SurgVLA-Bench: Towards Evaluating Vision-Language-Action Models for Laparoscopic Surgical Robotics

Jiashuo Sun[1,2⋆], Yue He[1,3⋆], Wenxuan Liu[1,3], Tao Mao[1,3], Jiazheng Wang[1,3], Xiang Chen[1,3], and Min Liu[1,2,3†]

[1] the National Engineering Research Center of Robot Visual Perception and Control Technology, Hunan, China
[2] the National Graduate College for Elite Engineers, Hunan University, Hunan, China
[3] the School of Artificial Intelligence and Robotics, Hunan University, Hunan, China
SunJiashuo@hnu.edu.cn, liu_min@hnu.edu.cn

Abstract. Vision-Language-Action (VLA) models represent a promising direction for embodied intelligence in surgical robotics. Despite the prevalence of VLA benchmarks for general robotics, standardized evaluation platforms specifically designed for surgical contexts remain absent. To address this limitation, we present SurgVLA-Bench, the first comprehensive benchmark for evaluating VLA models in laparoscopic surgical robotics. Leveraging the SurRoL simulation platform, we construct a hierarchical task taxonomy ranging from atomic actions to complete surgical procedures, complemented by a multi-dimensional evaluation framework assessing action accuracy and semantic consistency. We then systematically evaluate two representative paradigms, including autoregressive models such as OpenVLA, and flow matching models such as π_0, $\pi_{0.5}$, and SmolVLA. Our experiments show that autoregressive models tend to excel in semantic understanding, while flow matching models often achieve higher task precision but may face generalization trade-offs. However, even the best-performing models remain far from satisfactory, as the constrained endoscopic field of view, restricted viewing angles, and frequent occlusions persist as fundamental physical bottlenecks. The code and data are available at https://github.com/VCL-HNU/SurgVLA .

Keywords: Vision-Language-Action Models · Benchmark Dataset · Surgical Benchmarking · Simulation Environment.

1 Introduction

Surgical robots have become increasingly integral to modern clinical practice, offering enhanced precision, dexterity, and stability over conventional manual procedures. However, current systems still rely heavily on continuous teleoperation by experienced surgeons, motivating pursuit of higher autonomy levels.

⋆ These authors contributed equally to this work.
† Corresponding author.

In this context, Vision-Language-Action (VLA) models [28] have emerged as a promising paradigm, demonstrating remarkable capabilities in understanding language instructions, executing task-specific actions, and generalizing across diverse scenarios [1]. This raises a question: Can VLA models enable intelligent surgical robots to understand commands and perform precise interventions?

The field of VLA models has recently experienced significant advances. RT-2 [30] pioneered the approach of using vision-language model output tokens directly as robot actions. OpenVLA [12] introduced the first open-source VLA framework, providing architectural insights for subsequent model development. PaLM-E [5] demonstrated the feasibility of training VLA models on internet-scale data. π_0 [2] introduced flow matching [14] with multi-stage training, showing positive effects on task generalization. SmolVLA [21] and VLA-Adapter [24] explored lightweight approaches for efficient VLA deployment, while RyNN-VLA [3] combined embodied models with world models to enhance environmental understanding. However, these models are predominantly developed for general-purpose scenarios and lack applicability to surgical contexts, including the ability to discriminate surgical-specific objects, meet millimeter-level precision requirements, and address safety concerns arising from model hallucinations [13].

Research on surgical autonomy has primarily relied on task-specific methods, including imitation learning for suturing and peg transfer [20, 22], reinforcement learning in simulation [25, 26] and on physical platforms [11], and vision-language pretraining for surgical scene understanding [27]. While effective within their respective scopes, these approaches each address only a subset of the perception-language-action pipeline, and none provides a unified framework that jointly integrates visual understanding, language instruction following, and action generation. On the evaluation side, general-purpose robotics benchmarks such as LIBERO [15], CALVIN [17], and RLBench [9] have been instrumental in driving VLA research by providing standardized task suites and reproducible evaluation protocols. In the surgical domain, datasets such as Cholec80 [23] and CholecT50 [18] have advanced skill assessment and scene recognition, yet they focus on understanding rather than closed-loop action generation. No existing benchmark simultaneously evaluates visual perception, language instruction following, and action generation in surgical environments.

To bridge this gap, we propose SurgVLA-Bench, aiming to explore the capability of VLA models to understand surgeon instructions and execute precise interventions in surgical scenarios. This is the first VLA evaluation benchmark specifically designed for surgical contexts. Built upon the SurRoL [25] platform, our benchmark encompasses a comprehensive suite of evaluations ranging from atomic action tasks to composite surgical procedure segments.

Our main contributions are summarized as follows:

1. We propose SurgVLA-Bench, the first benchmark for surgical VLA models, featuring a hierarchical task taxonomy from atomic actions to full procedures and built on simulated environments with surgical instruments and multi-organ anatomical structures.

2. We construct a comprehensive standardized dataset supporting multiple mainstream formats (RLDS [19], LeRobot [7], etc.), facilitating direct adoption and extension by the research community.
3. We benchmark four representative VLAs [12, 21, 8, 2], analyzing their limitations in surgical contexts and identifying key challenges for future research.

2 Benchmarking

2.1 Tasks

Surgical procedures are inherently hierarchical: clinical workflows combine basic actions into complex operations, with different levels imposing distinct capability demands. To systematically evaluate VLA models in surgical scenarios and identify their capability bottlenecks, we design a clinically relevant hierarchical evaluation system encompassing eight tasks. Based on complexity, tasks are categorized into three levels (Fig. 1). The design follows three principles: (1) gradual difficulty progression; (2) coverage of typical surgical operations; (3) support for independent evaluation of different VLA capability dimensions. Each level targets distinct capabilities: Level 1 focuses on action precision, Level 2 on target discrimination and instruction following, and Level 3 on multi-step planning and spatial localization.

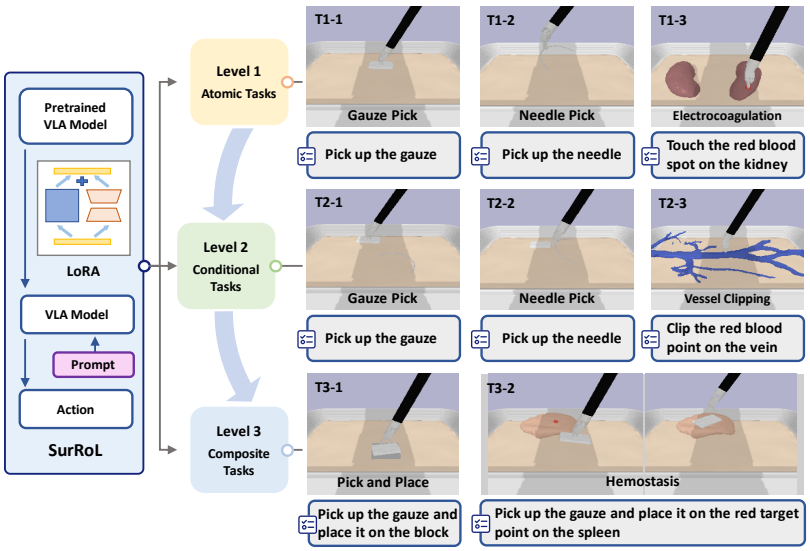

Fig. 1. Overview of SurgVLA-Bench. The left panel illustrates the fine-tuning and deployment pipeline, where a pretrained VLA model is fine-tuned via LoRA and deployed in the SurRoL simulation environment. The right panel presents the hierarchical task taxonomy spanning three difficulty levels: Atomic, Conditional, and Composite Tasks.

Level 1: Atomic Tasks. The scene contains only one target object without distractors, and language instructions are direct and unambiguous. This level primarily evaluates action precision and basic control capabilities. It includes three tasks: Gauze Pick (T1-1) and Needle Pick (T1-2), which simulate basic grasping of surgical consumables, and Electrocoagulation (T1-3), which simulates contact-based cauterization of a target tissue site.

Level 2: Conditional Tasks. The scene contains distractors, or the target object itself presents multiple confounding factors, requiring the model to accurately identify the manipulation target based on language instructions. This level primarily evaluates target discrimination capability and instruction-following accuracy. It includes: Gauze Pick (T2-1) and Needle Pick (T2-2), which extend their Level 1 counterparts by introducing visually similar distractors into the scene, and Vessel Clipping (T2-3), which simulates the placement of a hemostatic clip on a designated vessel segment among multiple candidate sites.

Level 3: Composite Tasks. The model must complete comprehensive operational workflows, placing objects at designated target positions. This level primarily evaluates multi-step sequential execution capability and precise spatial localization ability. It includes: Pick and Place (T3-1), which requires grasping a target object and transferring it to a specified location, and Hemostasis (T3-2), which simulates a complete hemostatic procedure involving sequential gauze grasping, transport, and placement onto a bleeding site on an organ surface.

2.2 Dataset Construction

To support the evaluation of SurgVLA-Bench, we construct a dedicated surgical task dataset. Unlike general robotics, where large-scale open-source datasets such as Open X-Embodiment [19] are readily available for VLA model training and evaluation, the surgical domain currently lacks standardized datasets suitable for this purpose. Therefore, we collect and construct our dataset from scratch based on task scenarios within the SurRoL [25] simulation environment.

Data Collection. We collect trajectories for all eight tasks defined in Section 2.1, organized into three datasets corresponding to the three difficulty levels. Taking T3-2 as an example, the data collection process is illustrated in the Fig. 2. We randomly position the target within a predefined area and execute the specified actions along preset trajectories based on the object's location. Each trajectory contains RGB images, depth images, robot proprioceptive states, and corresponding action sequences. For consistency, all RGB images are captured at a resolution of $[3 \times 224 \times 224]$. In total, we collect over 800 complete trajectories comprising approximately 40,000 action frames across all eight tasks, with each trajectory requiring manual verification to ensure action quality. Since the primary objective at this stage is to verify the basic feasibility of VLA models on the benchmark, we adopt concise and unambiguous instruction descriptions for each task (see Fig. 1), ensuring a clear correspondence between language instructions and task objectives. Additionally, we develop an integrated data collection pipeline that enables future developers to expand the dataset scale according to

their needs. We provide the data in LeRobot [7], RLDS [19], and 3D point cloud formats to facilitate direct adoption across different VLA paradigms.

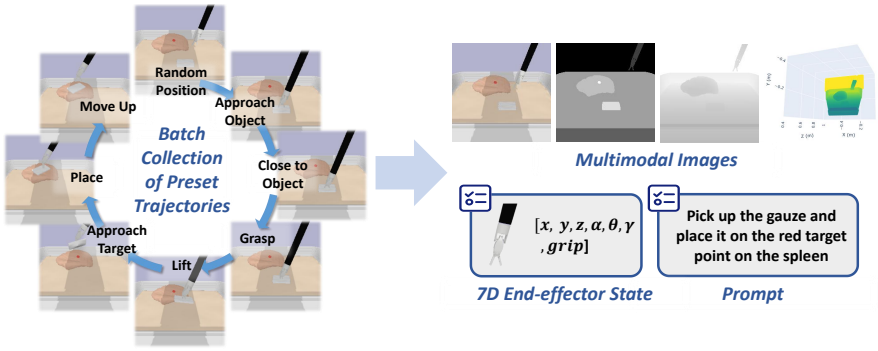

Fig. 2. Data Acquisition Pipeline and Dataset Composition.

Gripper State Representation. To mitigate the sparsity of the original pulse-like gripper signal, we convert the gripper action from instantaneous change values to continuous state values (i.e., persistently outputting closing or opening signals), which significantly improves training efficiency.

2.3 Simulation Environment

We build our simulation scenarios upon the SurRoL platform, which has shown efficacy in enabling sim-to-real transfer for surgical tasks [16]. First, we integrate multiple high-fidelity anatomical organ models, including liver, kidney, and venous vasculature, and combine them with existing instrument models from the SurRoL platform to build diverse surgical scenes. To enhance anatomical realism, we ensure that all organ models maintain accurate anatomical proportions. Beyond the organ models used in task scenarios, we provide an additional library of abdominal cavity organ models covering major intra-abdominal organs, facilitating future extensions by developers.

Furthermore, SurRoL natively employs the PyBullet [4] physics engine, whose contact detection is primarily designed for rigid bodies and has limited effectiveness for simulating deformable objects. So we optimize the grasping mechanism to address PyBullet's limited effectiveness for deformable objects. Informed by real-world experiments, we replace physics-based contact detection with a task-specific binding mechanism that triggers grasping when spatial and kinematic conditions derived from real-world observations are jointly satisfied, yielding manipulation outcomes that more faithfully reflect real surgical robot behavior.

3 Experiments

3.1 Implementation Details

Model Selection. To comprehensively evaluate VLA model performance in surgical scenarios, We select four representative VLA models as baselines, covering two mainstream paradigms: OpenVLA [12], built on Prismatic-7B [10], represents the autoregressive approach; the π-series [8, 2] and SmolVLA [21] ($\sim$0.5B parameters) adopt Flow Matching action heads, with $\pi_{0.5}$ adding a dual-layer policy architecture. Architectural details are summarized in Table 1.

Table 1. Architectural details of baseline VLA models evaluated in SurgVLA-Bench.

VLA Model	Year	Vision Encoder	LLM	Action Head
OpenVLA [12]	2024	SigLIP + DINOv2	Llama 2 7B	Autoregressive
π_0 [2]	2024	SigLIP 400M	Gemma 2B	Flow Matching
$\pi_{0.5}$ [8]	2025	SigLIP2	Gemma 2	Flow Matching
SmolVLA [21]	2025	SigLIP	SmolLM2	Flow Matching

Experimental Setup. As illustrated in Fig. 1, we design language instructions in a colloquial style for both atomic actions and tasks that simulate real surgical operations. The instructions are phrased in natural conversational language rather than strict medical terminology to facilitate LLM comprehension of surgical scenarios, thereby maximizing instruction following in VLA models.

All experiments are conducted on 4 NVIDIA A6000 GPUs. For each task, 50 independent trials were conducted per model. Following the evaluation protocols established in prior work [12, 21, 2, 8], we report the resulting success rate (SR) as the primary evaluation metric. The evaluation employs a binary success/failure judgment under step limits without time constraints. The maximum allowable steps are set to 100 for Level 1 and Level 2 tasks, and 150 for Level 3 tasks. To ensure fair comparison, all models are fine-tuned using Low-rank adaptation (LoRA) [6] until convergence. Specifically, each model is trained separately on a per-level basis and subsequently evaluated on individual sub-tasks within either the same level or across different levels. All models are evaluated under identical simulation environments, task configurations, and evaluation protocols, allowing each model to fully demonstrate its performance potential. We omit zero-shot baselines as preliminary tests yielded a 0% success rate. This stems from a significant domain gap, as models struggle with surgical instruments and kinematics distinct from industrial settings.

3.2 Experiments and Analysis

Analysis of Success Rate. As shown in Table 2, OpenVLA achieves competitive overall performance after fine-tuning. While its precision on atomic tasks

Table 2. Results on SurgVLA-Bench. For each evaluation task, best results per column are highlighted in bold.

Model	Level 1			Level 2			Level 3	
	T1-1	T1-2	T1-3	T2-1	T2-2	T2-3	T3-1	T3-2
OpenVLA [12]	36%	2%	**76%**	8%	2%	**72%**	0%	0%
π_0 [2]	76%	0%	0%	0%	4%	6%	**56%**	**2%**
$\pi_{0.5}$ [8]	**78%**	**8%**	0%	**36%**	**10%**	14%	12%	0%
SmolVLA [21]	0%	8%	0%	0%	0%	10%	0%	0%

is moderate, it effectively identifies surgical tools and anatomical structures, excelling on organ-involved tasks T1-3 and T2-3. OpenVLA localizes well but struggles with gripper closure, impairing gauze grasping. For needle grasping, the needle's slender, curved geometry hinders all models from learning correct gripper rotation, causing uniformly poor performance.

The π-series models handle single-object gauze grasping (T1-1) but fail on needle (T1-2) and electrocautery (T1-3) tasks, reflecting limited multi-task semantic discrimination, indicating limited multi-task semantic discrimination. This limitation is further evidenced by their degraded performance on Level 2 tasks involving multiple objects, where $\pi_{0.5}$ shows improvement over π_0. However, π_0 achieves strong results on the longer-horizon task T3-1 at Level 3, indicating its ability to execute multi-step tasks without distractions. In contrast, $\pi_{0.5}$ exhibits lower success rates on this task, likely due to its hybrid action representation combining discrete tokens for high-level semantic planning and continuous flow matching for low-level control. The lack of high-level supervision during fine-tuning may hinder its performance on multi-stage long-horizon tasks. Both π-series models perform poorly on organ-involved tasks, likely attributed to the domain gap between surgical scenes and pre-training data, where similar anatomical priors are absent.

As a lightweight model, SmolVLA struggles with scene understanding in multi-task settings, often failing to execute proper gripper actions. It consistently reaches the target position in simple scenarios like T1-1, with failures limited only to gripper closure. To investigate whether this performance gap stems from multi-task interference, we further evaluated SmolVLA in a single-task setting, where it achieved an 82% SR on T1-1, confirming its efficacy in simplified tasks.

Analysis of Prompt Robustness. To assess prompt ambiguity effects, we rephrased instructions for OpenVLA [12] and $\pi_{0.5}$[8] ("lift" instead of "pick up", "contact" instead of "touch"). As shown in Fig. 3(a), OpenVLA exhibits limited performance degradation under instruction perturbations, with improved success rates on modified gauze grasping commands. In contrast, $\pi_{0.5}$ shows larger performance drops on task T1-1, indicating that the LLM of OpenVLA maintains superior robustness after fine-tuning.

Analysis of Prompt Informativeness Utilization. To explore why the π-series models[2, 8] struggle with multi-task generalization after LoRA fine-tuning,

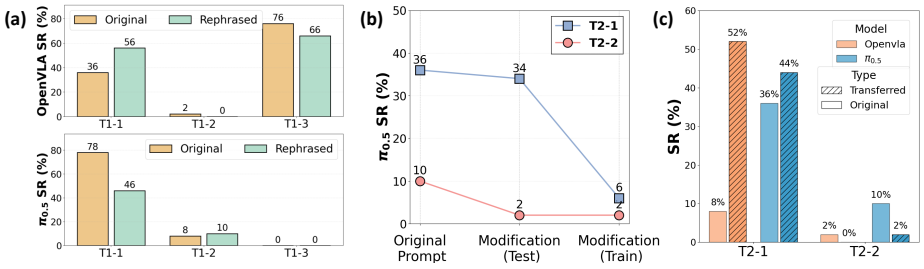

Fig. 3. Model analysis experiments. (a) Performance of OpenVLA and $\pi_{0.5}$ under prompt reformulation; (b) Impact of detailed prompts on $\pi_{0.5}$ during test-time and training-time; (c) Direct transfer results of OpenVLA and $\pi_{0.5}$ from Level 1 to Level 2.

we test whether prompt modifications could help using two strategies: (1) using more detailed instructions at test time, and (2) expanding task descriptions during training (e.g., "pick up the gauze, not the needle"). As a more complex and higher-performing model than π_0, $\pi_{0.5}$ exhibits stronger coupling between its performance and architectural design. Consequently, LoRA fine-tuning disrupts its internal structure, degrading the LLM's instruction-following capability. As shown in Fig. 3(b), modifying prompts for $\pi_{0.5}$ during testing or training not only fails to improve performance but leads to degradation, particularly when prompts are modified during training. The model favors simpler prompts over more informative but complex ones. This suggests that preserving pre-trained language capabilities under low-cost fine-tuning remains an open challenge for sophisticated VLA architectures [29].

Analysis of Generalization. Both Level 1 and Level 2 contain gauze and needle objects, with Level 1 being simpler. To assess whether models learn object semantics, we directly apply models trained on level 1 to Level 2 tasks without fine-tuning, with results shown in Fig. 3(c). For T1-1 (gauze grasping), OpenVLA shows significant performance gain after transfer. This improvement likely stems from cleaner training signals in Level 1, which lacks interference from other objects, suggesting OpenVLA effectively learns semantic representations that transfer to more complex scenarios. $\pi_{0.5}$ maintains its performance after transfer, indicating it also captures gauze semantics. For T2-2 (needle grasping), both models exhibit slight performance degradation, suggesting that direct transfer compromises spatial precision for fine-grained manipulation tasks.

4 Conclusion

In this paper, we introduce SurgVLA-Bench, the first benchmark for evaluating Vision-Language-Action models in surgical robotics. Built on SurRoL, our hierarchical evaluation system spans three levels and eight tasks with a dedicated dataset and multi-dimensional protocol. Systematic evaluation of four representative VLA models reveals that none perform satisfactorily under multi-task

training and testing. Autoregressive models tend to excel at semantic understanding, whereas flow matching models often deliver higher precision in task execution. However, this strength in specificity may come at the cost of limited generalization. In surgical scenarios involving multiple complex sub-tasks, addressing training conflicts among different tasks remains a key direction for future VLA research. These limitations stem from task interference under multi-task training, inherent differences in robot kinematics, limited surgical data, and most critically, the constrained endoscopic view with restricted angles and occlusions that impede depth perception. We believe SurgVLA-Bench can serve as a standardized evaluation platform for VLA model research in surgical scenarios and help drive further advancements in this field. Future work will extend the benchmark with greater task diversity, and real surgical data validation.

Acknowledgments. This work was supported by the National Natural Science Foundation of China under Grant U22B2050 and 62425305.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Attanasio, A., Scaglioni, B., De Momi, E., Fiorini, P., Valdastri, P.: Autonomy in surgical robotics. Annual Review of Control, Robotics, and Autonomous Systems **4**(1), 651–679 (2021)
2. Black, K., Brown, N., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Groom, L., Hausman, K., Ichter, B., et al.: π_0: A vision-language-action flow model for general robot control. arXiv preprint arXiv:2410.24164 (2024)
3. Cen, J., Huang, S., Yuan, Y., Li, K., Yuan, H., Yu, C., Jiang, Y., Guo, J., Li, X., Luo, H., et al.: Rynnvla-002: A unified vision-language-action and world model. arXiv preprint arXiv:2511.17502 (2025)
4. Coumans, E., Bai, Y.: Pybullet, a python module for physics simulation for games, robotics and machine learning. https://pybullet.org (2016)
5. Driess, D., Xia, F., Sajjadi, M.S., Lynch, C., Chowdhery, A., Wahid, A., Tompson, J., Vuong, Q., Yu, T., Huang, W., et al.: Palm-e: An embodied multimodal language model (2023)
6. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=nZeVKeeFYf9
7. Hugging Face Team: LeRobot: A Library for Real-World Robot Learning. https://github.com/huggingface/lerobot (2024), version 0.4+ includes v3.0 dataset format
8. Intelligence, P., Black, K., Brown, N., Darpinian, J., Dhabalia, K., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., et al.: $\pi_{0.5}$: a vision-language-action model with open-world generalization. $\pi_{0.5}$: a vision-language-action model with open-world generalization (2025)

9. James, S., Ma, Z., Arrojo, D.R., Davison, A.J.: Rlbench: The robot learning benchmark & learning environment. IEEE Robotics and Automation Letters **5**(2), 3019–3026 (2020)
10. Karamcheti, S., Nair, S., Balakrishna, A., Liang, P., Kollar, T., Sadigh, D.: Prismatic vlms: Investigating the design space of visually-conditioned language models (2024), https://arxiv.org/abs/2402.07865
11. Kim, J.W., Zhao, T.Z., Schmidgall, S., Deguet, A., Kobilarov, M., Finn, C., Krieger, A.: Surgical robot transformer (srt): Imitation learning for surgical tasks. arXiv preprint arXiv:2407.12998 (2024)
12. Kim, M.J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E., Lam, G., Sanketi, P., et al.: Openvla: An open-source vision-language-action model. arXiv preprint arXiv:2406.09246 (2024)
13. Kim, Y., Jeong, H., Chen, S., Li, S.S., Park, C., Lu, M., Alhamoud, K., Mun, J., Grau, C., Jung, M., et al.: Medical hallucinations in foundation models and their impact on healthcare. arXiv preprint arXiv:2503.05777 (2025)
14. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
15. Liu, B., Zhu, Y., Gao, C., Feng, Y., Liu, Q., Zhu, Y., Stone, P.: Libero: Benchmarking knowledge transfer for lifelong robot learning. Advances in Neural Information Processing Systems **36**, 44776–44791 (2023)
16. Long, Y., Lin, A., Kwok, D.H.C., Zhang, L., Yang, Z., Shi, K., Song, L., Fu, J., Lin, H., Wei, W., et al.: Surgical embodied intelligence for generalized task autonomy in laparoscopic robot-assisted surgery. Science Robotics **10**(104), eadt3093 (2025)
17. Mees, O., Hermann, L., Rosete-Beas, E., Burgard, W.: Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. IEEE Robotics and Automation Letters **7**(3), 7327–7334 (2022)
18. Nwoye, C.I., Alapatt, D., Yu, T., Vardazaryan, A., Xia, F., Zhao, Z., Xia, T., Jia, F., Yang, Y., Wang, H., et al.: Cholectriplet2021: A benchmark challenge for surgical action triplet recognition. Medical Image Analysis **86**, 102803 (2023)
19. O'Neill, A., Rehman, A., Maddukuri, A., Gupta, A., Padalkar, A., Lee, A., Pooley, A., Gupta, A., Mandlekar, A., Jain, A., et al.: Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 6892–6903. IEEE (2024)
20. Pore, A., Tagliabue, E., Piccinelli, M., Dall'Alba, D., Casals, A., Fiorini, P.: Learning from demonstrations for autonomous soft-tissue retraction. In: 2021 international symposium on medical robotics (ISMR). pp. 1–7. IEEE (2021)
21. Shukor, M., Aubakirova, D., Capuano, F., Kooijmans, P., Palma, S., Zouitine, A., Aractingi, M., Pascal, C., Russi, M., Marafioti, A., et al.: Smolvla: A vision-language-action model for affordable and efficient robotics. arXiv preprint arXiv:2506.01844 (2025)
22. Soper, N.J., Fried, G.M.: The fundamentals of laparoscopic surgery: its time has come. Bull Am Coll Surg **93**(9), 30–32 (2008)
23. Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M., Padoy, N.: Endonet: a deep architecture for recognition tasks on laparoscopic videos. IEEE transactions on medical imaging **36**(1), 86–97 (2016)
24. Wang, Y., Ding, P., Li, L., Cui, C., Ge, Z., Tong, X., Song, W., Zhao, H., Zhao, W., Hou, P., et al.: Vla-adapter: An effective paradigm for tiny-scale vision-language-action model. arXiv preprint arXiv:2509.09372 (2025)

25. Xu, J., Li, B., Lu, B., Liu, Y.H., Dou, Q., Heng, P.A.: Surrol: An open-source reinforcement learning centered and dvrk compatible platform for surgical robot learning. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 1821–1828. IEEE (2021)
26. Yu, Q., Moghani, M., Dharmarajan, K., Schorp, V., Panitch, W.C.H., Liu, J., Hari, K., Huang, H., Mittal, M., Goldberg, K., et al.: Orbit-surgical: An open-simulation framework for learning surgical augmented dexterity. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 15509–15516. IEEE (2024)
27. Yuan, K., Srivastav, V., Navab, N., Padoy, N.: Hecvl: Hierarchical video-language pretraining for zero-shot surgical phase recognition. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 306–316. Springer (2024)
28. Zhang, D., Sun, J., Hu, C., Wu, X., Yuan, Z., Zhou, R., Shen, F., Zhou, Q.: Pure vision language action (vla) models: A comprehensive survey. arXiv preprint arXiv:2509.19012 (2025)
29. Zhuoling, L., Liangliang, R., Jinrong, Y., Yong, Z., et al.: Vip: Vision instructed pre-training for robotic manipulation. ICML (2025)
30. Zitkovich, B., Yu, T., Xu, S., Xu, P., Xiao, T., Xia, F., Wu, J., Wohlhart, P., Welker, S., Wahid, A., et al.: Rt-2: Vision-language-action models transfer web knowledge to robotic control. In: Conference on Robot Learning. pp. 2165–2183. PMLR (2023)