.
53. C. De Farias, M. Adjigble, B. Tamadazte, R. Stolkin, N. Marturi, "Dual quaternion-based visual servoing for grasping moving objects" in *IEEE International Conference on Automation Science and Engineering (CASE)* (IEEE. 2021), pp.151–158.
54. F. Zhong, B. Li, W. Chen, Y. H. Liu, Robot–camera calibration in tightly constrained environment using interactive perception. *IEEE Trans. Robot.* **39**, 4952–4970 (2023).

Acknowledgments: We acknowledge Cornerstone Robotics for providing the Sentire surgical system and all necessary instruments and accessories and APIs for our experiments. We express our gratitude to W. Luo for developing the Python API for the Sentire surgical system. We thank Y. Chan for preparing ex vivo animal tissues and supporting the in vivo trial and L. Jiang, H. Wang, and Q. Wang for the help in creating the illustrations in the manuscript. We appreciate K. Leung for helping to apply for ethics approval. **Funding:** This work was funded by the National Natural Science Foundation of China (project no. 62322318), the Research Grants Council of the Hong Kong Special Administrative Region (project nos. 24209223, 14208424, T45-401/22-N, T42-409/18-R, and AoE/E-407/24-N), the Hong Kong Innovation and Technology Fund (project no. ITS/223/22), and the industry sponsorship from Cornerstone Robotics. The collaboration between CUHK and JHU (Johns Hopkins University) researchers is under funding support from the Multi-Scale Medical Robotics Center under the InnoHK Initiative of the Innovation and Technology Commission of HKSAR (Hong Kong Special Administrative Region) Government. **Author contributions:** Q.D., K.W.S.A., and P.W.Y.C. conceived the study. Q.D., K.W.S.A., Z.W., Z.C., Y. Liu, R.H.T., P.K., P.W.Y.C., and H.C.Y. designed the work. Y. Long, A.L., Z.Y., K.S., L.S., J.F., H.L., W.W., K.C., and X.C. developed the methodology and conducted experiments with data analysis. Y. Long, L.Z., X.C., H.C.Y., Z.W., and Q.D. proposed specific surgical assistive tasks for autonomy. Z.C., Y.H., L.Z., and D.H.C.K. provided technical support for the Sentire surgical system and its accessories during both ex vivo and in vivo trials. H.C.Y. and P.W.Y.C. provided clinical perspectives. K.S., H.L., L.S., Y. Long, X.C., and Q.D. designed supplementary videos. Y. Long, A.L., K.C., H.C.Y., Z.W., and Q.D. cowrote the initial manuscript, with all coauthors providing constructive comments and editing. **Competing interests:** K.W.S.A. is an employee of Cornerstone Robotics and holds an academic professorship at the Chinese University of Hong Kong. D.H.C.K., L.Z., Y.H., Z.C., and Z.W. are solely employed by Cornerstone Robotics. The other authors declare no conflicts of interest. The authors declare that this research was conducted independently of the funding source and that the results are not influenced by these affiliations. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper or the Supplementary Materials. All of the materials and source codes are available at Zenodo: https://doi.org/10.5281/zenodo.15615164 and GitHub website: https://github.com/med-air/SurRoL.

Submitted 30 September 2024
Accepted 17 June 2025
Published 16 July 2025
10.1126/scirobotics.adt3093

Science Robotics

Surgical embodied intelligence for generalized task autonomy in laparoscopic robot-assisted surgery

Yonghao Long, Anran Lin, Derek Hang Chun Kwok, Lin Zhang, Zhenya Yang, Kejian Shi, Lei Song, Jiawei Fu, Hongbin Lin, Wang Wei, Kai Chen, Xiangyu Chu, Yang Hu, Hon Chi Yip, Philip Wai Yan Chiu, Peter Kazanzides, Russell H. Taylor, Yunhui Liu, Zihan Chen, Zerui Wang, Samuel Kwok Wai Au, and Qi Dou

Sci. Robot. **10** (104), eadt3093. DOI: 10.1126/scirobotics.adt3093

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.