Surgical Robot Transformer (SRT): Imitation Learning for Surgical Tasks

Ji Woong Kim[1] **Tony Z. Zhao**[2] **Samuel Schmidgall**[1] **Anton Deguet**[1]
Marin Kobilarov[1] **Chelsea Finn**[2] **Axel Krieger**[1]
Johns Hopkins University[1] Stanford University[2]
https://surgical-robot-transformer.github.io/

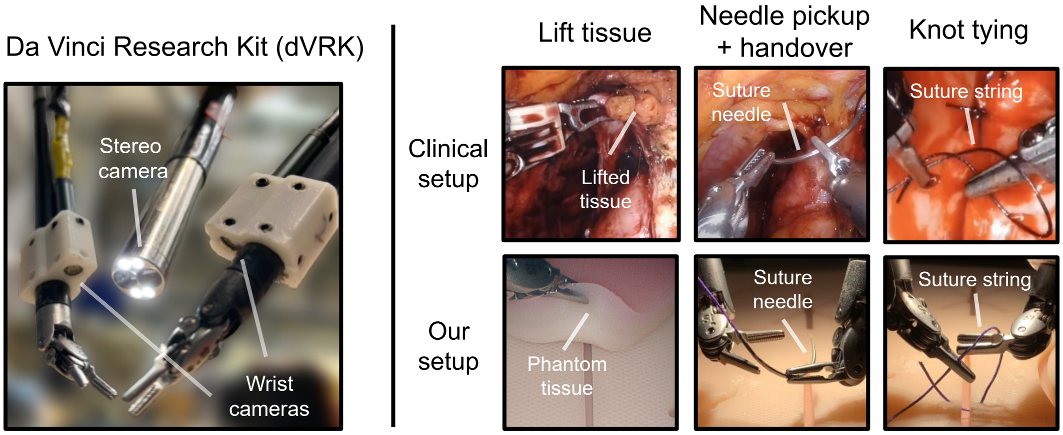

Figure 1: (*Left*): The da Vinci Surgical Research Kit (dVRK) system is equipped with a surgical endoscope and wrist cameras. (*Right*): Three fundamental surgical tasks are learned, including lift tissue (i.e. tissue retraction), needle-pickup and handover, and knot-tying which are among the most common surgical tasks.

Abstract: We explore whether surgical manipulation tasks can be learned on the da Vinci robot via imitation learning. However, the da Vinci system presents unique challenges which hinder straight-forward implementation of imitation learning. Notably, its forward kinematics is inconsistent due to imprecise joint measurements, and naively training a policy using such approximate kinematics data often leads to task failure. To overcome this limitation, we introduce a relative action formulation which enables successful policy training and deployment using its approximate kinematics data. A promising outcome of this approach is that the large repository of clinical data, which contains approximate kinematics, may be directly utilized for robot learning without further corrections. We demonstrate our findings through successful execution of three fundamental surgical tasks, including tissue manipulation, needle handling, and knot-tying.

Keywords: Imitation Learning, Manipulation, Medical Robotics

1 Introduction

Recently, large-scale imitation learning has shown great promise in creating generalist systems for manipulation tasks [1]. Prior research in this area has mostly focused on learning day-to-day household activities. However, an under-explored area with high potential is the surgical domain, particularly with the use of Intuitive Surgical's da Vinci robot. These robots are deployed globally and possess immense scaling potential: as of 2021, over 10 million surgeries have been performed using 6,500 da Vinci systems in 67 countries, with 55,000 surgeons trained on the system [2]. Often, the video and kinematics data are recorded for post-operative analysis, resulting in a large repository of demonstration data. Utilizing such large scale data holds significant potential for building generalist systems for autonomous surgery [3].

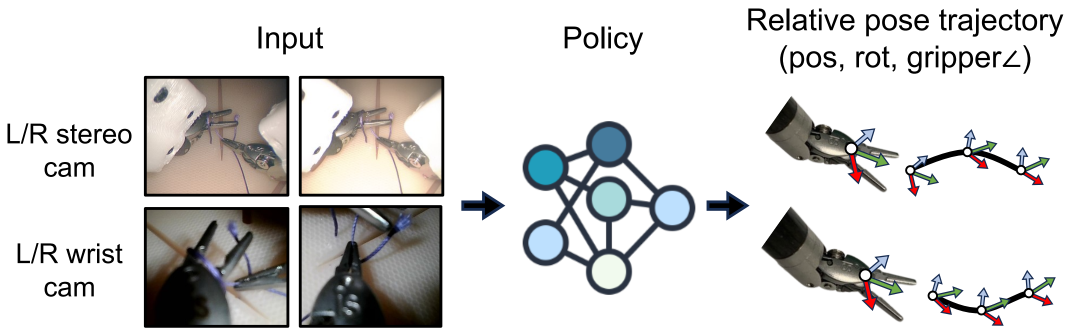

Figure 2: We propose a policy design which only takes images as input and outputs relative pose trajectories for both arms. Modeling policy actions as relative motion is a key ingredient that makes robot learning work on the dVRK.

However, robot learning on the da Vinci presents unique challenges. The hardware suffers from inaccurate forward kinematics due to potentiometer-based joint measurements, hysteresis, and overall flexibility and slack in its mechanism [4]. These limitations result in the robot's failure to perform simple visual-servoing tasks [5]. As we discover in this work, naively training a policy using such approximate kinematics data almost always leads to task failure. For instance, a policy trained to output absolute end-effector poses, which is a common strategy to train robot policies, achieves near-zero success rates across all tasks explored in this work, including tissue lift, needle pickup and handover, and knot-tying (Fig. 1). To achieve robot learning at scale, we must devise a strategy that leverages such approximate kinematics data effectively.

Towards this end, we present an approach for robot learning on the da Vinci using its approximate kinematics data. Intuitively, our approach is based on the observation that the relative motion of the robot is much more consistent than its absolute forward kinematics. We thus model policy actions as differential motion and further explore its variants to design the most effective action representation for the da Vinci. We find that training an imitation learning algorithm using such relative formulation shows robustness to various configuration changes to the robot, even those known to significantly disrupt the robot's forward kinematics. Specifically, the da Vinci tools can be removed and reinstalled and all the robot joints can be freely moved, including the notoriously inaccurate set-up joints [4], without significantly impacting policy performance.

Additionally, we explore the use of wrist cameras in the surgical workflow. While not commonly employed in clinical settings, wrist cameras have demonstrated effectiveness in improving policy performance and facilitating generalization to out-of-distribution scenarios, such as varying workspace heights or unfamiliar visual distractions [6]. We thus evaluate their impact on performance and practical potential by designing removable brackets that enable easy sharing across various surgical instruments.

Overall, our results indicate that the relative motion on the da Vinci is more consistent than its absolute motion. Following this result, we further observe that a carefully chosen relative action representation can sufficiently train policies that achieve high success rates in surgical manipulation tasks. Additionally, using wrist cameras significantly improves policy performance, especially during phases of the procedure when precise depth estimation is crucial. In robustness tests, our model demonstrates the ability to generalize to novel scenarios, such as in the presence of unseen 3D suture pads and animal tissues, showing promise for future extensions into pre-clinical research.

Our main contributions are: (i) a successful demonstration of imitation learning on the da Vinci while using its approximate kinematics data and without requiring further kinematics corrections, while drastically outperforming the baseline approach; (ii) experiments showing that imitation learning can effectively learn complex surgical tasks and generalize to novel scenarios such as in the presence of unseen realistic tissue; (iii) ablative experiments demonstrating the importance of wrist cameras for learning surgical manipulation tasks.

2 Related Work

Manipulation and Imitation Learning Imitation learning enables robots to learn from expert demonstrations [7]. Behavioral cloning (BC) is a simple instantiation of imitation learning that directly predicts actions from observations. Early works tackle this problem through the lens of motor primitives [8, 9, 10, 11]. With the development of deep learning and generative modeling, different architectures and training objectives have been proposed to model the demonstrations end-to-end. This includes the use of ConvNets or ViT [12] for image processing [13, 14, 15], RNN or transformers for fusing history of observations [16, 17, 18], tokenization of the action space [19], generative modeling techniques such as energy-based models [20], diffusion [21] and VAEs [22, 23]. Prior works also focus on the few-shot aspect of imitation learning, [24, 25, 26, 27], language conditioning [15, 18, 19, 28], co-training [19, 29, 28], retrieval [30, 31, 32], using play data [33, 34, 35, 36], using human videos [37, 38, 39, 40, 41, 42], and exploiting task-specific structures [43, 44, 45].

However, most of these prior works focus on table-top manipulation in home settings. Surgical tasks, on the other hand, pose a unique set of challenges. They require precise manipulation of deformable objects, involve hard perception problems with inconsistent lighting and occlusions, and surgical robots may often have inaccurate proprioception and hysteresis [4] that are less pronounced in industrial arms. While in principle end-to-end imitation learning could capture these variations implicitly, it is unclear what design choices are important to enable effective learning in this regime.

We also note that in the dVRK community, the inaccuracies of the robot have been addressed via hand-eye calibration [46, 47, 48]. However, hand-eye calibration is effective if it is performed during both data collection and inference, in which case ground-truth kinematics data would be available at all times. In our scenario, however, we assume that the demonstration dataset has been already been collected without hand-eye calibration e.g., the large-scale clinical data and thus assume that precise ground-truth kinematics is not available during training. While hand-eye calibration can still be performed during inference and may possibly help, it is not a fundamental solution to the problem.

Autonomous Surgery Prior works in autonomous surgery primarily focus on designing task-specific policies for specific tasks, such as for suturing [49, 50, 51, 52], endoscope control [53, 54], navigation [55, 56, 57], and tissue manipulation [58, 59]. Such developments have led to impressive demonstrations such as automating intestinal anastomosis (suturing of two tubular structures) in-vivo on a pig [60]. However, these methods do not typically scale well across various tasks or generalize well to varying environmental conditions. In contrast, end-to-end imitation learning offers a relatively simple solution to these shortcoming by only requiring good robot demonstrations. While prior works have also explored the use of imitation learning for surgical tasks [61, 62, 63, 55], its application to complex manipulation tasks like knot-tying remains unexplored, and practical design choices for implementing on the da Vinci have not been addressed.

3 Technical Approach

Consider the dVRK system, as illustrated in Fig. 3, which includes both the robot and a teleoperation console for user interaction. The dVRK features an endoscopic camera maipulator (ECM) and two patient side manipulators (PSM1, PSM2) sharing the same robot base. Each arm is a sequential combination of set-up joints (SUJ) which are passive, followed by active joints which are motorized (Fig. 3). The passive joints are notoriously inaccurate due to using only potentiometers for joint measurements. The active joints use both potentiometers and motor encoders, providing improved precision. However, in general, the use of potentiometers throughout all the joints causes the forward kinematics of the arms to be inaccurate, even up to 5cm error [4].

Using the dVRK console, the user collects many demonstrations of a task, acquiring a dataset $D = \{\tau_1, ..., \tau_N\}$, where each trajectory $\tau_i = \{(o_1, x_1, a_1), ..., (o_T, x_T, a_T)\}$ is a collection of observation o_t, proprioception x_t, and actions a_t, collected at time step t. Specifically, the observation o_t includes left/right surgical endoscope images and left/right wrist camera images, totaling four images (Fig. 2), proprioception is the current pose of the PSMs w.r.t surgical endoscope tip frame denoted as $x_t = \{g_t^l,\ g_t^r\}$, where $g = (p, R) \in SE(3)$ and $\{l, r\}$ denotes the left and right grippers

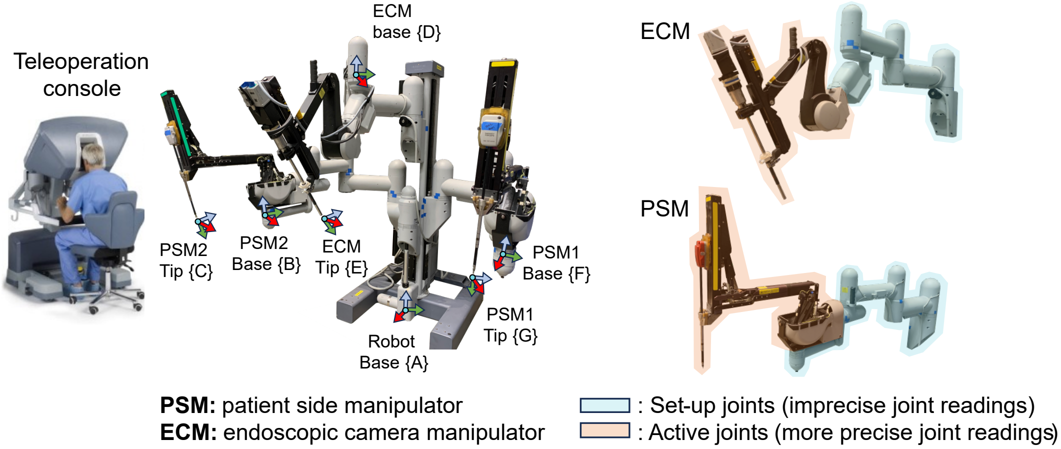
Teleoperation console
ECM base {D}
PSM2 Tip {C}
PSM2 Base {B}
ECM Tip {E}
PSM1 Base {F}
PSM1 Tip {G}
Robot Base {A}
ECM
PSM
PSM: patient side manipulator
ECM: endoscopic camera manipulator
▭ : Set-up joints (imprecise joint readings)
▭ : Active joints (more precise joint readings)

Figure 3: The dVRK system consists of an endoscopic camera manipulator (ECM) and two patient side manipulators (PSM1, PSM2). Unfortunately, the dVRK arms are notorious for providing inconsistent forward kinematics. This is due to the setup joints (blue) only using potentiometers for joint measurements, which can be unrelible. The active joints (pink) use both potentiometers and motor encoders, improving precision.

respectively, and actions are the commanded desired waypoints to reach specified via teleoperation control, denoted as $a_t = \{\hat{g}_t^l,\ \hat{g}_t^r\}$.

Our objective is to learn surgical manipulation tasks via imitation learning. Given the robot's inaccurate forward kinematics, choosing the appropriate action representation is crucial. To illustrate this, we investigate three action representations: camera-centric, tool-centric, and hybrid-relative as shown in Fig. 4. The camera-centric approach serves as a baseline, highlighting the limitations of modeling actions as absolute poses of the end-effectors. The tool-centric approach offer an improved formulation by modeling actions as relative motion, leading to higher success rates. The hybrid-relative approach further improves beyond tool-centric approach by modeling translation actions with respect to a fixed reference frame, further improving accuracy in translation movements. These approaches are detailed below:

1. ***Camera-centric actions***: We model camera-centric actions as absolute poses of the end-effectors w.r.t the endoscope tip frame. The setup is similar to how position-based visual servoing applications (PBVS) are implemented and is a natural choice on the dVRK. Specifically, the objective is to learn a policy π that, given an observation o_t at time t, predicts an action sequence $A_{t,C} = (a_t, ..., a_{t+C})$, where C denotes the action prediction horizon. The policy can thus be defined as $\pi : o_t \mapsto A_{t,C}$. This formulation is visually shown in Fig. 4.
2. ***Tool-centric actions***: We model tool-centric actions as relative motion w.r.t the *current* end-effector frame, which is a moving body frame. Tool-centric actions can be defined as:
$$A_{t,C}^{tool} = \big\{(g_t^i)^T \hat{g}_s^i \mid s \in [t, t+C];\ i \in \{l, r\}\big\} \tag{1}$$
Intuitively, the desired poses $\hat{g}_s^i$ are subtracted by the current end-effector poses g_t^i using the $SE(3)$ subtraction rule, for each time up to horizon C and for each corresponding left and right grippers. There are largely two benefits to adopting this action representation. A relative motion formulation is used, which we show later in Section 5, is more consistent compared to the absolute forward kinematics of the arms. Also, the subtraction cancels out the endoscope forward kinematics terms and the actions can be expressed in terms of the PSMs forward kinematics only. This effectively reduces the margin for errors since less joints are involved in representing actions. However, one caveat of this approach is that the delta motion is defined w.r.t a moving reference frame. This requires the policy to implicitly localize the current end-effector orientation from image observations, and output translations and rotations along the localized principal axes, which can be a challenging task. The policy objective can be defined as $\pi : o_t \mapsto A_{t,C}^{tool}$. This formulation is visually shown in Fig. 4.

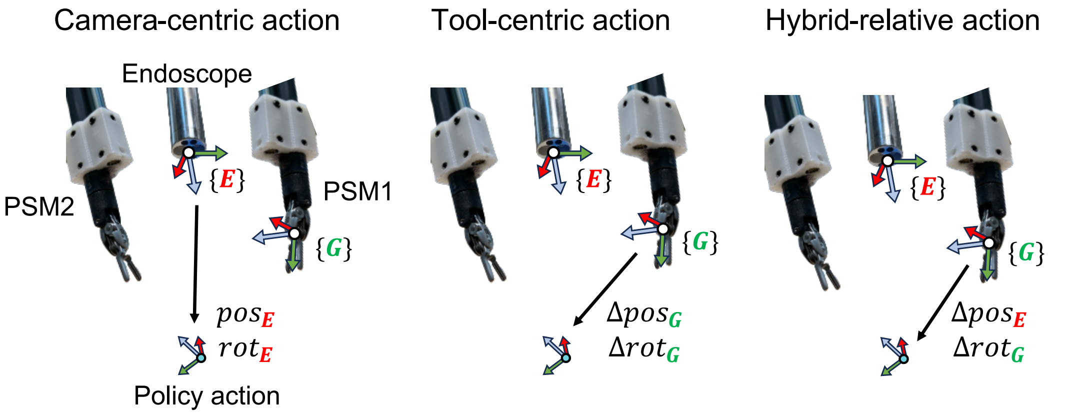

Figure 4: We consider three options for modeling policy actions. *(Left):* Camera-centric approach models actions as absolute end-effector poses w.r.t the endoscope tip frame. *(Middle):* Tool-centric approach models actions as delta positions and delta rotations defined w.r.t the current end-effector frame. *(Right):* Hybrid relative approach models actions as delta positions defined w.r.t the endoscope tip frame and delta rotations defined w.r.t the current end-effector frame.

3. ***Hybrid Relative Actions***: Similar to tool-centric actions, hybrid relative actions are modeled as relative motion but w.r.t two different reference frames. Specifically, the delta translations are defined w.r.t endoscope tip frame and delta rotations are defined w.r.t the current end-effector frame. This formulation can be defined as follows:

$$A_{t,C}^{hybrid} = \{\hat{g}_s^i \ominus g_t^i \mid s \in [t, t+C];\ i \in \{l, r\}\} \tag{2}$$

Where the subtraction operation $\ominus$ defined as:

$$\hat{g}_s^i \ominus g_t^i = \left(\hat{p}_s^i - p_t^i, (R_t^i)^T \hat{R}_s^i\right) \tag{3}$$

Intuitively, the subtraction is performed between the corresponding translation and rotation elements i.e. vector subtraction for positions and $SO(3)$ subtraction for rotations. A key distinction of this approach from the tool-centric approach lies in modeling the delta translation with respect to the fixed frame of the endoscope-tip, rather than the moving frame of the end-effector. This approach removes the burden for the policy to localize the end-effector's orientation to generate delta translations along the localized axes, thereby improving the quality of translation motion. The policy can be defined as $\pi : o_t \mapsto A_{t,C}^{hybrid}$. This formulation is visually shown in Fig. 4.

4 Implementation Details

To train our policies, we use action chunking with transformers (ACT) [23] and diffusion policy [64]. The policies were trained using the endoscope and wrist cameras images as input, which are all downsized to image size of 224 × 224 × 3. The original input size of the surgical endoscope images were 1024 × 1280 × 3 and the wrist images were 480 × 640 × 3. Kinematics data is not provided as input as commonly done in other imitation learning approaches because it is generally inconsistent due to the design limitations of the dVRK. The policy outputs include the end-effector (delta) position, (delta) orientation, and jaw angle for both arms. We leave further specific implementation details in Appendix A.

5 Experiments

In our experiments, we aim to understand the following key questions: (1) is imitation learning sufficient to learn challenging surgical manipulation tasks? (2) Is the dVRK's relative motion more consistent than its absolute forward kinematics? (3) Are using wrist cameras critical to achieving high success rates? (4) How well does our proposed model generalize to unseen novel scenarios? To answer these questions, we compare the task success rates of the policies trained using various

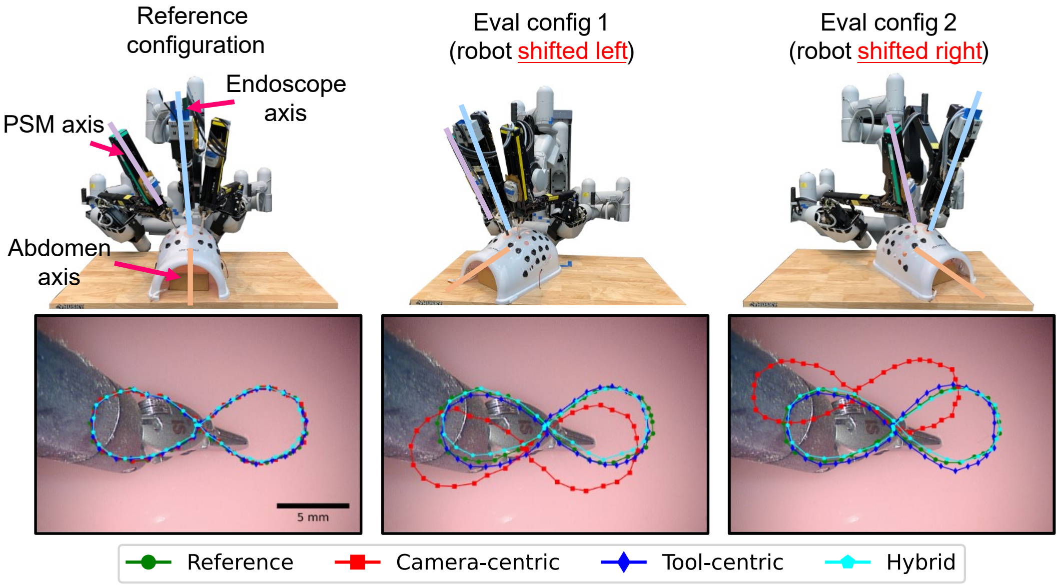

Figure 5: The repeatability of all action representations are tested by repeating a recorded reference trajectory under various robot configurations. (*Left*): The first column shows perfect reconstruction of the reference trajectory for all action representations since the robot joints have not moved since when the reference trajectory was collected. *(Middle, Right)* When the robot is shifted to the left or to the right, the camera-centric action representation fails to track the reference trajectory while the relative action representations track them quite closely. This is primarily due to the set-up joints being moved, which causes significant joint measurement errors. This experiment proves that in the presence of inconsistent joint measurements, relative motion can be more consistent.

Table 1: Trajectory tracking RMSE (mm) under various robot configurations

	Ref config	Eval config 1	Eval config 2
Camera-centric	0.6	1.9	2.8
Tool-centric	0.9	0.9	0.7
Hybrid-relative	0.9	0.8	0.8

action representations. We also directly compare the consistency of relative versus absolute motion by tracking a reference trajectory using the various action representations and comparing their tracking errors. We also explore the importance of wrist cameras by comparing the policy performance with and without them. Finally, we consider whether the proposed models can generalize to novel unseen scenarios, such as in the presence of animal tissues. These experiments are explored in the context of three tasks: lift tissue, needle pickup and handover, and knot-tying.

Experiment Setup During data collection, the robot is set up in a reference configuration as shown in Fig. 5. In this configuration, 224 trials were collected for tissue lift, 250 trials for needle pickup and handover, and 500 trials for knot-tying, all collected by a single user across multiple days. During all experiments, a dome simulating the human abdomen (Fig. 5) was used to roughly place the arms and the endoscope in an approximately similar location using the same holes. The placement is only approximate because the holes are much larger than the endoscope and tool shaft size, and the tools have to be manually placed into the holes by moving the set-up joints.

Evaluating the Consistency of Relative Motion vs. Absolute Forward Kinematics In this section we seek to understand whether relative motion on the dVRK is more consistent than its absolute forward kinematics. To test our hypothesis, we teleoperate a reference trajectory e.g., an infinity sign as shown in Fig. 5. This trajectory is then represented in various action representations using the formulas presented in Section 3. Then, we place the end-effector in the same initial pose and replay the trajectories using the various action representations under different robot configurations. These different configurations include shifting the robot workspace to the left and to the right side

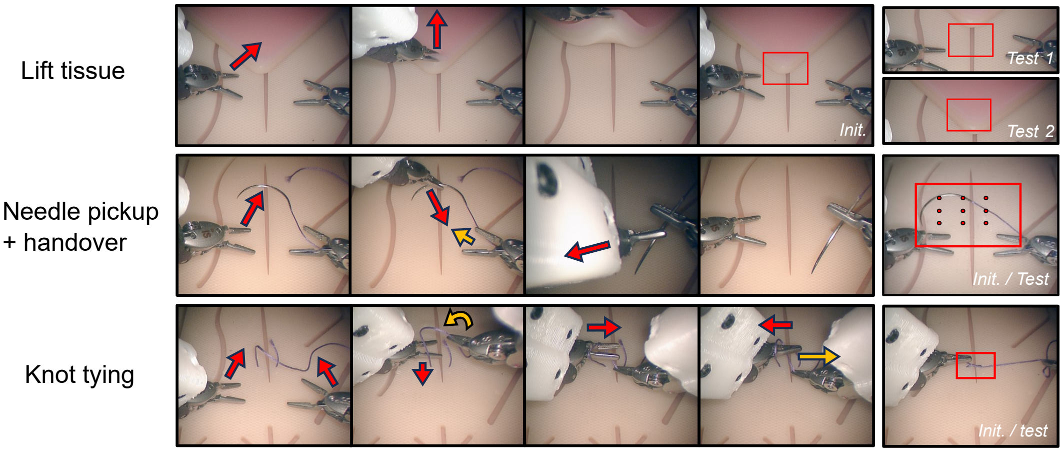

Figure 6: *(Top):* Tissue lift task requires grabbing the corner of the rubber pad (i.e. tissue) and lifting it upwards. During training the corner is kept within the red box and the configuration of the corners at test time is shown. *(Middle):* Needle pickup and handover is self-explanatory. The needle was placed randomly inside the red box during training. At test time, the center hump of the needle was placed at nine locations as shown, to enforce consistent setup during evaluation. *(Bottom):* Knot-tying requires creating a loop using the left string, grabbing the terminal end of the string through the loop, and pulling the grippers away from each other. During training, the location of the strings originating from the pads were randomly placed inside the red box, and at test time, it was centered in the red box as shown.

Table 2: Success rates on three surgical tasks using various action representations

		Tissue lift		Needle pick + handover		Knot tying		
		Test 1	Test 2	Grasp	Handover	Grasp String	Loop	Whole Task
ACT [23]	Camera-centric	0/5	0/5	0/9	0/9	0/20	0/20	0/20
	Tool-centric	**5/5**	**5/5**	**9/9**	5/9	**20/20**	**20/20**	**18/20**
	Hybrid-relative	**5/5**	**5/5**	**9/9**	**9/9**	**20/20**	**20/20**	**18/20**
	Hybrid-relative (no wrist cam)	-	-	**9/9**	6/9	8/20	4/20	4/20
	Hybrid-relative (pork backgrd)	-	-	**9/9**	**9/9**	-	-	-
Diffusion Policy [21]	Hybrid-relative	**5/5**	**5/5**	8/9	4/9	10/20	7/20	4/20

(Fig. 5). These workspace shifts cause the robot set-up joints to move, which are the joints prone to cause large joint measurement errors due to using only potentiometers for joint measurements. To compare their tracking errors, the replayed trajectories are annotated at the end-effector (i.e. control point) in image coordinates and plotted, as shown in the bottom row of Fig. 5.

The plots in Fig. 5 and the numeric RMSE results in Table 1 show that in the reference configuration, all action representation precisely reconstruct the reference trajectory, since the set-up joints have not yet moved. However, when the robot configuration is changed by moving the set-up joints and new erroneous joint measurments are obtained, the camera-centric action representation fails to reconstruct the reference trajectory. Also, this error is not consistent for different robot configurations as shown in the trajectory plots in Fig. 5. For relative action representations, which include tool-centric and hybrid-relative action formulations, the reference trajectory is repeated more consistently, and their numeric errors do not vary significantly as observed in Table 1. In summary, this experiment shows that relative motion on the dVRK is more consistent compared to its absolute forward kinematics in the presence of inconsistent joint measurement errors.

Policy Performance Using Various Action Representations We evaluate the policy performance using the various action representations on tissue lift, needle pickup and handover, and knot-tying as shown in Table 2. The camera-centric action representation performed poorly across all three tasks. Because the joint measurements of the dVRK are inconsistent, the end-effectors almost always failed to reach the target objects (e.g., tissue corner, needle, and string) and often dangerously collided with the underlying rubber pads. The policy trained using tool-centric action representation showed improved performance across all three tasks. However, during needle pickup + handover when large rotations were involved, the handover phase of the task often failed. In particular, after picking up the needle, the left gripper had to make a ~ 90 degree rotation to transfer the needle to the opposing arm (Fig. 6). During this phase of the motion, the orientations of the grippers appeared correct, however, the translation motion appeared incorrect and seemed to be the cause of task failure. We conjecture this reason was due to grounding the policy actions to a moving end-effector frame. The policy is required to localize the moving end-effector orientation using image observations and generate delta translations along the localized principal axes of direction, which can be a challenging task. To fix this issue, the hybrid relative motion action was used, which grounds the translation motion to a fixed frame of the endoscope-tip. This formulation improved the translation errors during the aforementioned needle handover phase and ultimately achieved the highest success rates across all three tasks. This best performing action representation was also implemented on diffusion policy, however, the performance was not as high as ACT.

Evaluating the Importance of Wrist Camera We also evaluate the importance of adding wrist cameras and their contribution to task success. To demonstrate this, we trained policies without wrist cameras on the needle pickup and handover and knot-tying tasks using the hybrid relative action formulation (Table 2). Overall, we observed that omitting wrist cameras lead to significant drop in performance. We conjecture that wrist cameras aid in scenarios where precise depth estimation is necessary. For instance, during the needle pickup and handover task, specifically during the latter phase transferring the needle, the wrist views clearly showed whether the needle was being navigated into the opposing grippers from afar. This additional view may have provided better context for task success. However, we observed that this level of information was difficult to discern from the third-person endoscopic view.

Evaluating Generalization We also evaluate the ability of our models to generalize to novel scenarios, such as under more clinically relevant background using animal tissues (pork and chicken) and an unseen 3D suture pad. Most of this evaluation remains qualitative and greater details are elaborated in Appendix B. In terms of quantitative results, we evaluate the hybrid-relative action formulation on the needle pick-up and handover task on a pig loin background. We observe that its overall success rate is quite high (Table 2), however, the quality of the motion and the accuracy of the needle grasps were much lower compared to those observed in the core experiments. In terms of qualitative results, we observe multiple instances of successful knot-tying achieved on pork tissue, and needle grasps on chicken background and on an unseen 3D pad (Appendix B).

6 Limitations and Conclusion

In this work, we opt for using off-the-shelf large wrist cameras which are not clinically relevant. However, the cameras may be replaced with much smaller ones (1-2mm diameter) and its mount can be further optimized by integrating quick-release mechanisms for swift transfer between surgical tools. Also, our model is limited as it can only act based on current observations and does not have the ability to modulate different behavior based on human instruction. We hope to address these issues in future work to further advance the autonomy of surgical robots.

In summary, we demonstrated an approach for imitation learning on the dVRK using its approximate kinematics data, without providing further post-processing corrections. The key idea of our approach was to rely on the more consistent relative motion of the robot, achieved by modeling policy actions as relative motion such as tool-centric and hybrid-relative actions. As mentioned in the introduction, we believe that our work is a step towards leveraging the large repository of approximate surgical data for robot learning at scale, without providing further kinematics corrections. We believe more research in this direction can further guide the path towards building general-purpose systems towards autonomous surgery.

References

[1] E. Collaboration, A. Padalkar, A. Pooley, A. Mandlekar, A. Jain, A. Tung, A. Bewley, A. Herzog, A. Irpan, A. Khazatsky, A. Rai, A. Singh, A. Garg, A. Brohan, A. Raffin, A. Wahid, B. Burgess-Limerick, B. Kim, B. Schölkopf, B. Ichter, C. Lu, C. Xu, C. Finn, C. Xu, C. Chi, C. Huang, C. Chan, C. Pan, C. Fu, C. Devin, D. Driess, D. Pathak, D. Shah, D. Büchler, D. Kalashnikov, D. Sadigh, E. Johns, F. Ceola, F. Xia, F. Stulp, G. Zhou, G. S. Sukhatme, G. Salhotra, G. Yan, G. Schiavi, G. Kahn, H. Su, H.-S. Fang, H. Shi, H. B. Amor, H. I. Christensen, H. Furuta, H. Walke, H. Fang, I. Mordatch, I. Radosavovic, I. Leal, J. Liang, J. Abou-Chakra, J. Kim, J. Peters, J. Schneider, J. Hsu, J. Bohg, and J. Bingham. Open x-embodiment: Robotic learning datasets and rt-x models, 2023.

[2] Intuitive Surgical. 2021 intuitive sustainability report. https://www.intuitive.com/en-us/-/media/ISI/Intuitive/Pdf/2021-intuitive-sustainability-report.pdf, 2021.

[3] S. Schmidgall, J. W. Kim, A. Kuntz, A. E. Ghazi, and A. Krieger. General-purpose foundation models for increased autonomy in robot-assisted surgery. *arXiv preprint arXiv:2401.00678*, 2024.

[4] Z. Cui, J. Cartucho, S. Giannarou, and F. Rodriguez y Baena. Caveats on the first-generation da vinci research kit: Latent technical constraints and essential calibrations. *IEEE Robotics amp; Automation Magazine*, page 2–17, 2023. ISSN 1558-223X. doi:10.1109/mra.2023.3310863. URL http://dx.doi.org/10.1109/MRA.2023.3310863.

[5] M. Hwang, B. Thananjeyan, S. Paradis, D. Seita, J. Ichnowski, D. Fer, T. Low, and K. Goldberg. Efficiently calibrating cable-driven surgical robots with rgbd fiducial sensing and recurrent neural networks. *IEEE Robotics and Automation Letters*, 5(4):5937–5944, 2020. doi:10.1109/LRA.2020.3010746.

[6] K. Hsu, M. J. Kim, R. Rafailov, J. Wu, and C. Finn. Vision-based manipulators need to also see from their hands, 2022.

[7] D. A. Pomerleau. Alvinn: An autonomous land vehicle in a neural network. In *NIPS*, 1988.

[8] A. J. Ijspeert, J. Nakanishi, H. Hoffmann, P. Pastor, and S. Schaal. Dynamical movement primitives: learning attractor models for motor behaviors. *Neural computation*, 2013.

[9] P. Pastor, H. Hoffmann, T. Asfour, and S. Schaal. Learning and generalization of motor skills by learning from demonstration. *2009 IEEE International Conference on Robotics and Automation*, pages 763–768, 2009.

[10] J. Kober and J. Peters. Learning motor primitives for robotics. In *2009 IEEE International Conference on Robotics and Automation*, 2009.

[11] A. Paraschos, C. Daniel, J. Peters, and G. Neumann. Using probabilistic movement primitives in robotics. *Autonomous Robots*, 42:529–551, 2018.

[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. 2021.

[13] R. Rahmatizadeh, P. Abolghasemi, L. Bölöni, and S. Levine. Vision-based multi-task manipulation for inexpensive robots using end-to-end learning from demonstration. *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3758–3765, 2017.

[14] T. Zhang, Z. McCarthy, O. Jow, D. Lee, K. Goldberg, and P. Abbeel. Deep imitation learning for complex manipulation tasks from virtual reality teleoperation. *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–8, 2017.

[15] E. Jang, A. Irpan, M. Khansari, D. Kappler, F. Ebert, C. Lynch, S. Levine, and C. Finn. Bc-z: Zero-shot task generalization with robotic imitation learning. In *Conference on Robot Learning*, 2022.

[16] A. Mandlekar, D. Xu, J. Wong, S. Nasiriany, C. Wang, R. Kulkarni, L. Fei-Fei, S. Savarese, Y. Zhu, and R. Mart'in-Mart'in. What matters in learning from offline human demonstrations for robot manipulation. In *Conference on Robot Learning*, 2021.

[17] N. M. M. Shafiullah, Z. J. Cui, A. Altanzaya, and L. Pinto. Behavior transformers: Cloning k modes with one stone. *ArXiv*, abs/2206.11251, 2022.

[18] 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. Ryoo, G. Salazar, P. Sanketi, K. Sayed, J. Singh, S. Sontakke, A. Stone, C. Tan, H. Tran, V. Vanhoucke, S. Vega, Q. Vuong, F. Xia, T. Xiao, P. Xu, S. Xu, T. Yu, and B. Zitkovich. Rt-1: Robotics transformer for real-world control at scale. In *arXiv preprint arXiv:2212.06817*, 2022.

[19] 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, and B. Zitkovich. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *arXiv preprint arXiv:2307.15818*, 2023.

[20] P. R. Florence, C. Lynch, A. Zeng, O. Ramirez, A. Wahid, L. Downs, A. S. Wong, J. Lee, I. Mordatch, and J. Tompson. Implicit behavioral cloning. *ArXiv*, abs/2109.00137, 2021.

[21] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song. Diffusion policy: Visuomotor policy learning via action diffusion, 2023.

[22] C. Lynch, M. Khansari, T. Xiao, V. Kumar, J. Tompson, S. Levine, and P. Sermanet. Learning latent plans from play. In *Conference on Robot Learning*, 2019.

[23] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *RSS*, 2023.

[24] Y. Duan, M. Andrychowicz, B. C. Stadie, J. Ho, J. Schneider, I. Sutskever, P. Abbeel, and W. Zaremba. One-shot imitation learning. *ArXiv*, abs/1703.07326, 2017.

[25] S. James, M. Bloesch, and A. J. Davison. Task-embedded control networks for few-shot imitation learning. *ArXiv*, abs/1810.03237, 2018.

[26] C. Finn, T. Yu, T. Zhang, P. Abbeel, and S. Levine. One-shot visual imitation learning via meta-learning. In *Conference on robot learning*, 2017.

[27] P. Englert and M. Toussaint. Learning manipulation skills from a single demonstration. *The International Journal of Robotics Research*, 37(1):137–154, 2018.

[28] Octo Model Team, D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, C. Xu, J. Luo, T. Kreiman, Y. Tan, D. Sadigh, C. Finn, and S. Levine. Octo: An open-source generalist robot policy. https://octo-models.github.io, 2023.

[29] Z. Fu, T. Z. Zhao, and C. Finn. Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation, 2024.

[30] J. Pari, N. M. Shafiullah, S. P. Arunachalam, and L. Pinto. The surprising effectiveness of representation learning for visual imitation. *arXiv preprint arXiv:2112.01511*, 2021.

[31] S. Nasiriany, T. Gao, A. Mandlekar, and Y. Zhu. Learning and retrieval from prior data for skill-based imitation learning, 2022.

[32] M. Du, S. Nair, D. Sadigh, and C. Finn. Behavior retrieval: Few-shot imitation learning by querying unlabeled datasets, 2023.

[33] C. Lynch, M. Khansari, T. Xiao, V. Kumar, J. Tompson, S. Levine, and P. Sermanet. Learning latent plans from play. In *Conference on robot learning*, pages 1113–1132. PMLR, 2020.

[34] C. Wang, L. Fan, J. Sun, R. Zhang, L. Fei-Fei, D. Xu, Y. Zhu, and A. Anandkumar. Mimicplay: Long-horizon imitation learning by watching human play. *arXiv preprint arXiv:2302.12422*, 2023.

[35] Z. J. Cui, Y. Wang, N. M. M. Shafiullah, and L. Pinto. From play to policy: Conditional behavior generation from uncurated robot data. *arXiv preprint arXiv:2210.10047*, 2022.

[36] E. Rosete-Beas, O. Mees, G. Kalweit, J. Boedecker, and W. Burgard. Latent plans for task-agnostic offline reinforcement learning. In *Conference on Robot Learning*, pages 1838–1849. PMLR, 2023.

[37] H. Xiong, Q. Li, Y.-C. Chen, H. Bharadhwaj, S. Sinha, and A. Garg. Learning by watching: Physical imitation of manipulation skills from human videos. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7827–7834. IEEE, 2021.

[38] N. Das, S. Bechtle, T. Davchev, D. Jayaraman, A. Rai, and F. Meier. Model-based inverse reinforcement learning from visual demonstrations. In *Conference on Robot Learning*, pages 1930–1942. PMLR, 2021.

[39] L. Smith, N. Dhawan, M. Zhang, P. Abbeel, and S. Levine. Avid: Learning multi-stage tasks via pixel-level translation of human videos. *arXiv preprint arXiv:1912.04443*, 2019.

[40] L. Shao, T. Migimatsu, Q. Zhang, K. Yang, and J. Bohg. Concept2robot: Learning manipulation concepts from instructions and human demonstrations. *The International Journal of Robotics Research*, 40(12-14):1419–1434, 2021.

[41] A. S. Chen, S. Nair, and C. Finn. Learning generalizable robotic reward functions from" in-the-wild" human videos. *arXiv preprint arXiv:2103.16817*, 2021.

[42] A. D. Edwards and C. L. Isbell. Perceptual values from observation. *arXiv preprint arXiv:1905.07861*, 2019.

[43] A. Zeng, P. R. Florence, J. Tompson, S. Welker, J. Chien, M. Attarian, T. Armstrong, I. Krasin, D. Duong, V. Sindhwani, and J. Lee. Transporter networks: Rearranging the visual world for robotic manipulation. In *Conference on Robot Learning*, 2020.

[44] E. Johns. Coarse-to-fine imitation learning: Robot manipulation from a single demonstration. *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4613–4619, 2021.

[45] M. Shridhar, L. Manuelli, and D. Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. *ArXiv*, abs/2209.05451, 2022.

[46] O. Özgüner, T. Shkurti, S. Huang, R. Hao, R. C. Jackson, W. S. Newman, and M. C. Çavuşoğlu. Camera-robot calibration for the da vinci robotic surgery system. *IEEE Transactions on Automation Science and Engineering*, 17(4):2154–2161, 2020. doi:10.1109/TASE.2020.2986503.

[47] K. Pachtrachai, M. Allan, V. Pawar, S. Hailes, and D. Stoyanov. Hand-eye calibration for robotic assisted minimally invasive surgery without a calibration object. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2485–2491, 2016. doi:10.1109/IROS.2016.7759387.

[48] Z. Wang, Z. Liu, Q. Ma, A. Cheng, Y.-h. Liu, S. Kim, A. Deguet, A. Reiter, P. Kazanzides, and R. H. Taylor. Vision-based calibration of dual rcm-based robot arms in human-robot collaborative minimally invasive surgery. *IEEE Robotics and Automation Letters*, 3(2):672–679, 2018. doi:10.1109/LRA.2017.2737485.

[49] A. Shademan, R. S. Decker, J. D. Opfermann, S. Leonard, A. Krieger, and P. C. Kim. Supervised autonomous robotic soft tissue surgery. *Science translational medicine*, 8(337): 337ra64–337ra64, 2016.

[50] H. Saeidi, H. N. Le, J. D. Opfermann, S. Léonard, A. Kim, M. H. Hsieh, J. U. Kang, and A. Krieger. Autonomous laparoscopic robotic suturing with a novel actuated suturing tool and 3d endoscope. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 1541–1547. IEEE, 2019.

[51] F. Zhong, Y. Wang, Z. Wang, and Y.-H. Liu. Dual-arm robotic needle insertion with active tissue deformation for autonomous suturing. *IEEE Robotics and Automation Letters*, 4(3):2669–2676, 2019.

[52] S. A. Pedram, P. Ferguson, J. Ma, E. Dutson, and J. Rosen. Autonomous suturing via surgical robot: An algorithm for optimal selection of needle diameter, shape, and path. In *2017 IEEE International conference on robotics and automation (ICRA)*, pages 2391–2398. IEEE, 2017.

[53] C. Gruijthuijsen, L. C. Garcia-Peraza-Herrera, G. Borghesan, D. Reynaerts, J. Deprest, S. Ourselin, T. Vercauteren, and E. Vander Poorten. Robotic endoscope control via autonomous instrument tracking. *Frontiers in Robotics and AI*, 9:832208, 2022.

[54] X. Ma, C. Song, P. W. Chiu, and Z. Li. Autonomous flexible endoscope for minimally invasive surgery with enhanced safety. *IEEE Robotics and Automation Letters*, 4(3):2607–2613, 2019.

[55] J. W. Kim, P. Zhang, P. Gehlbach, I. Iordachita, and M. Kobilarov. Towards autonomous eye surgery by combining deep imitation learning with optimal control. In J. Kober, F. Ramos, and C. Tomlin, editors, *Proceedings of the 2020 Conference on Robot Learning*, volume 155 of *Proceedings of Machine Learning Research*, pages 2347–2358. PMLR, 16–18 Nov 2021. URL `https://proceedings.mlr.press/v155/kim21a.html`.

[56] J. W. Kim, C. He, M. Urias, P. Gehlbach, G. D. Hager, I. Iordachita, and M. Kobilarov. Autonomously navigating a surgical tool inside the eye by learning from demonstration. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7351–7357, 2020. doi:10.1109/ICRA40945.2020.9196537.

[57] J. W. Kim, S. Wei, P. Zhang, P. Gehlbach, J. U. Kang, I. Iordachita, and M. Kobilarov. Towards autonomous retinal microsurgery using rgb-d images. *IEEE Robotics and Automation Letters*, 9 (4):3807–3814, 2024. doi:10.1109/LRA.2024.3368192.

[58] E. Tagliabue, A. Pore, D. Dall'Alba, E. Magnabosco, M. Piccinelli, and P. Fiorini. Soft tissue simulation environment to learn manipulation tasks in autonomous robotic surgery. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3261–3266. IEEE, 2020.

[59] C. Shin, P. W. Ferguson, S. A. Pedram, J. Ma, E. P. Dutson, and J. Rosen. Autonomous tissue manipulation via surgical robot using learning based model predictive control. In *2019 International conference on robotics and automation (ICRA)*, pages 3875–3881. IEEE, 2019.

[60] H. Saeidi, J. D. Opfermann, M. Kam, S. Wei, S. Léonard, M. H. Hsieh, J. U. Kang, and A. Krieger. Autonomous robotic laparoscopic surgery for intestinal anastomosis. *Science robotics*, 7(62):eabj2908, 2022.

[61] P. M. Scheikl, N. Schreiber, C. Haas, N. Freymuth, G. Neumann, R. Lioutikov, and F. Mathis-Ullrich. Movement primitive diffusion: Learning gentle robotic manipulation of deformable objects. *IEEE Robotics and Automation Letters*, 9(6):5338–5345, June 2024. ISSN 2377-3774. doi:10.1109/lra.2024.3382529. URL http://dx.doi.org/10.1109/LRA.2024.3382529.

[62] K. Kawaharazuka, K. Okada, and M. Inaba. Robotic constrained imitation learning for the peg transfer task in fundamentals of laparoscopic surgery, 2024. URL https://arxiv.org/abs/2405.03440.

[63] B. Li, R. Wei, J. Xu, B. Lu, C.-H. Yee, C.-F. Ng, P.-A. Heng, Q. Dou, and Y.-H. Liu. 3d perception based imitation learning under limited demonstration for laparoscope control in robotic surgery, 2022. URL https://arxiv.org/abs/2204.03195.

[64] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023.

[65] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[66] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015.

[67] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. *ArXiv*, abs/1505.04597, 2015. URL https://api.semanticscholar.org/CorpusID:3719281.

A Implementation Details

For ACT, main modifications include changing the input layers to accept four images, which include left/right surgical endoscope views and left/right wrist camera views. The output dimensions are also revised to generate end-effector poses, which amounts to a 10-dim vector for each arm (position [3] + orientation [6] + jaw angle [1] = 10), thus amounting to a 20-dim vector total for both arms. The orientation was modeled using a 6D rotation representation following [21], where the 6 elements corrrespond to the first two columns of the rotation matrix. Since the network predictions may not generate orthonormal vectors, Gram-Schmidt process is performed to convert them to orthonormal vectors, and a cross product of the two vectors are performed to generate the remaining third column of the rotation matrix. For diffusion policy, similar modifications are made such as changing the input and the output dimensions of the network appropriately. The specific hyperparameters for training are shown in Table 3 and 4.

learning rate	1e-5
batch size	8
# encoder layers	4
# decoder layers	7
feedforward dimension	3200
hidden dimension	512
# heads	8
chunk size	100
beta	10
dropout	0.1

Table 3: Hyperparameters of ACT.

learning rate	1e-4
batch size	64
chunk size	32
scheduler	DDIM[65]
train and test diffusion steps	100, 100
ema power	0.75
backbone	ResNet18[66]
noise predictor	UNet[67]
image augmentation	RandomCrop(ratio=0.95) & ColorJitter(brightness=0.3, contrast=0.4, saturation=0.5) & RandomRotation(degrees=[-5.0, 5.0])

Table 4: Hyperparameters of Diffusion Policy.

B Generalization to Novel Settings

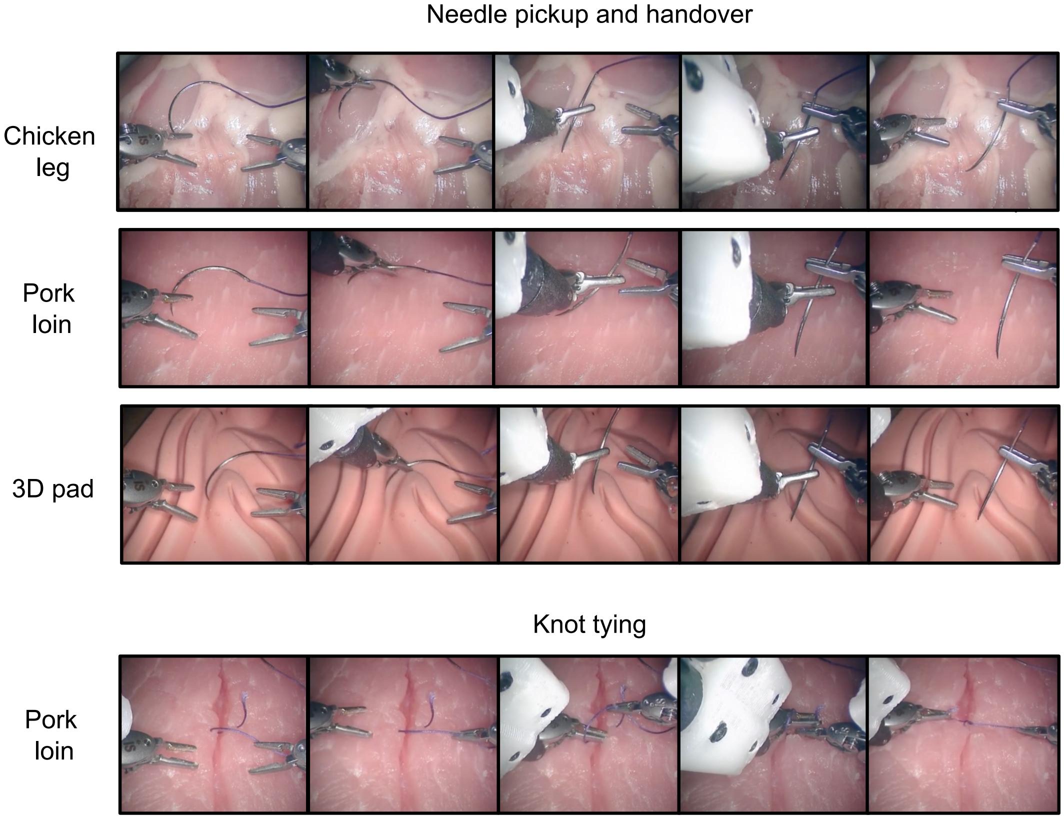

Figure 7: We show qualitative examples of our model generalizing to novel scenarios beyond training settings. *(Top):* Successful zero-shot needle pickup and handover on chicken leg, pork loin, and on a 3D pad. *(Bottom):* Successful zero-shot knot-tyng on pork loin.