arXiv:2604.21017v3 [cs.RO] 4 Jun 2026

Open-H-Embodiment: A Large-Scale Dataset for Enabling Foundation Models in Medical Robotics

Nigel Nelson[1,†], Juo-Tung Chen[2,†], Jesse Haworth[2,†], Xinhao Chen[2,†], Lukas Zbinden[1,†], Dianye Huang[3,†], Alaa Eldin Abdelaal[4], Alberto Arezzo[5], Ayberk Acar[6], Farshid Alambeigi[7], Carlo Alberto Ammirati[5], Yunke Ao[8,9,10], Pablo David Aranda Rodriguez[11], Soofiyan Atar[12], Mattia Ballo[13], Noah Barnes[2], Federica Barontini[5], Filip Binkiewicz[14], Peter Black[15,16], Sebastian Bodenstedt[17,18], Leonardo Borgioli[19], Nikola Budjak[11], Benjamin Calmé[20], Fabio Carrillo[8], Nicola Cavalcanti[8], Changwei Chen[12], Haoxin Chen[21], Sihang Chen[22], Qihan Chen[23], Zhongyu Chen[24,25], Ziyang Chen[26], Shing Shin Cheng[24], Meiqing Cheng[27], Min Cheng[28,22], Zih-Yun Sarah Chiu[2], Xiangyu Chu[24,25], Camilo Correa-Gallego[29], Giulio Dagnino[5,30], Anton Deguet[2], Jacob Delgado[2], Jonathan C. DeLong[31], Kaizhong Deng[32], Alexander Dimitrakakis[29], Qingpeng Ding[24], Hao Ding[2,33], Giovanni Distefano[5], Daniel Donoho[34], Anqing Duan[35], Marco Esposito[11], Shane Farritor[36], Jad Fayad[37], Zahi Fayad[29], Mario Ferradosa[38], Filippo Filicori[39], Chelsea Finn[4,40], Philipp Fürnstahl[8,41], Jiawei Ge[2], Stamatia Giannarou[32], Xavier Giralt Ludevid[38], Frederic Giraud[8], Aditya Amit Godbole[42], Ken Goldberg[26], Antony Goldenberg[2], Diego Granero Marana[14], Xiaoqing Guo[21], Tamás Haidegger[43,44], Evan Hailey[36], Pascal Hansen[18], Ziyi Hao[24], Kush Hari[26], Kengo Hayashi[5], Jonathon Hawkins[14], Shelby Haworth[2], Ortrun Hellig[17], S. Duke Herrell[6], Zhouyang Hong[24], Andrew Howe[14], Junlei Hu[20], Zhaoyang Jacopo Hu[32], Ria Jain[26], Mohammad Rafiee Javazm[7], Howard Ji[4], Rui Ji[45], Jianmin Ji[22], Zhongliang Jiang[46,3], Dominic Jones[20], Jeffrey Jopling[2], Britton Jordan[6], Ran Ju[46,28], Michael Kam[2], Luoyao Kang[24], Fausto Kang[2], Siddhartha Kapuria[7], Peter Kazanzides[2], Aimal Khan[47], Sonika Kiehler[7], Ethan Kilmer[2], Ji Woong (Brian) Kim[2,4], Przemysław Korzeniowski[1,13], Chandra Kuchi[40], Nithesh Kumar[6], Alan Kuntz[6], Federico Lavagno[5], Yu Chung Lee[15], Hao-Chih Lee[29], Hang Li[24], Zhen Li[45], Xiao Liang[12], Xinxin Lin[27], Jinsong Lin[24], Chang Liu[2], Fei Liu[31], Pei Liu[3], Yun-hui Liu[24], Wanli Liuchen[23], Eszter Lukács[43,44], Sareena Mann[26], Miles Mannas[15,16], Brett Marinelli[29], Sabina Martyniak[13], Francesco Marzola[5], Lorenzo Mazza[17], Xueyan Mei[29], Maria Clara Morais[42], Luigi Muratore[5], Chetan Reddy Narayanaswamy[4], Michał Naskręt[13], David Navarro-Alarcon[23], Cyrus Neary[15], Chi Kit Ng[24], Christopher Nguan[15,16], David Noonan[37], Ki Hwan Oh[19], Tom Christian Olesch[31], Allison M. Okamura[4], Justin Opfermann[2], Matteo Pescio[5], Doan Xuan Viet Pham[2], Tito Porras[33], Hongliang Ren[24], Ariel Rodriguez Jimenez[17], Ferdinando Rodriguez y Baena[32], Septimiu E. Salcudean[15], Asmitha Sathya[2], Preethi Satish[26], Lalithkumar Seenivasan[2], Jiaqi Shao[4], Yiqing Shen[2,33], Yu Sheng[22], Lucy XiaoYang Shi[4,40], Zoe Soulé[17], Stefanie Speidel[17,18], Mingwu Su[24], Jianhao Su[32], Idris Sunmola[2], Kristóf Takács[43], Yunxi Tang[24,25], Patrick Thornycroft[14], Yu Tian[24], Jordan Thompson[6], Mehmet K. Turkcan[48], Mathias Unberath[2,33], Pietro Valdastri[20], Carlos Vives[38], Quan Vuong[40], Martin Wagner[17], Farong Wang[31], Wei Wang[27], Lidian Wang[22], Chung-Pang Wang[12], Guankun Wang[24], Junyi Wang[46], Erqi Wang[24], Ziyi Wang[24], Tanner Watts[6], Wolfgang Wein[11], Yimeng Wu[2], Zijian Wu[15], Hongjun Wu[2], Luohong Wu[8], Jie Ying Wu[6], Junlin Wu[2], Victoria Wu[37], Kaixuan Wu[24], Mateusz Wójcikowski[13], Yunye Xiao[11], Nan Xiao[31], Wenxuan Xie[24], Hao Yang[6], Tianqi Yang[24,25], Yinuo Yang[12], Menglong Ye[37], Ryan S. Yeung[15], Nural Yilmaz[2], Chim Ho Yin[24], Michael Yip[12], Rayan Younis[17], Chenhao Yu[33], Sayem Nazmuz Zaman[15], Milos Zefran[19], Han Zhang[2], Yuelin Zhang[24], Yidong Zhang[24], Yanyong Zhang[22], Xuyang Zhang[22], Yameng Zhang[46,25], Joyce Zhang[14], Ning Zhong[45], Peng Zhou[49], Haoying Zhou[2,50], Xiuli Zuo[45], Nassir Navab[3,‡], Mahdi Azizian[1,‡], Sean D. Huver[1,‡], Axel Krieger[2,33,‡]

[1]NVIDIA, [2]Johns Hopkins University, [3]Technical University of Munich, [4]Stanford University, [5]University of Turin, [6]Vanderbilt University, [7]The University of Texas at Austin, [8]Balgrist University Hospital, [9]ETH Zurich, [10]ETH AI Center, [11]ImFusion GmbH, [12]University of California San Diego, [13]Sano Centre for Computational Medicine, [14]CMR Surgical, [15]University of British Columbia, [16]Vancouver General Hospital, [17]CeTI/TU Dresden, [18]German Cancer Research Center, [19]University of Illinois Chicago, [20]University of Leeds, [21]Hong Kong Baptist University, [22]University of Science and Technology of China, [23]The Hong Kong Polytechnic University, [24]The Chinese University of Hong Kong, [25]Multi-scale Medical Robotics Center, [26]University of California Berkeley, [27]Sun Yat-Sen University, [28]Tuodao Medical Technology Co., Ltd, [29]Icahn School of Medicine at Mount Sinai, [30]University of Twente, [31]University of Tennessee Knoxville, [32]Imperial College London, [33]Semaphor Surgical, [34]Surgical Data Science Collective, [35]Mohamed bin Zayed University of Artificial Intelligence, [36]Virtual Incision, [37]Moon Surgical, [38]Rob Surgical, [39]Hofstra/Northwell School of Medicine, [40]Physical Intelligence, [41]University of Zurich, [42]Northwell Health, [43]Óbuda University, [44]Austrian Center for Medical Innovation and Technology, [45]Qilu Hospital of Shandong University, [46]The University of Hong Kong, [47]Vanderbilt University Medical Center, [48]Columbia University, [49]Great Bay University, [50]Worcester Polytechnic Institute

† Co-first authors. ‡ Co-senior authors.

Autonomous medical robots hold promise to improve patient outcomes, reduce provider workload, democratize access to care, and enable superhuman precision. However, autonomous medical robotics has been limited by a fundamental data problem: existing medical robotic datasets are small, single-embodiment, and rarely shared openly, restricting the development of foundation models that the field needs to advance. We introduce Open-H-Embodiment, the largest open dataset of medical robotic video with synchronized kinematics to date, spanning more than 50 institutions and multiple robotic platforms including the CMR Versius, Intuitive Surgical's da Vinci, da Vinci Research Kit (dVRK), Rob Surgical BiTrack, Virtual Incision's MIRA, Moon Surgical Maestro, and a variety of custom systems, spanning surgical manipulation, robotic ultrasound, and endoscopy procedures. We demonstrate the research enabled by this dataset through two foundation models. GR00T-H is the first open foundation vision-language-action model for medical robotics, which is the only evaluated model to achieve full end-to-end task completion on a structured suturing benchmark (25% of trials vs. 0% for all others) and achieves 64% average success across a 29-step ex vivo suturing sequence. We also train Cosmos-H-Surgical-Simulator, the first action-conditioned world model to enable multi-embodiment surgical simulation from a single checkpoint, spanning nine robotic platforms and supporting in silico policy evaluation and synthetic data generation for the medical domain. These results suggest that open, large-scale medical robot data collection can serve as critical infrastructure for the research community, enabling advances in robot learning, world modeling, and beyond.

Project page: `https://open-h.github.io/open-h-embodiment/`

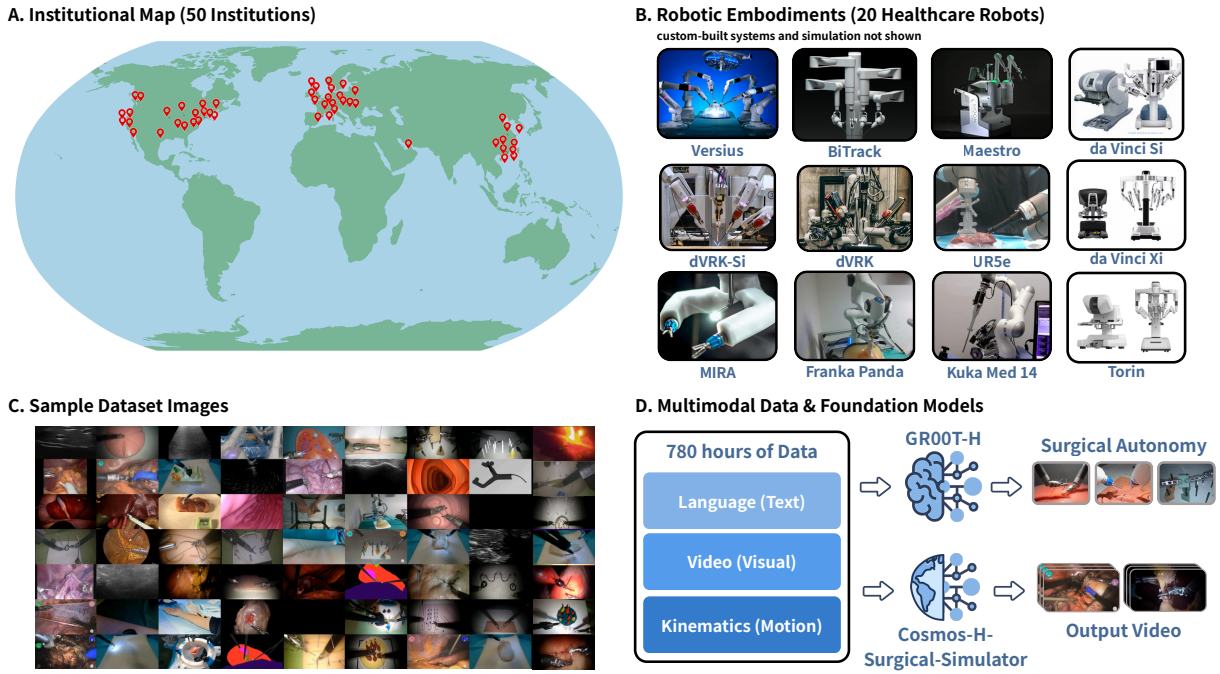


Figure 1: **Open-H-Embodiment overview.** (**A**) Geographic distribution of the 50 participating institutions across North America, Europe, the Middle East, and Asia. (**B**) The 20 healthcare robotic platforms represented in the dataset, spanning surgical systems (da Vinci Si, da Vinci Xi, dVRK, dVRK-Si, MIRA, Versius, BiTrack, Maestro, Torin), general-purpose manipulators adapted for clinical use (Franka Panda, UR5e, Kuka Med 14), and emerging platforms. (**C**) Representative frames from the dataset illustrating the diversity of tasks, viewpoints, and tissue types covered, including robotic surgery, robotic ultrasound, and related healthcare manipulation tasks. (**D**) The dataset comprises 780 hours of synchronized multimodal demonstrations spanning language annotations, video observations, and kinematic trajectories. This corpus supports two downstream directions: training GR00T-H, a healthcare-focused vision-language-action model targeting surgical autonomy, and training Cosmos-H-Surgical-Simulator, a multi-embodiment, action-conditioned world model for surgical scene synthesis.

Summary

Open-H: The largest medical robotic dataset of paired video and kinematics enabling multi-embodiment foundation models.

Introduction

In the global healthcare system, surgical patient outcomes vary tremendously and are heavily dependent on factors such as surgeon experience, workload and fatigue, and disparities in institutional resources. This variability has the potential to be exacerbated by a mounting global healthcare workforce crisis; in the United States alone, a deficit of 13,000 to 86,000 surgeons is projected by 2036 (*1, 2*). Autonomous robotic systems offer a critical pathway to mitigate surgeon workload, reduce burnout, and expand access to high-quality care, particularly in underserved regions or ambulatory surgery centers lacking specialized expertise. The path toward such autonomy is already underway; medical robots are increasingly deployed across the healthcare spectrum, taking on diverse roles ranging from patient rehabilitation to precise diagnostic imaging. While many of these systems operate with significant independence, surgical robots present a stark contrast. They remain almost entirely dependent on real-time human teleoperation, functioning at Level 1 (robot assistance) on surgical levels of autonomy (*3, 4*). Developing autonomous and semi-autonomous capabilities for high-volume procedures, such as cholecystectomy and suturing, provides a vital strategy to reduce performance variability and extend surgical capacity to underserved populations (*5*).

Recent advances in robot learning have enabled the autonomous execution of complex surgical tasks, including portions of ex vivo cholecystectomy (*6*) and in vivo tissue retraction and gauze packing (*7*). While these results demonstrate the potential for autonomous systems in clinical procedures, current models remain fragile, failing when confronted with procedural variations and struggling to generalize across different surgical tasks. To achieve scalable autonomy, the surgical robotics field must adopt the paradigms currently driving general artificial intelligence. The broader machine learning community has established a powerful empirical principle: large-scale, diverse pretraining datasets unlock generalist mod-

els that significantly outperform narrow policies, fine-tune efficiently, and generalize to novel domains. This trend is well-established in natural language processing (*8, 9*) and computer vision (*10, 11, 12*), and is increasingly evident in general-purpose robotics (*13*). The evolution of Vision-Language-Action (VLA) models perfectly illustrates this trajectory. Foundational work like RT-1 (*14*) and RT-2 (*15*) demonstrated that transformer-based visuomotor policies exhibit clear scaling behavior, effectively transferring web-scale semantic knowledge to robotic control. Subsequent efforts proved the immense value of cross-embodiment data sharing. Open-X-Embodiment (*13*) pooled over one million trajectories, enabling the RT-2-X model to improve the success rate upon the single-embodiment RT-2 by approximately 50% on emergent tasks. This breakthrough led to a rapid progression of highly efficient, generalist architectures. Models such as Octo (*16*), CrossFormer (*17*), and the 7B-parameter OpenVLA (*18*) showed that policies pretrained on massive corpora can be efficiently fine-tuned to new embodiments, and outperform narrowly trained models. Frameworks like AgiBot World (*19*), Google's Gemini Robotics (*20*), and NVIDIA's GR00T (*21*) pushed these boundaries further, training on vast multimodal and synthetic datasets that span tens of thousands of hours without hitting performance saturation. Most recently, works such as Cosmos-Policy (*22*), DreamZero (*23*), and LingBot-VA (*24*) have shifted from semantic to spatiotemporal priors, replacing language-pretrained backbones with video foundation models that internalize physical dynamics from internet-scale video and show strong initial results on general manipulation benchmarks. The central lesson across this progression is that data diversity, multimodality richness, and scale compound. Each generation of larger, more varied corpora yields stronger generalization, higher absolute performance, and more efficient fine-tuning.

Surgical robotics, however, has not shared in these advances. The SutureBot benchmark (*25*) evaluated state-of-the-art general-purpose VLAs, including OpenVLA (*18*), GR00T-N1 (*21*), and π_0 (*26*), on autonomous suturing and found that all were significantly outperformed by a multitask Action Chunking Transformer (ACT) (*27*) policy trained from scratch on SutureBot data alone. This failure stems from a huge mismatch in both task complexity and physical environment. Unlike most actions in general robot datasets that require minimal instruction, surgical procedures demand years of intensive practice and specialized expertise

to execute safely. Furthermore, surgical instruments interact with deformable, viscoelastic tissues rather than rigid objects; visual feedback relies on complex endoscopic optics rather than standard fixed cameras; and procedures contain multi-step sequences with irreversible actions and narrow tolerance for error. Consequently, most capable general-purpose policies fail at entry-level surgical tasks, not because their architectures are inadequate, but because the gap between general manipulations and surgical procedures is too vast to bridge with out-of-domain pretraining.

The root cause of this domain gap is a structural data bottleneck. Unlike general robotics, acquiring synchronized kinematic and multimodal data in a clinical environment requires navigating stringent safety protocols, institutional review boards, and patient privacy regulations. Instrumenting proprietary surgical robots to record high-fidelity kinematic streams adds further technical complexity rarely encountered in standard robotic data collection. The resulting scarcity is evident in the scale of existing datasets. For over a decade, JIGSAWS (*28*), providing merely 3 hours of demonstrations on the da Vinci system, remained the primary surgical manipulation benchmark. Subsequent works have made meaningful progress. The Expanded CRCD (*29*) added kinematic recordings of pseudo-cholecystectomy, SutureBot (*25*) provided around 6 hours of end-to-end suturing data, and ImitateCholec (*30*) reached approximately 20 hours of segmented demonstrations of cholecystectomy with paired endoscopic video and dVRK kinematics, making it the largest publicly available single-robot surgical dataset. However, even these important contributions remain orders of magnitude smaller than the massive corpora driving general-purpose manipulation. They are also single-embodiment, single-institution, and narrow in task scope. Prior to the present work, no cross-embodiment surgical dataset has been assembled, leaving the field without the infrastructure that can determine if domain-focused pretraining can overcome the transfer deficit.

Open-H-Embodiment is a direct response to this bottleneck. Inspired by the collaborative success of Open-X-Embodiment (*13*), but rigorously adapted to the unique sensing, safety, and governance constraints of the medical domain, this community-driven initiative aggregates data from dozens of global institutions. By unifying a large corpus of real and synthetic data across a diverse range of robotic surgery and healthcare tasks, the dataset provides synchronized, multimodal observations under a standardized schema.

Pairing high-capacity VLA models with the Open-H dataset finally allows us to test whether the scaling laws of general robotics apply to the healthcare domain. To investigate this, we introduce GR00T-H, a healthcare-focused foundational VLA developed by post-training the GR00T-N1.6 model on the Open-H dataset. We rigorously evaluate GR00T-H against both generalist and specialist baselines on complex, long-horizon tasks across multiple robotic platforms. Our results show that GR00T-H has significant advantages in overall task success, data efficiency during fine-tuning, and cross-embodiment generalization. Beyond policy learning, we demonstrate that the Open-H enables fine-tuning of world foundation models for surgical robotics simulation. We produce Cosmos-H-Surgical-Simulator (C-H-S-S), the first multi-embodiment, kinematic action-conditioned world model for surgical simulation, built by fine-tuning Cosmos-Predict 2.5 (*31*) on the Open-H surgical data mixture spanning nine robotic platforms. C-H-S-S provides a publicly available, commercially usable frontier for in silico policy evaluation and synthetic data generation in the surgical domain.

In summary, the primary contributions of this work include:

- We introduce Open-H, the first large-scale, cross-embodiment, and multimodal dataset for surgical and healthcare robotics.
- We present GR00T-H, an open foundational surgical VLA demonstrating superior task success, fine-tuning data efficiency, and cross-embodiment generalization.
- We develop C-H-S-S, the first multi-embodiment, kinematic action-conditioned world model for surgical simulation, enabling in silico policy evaluation and synthetic data generation.

Results

The Open-H-Embodiment Dataset

Open-H-Embodiment is the first cross-embodiment, multi-institution dataset for healthcare robotics. The corpus comprises 119 datasets totaling 780 hours of paired video and kinematic data, contributed by more than 50 institutions worldwide. It spans 20 distinct robot

platforms, 33 task families, and 5 environment types ranging from digital simulation to live clinical procedures (Figure 2). Table 1 contextualizes this scale against prior surgical datasets, all of which are single-embodiment and single-institution.

Embodiment Diversity: The 20 robot platforms span five structural families: surgical robotic systems (618 hours across 76 datasets), industrial arms modified for healthcare use (45 hours, 12 datasets), flexible endoscope robots (30 hours, 8 datasets), simulated robots (86 hours, 21 datasets), and manual instrumentation (1 hour, 2 datasets). Commercial platforms include the Versius (CMR Surgical, Cambridge, UK), da Vinci Si and Xi (Intuitive Surgical, Sunnyvale, USA), BiTrack (Rob Surgical, Barcelona, Spain), MIRA (Virtual Incision, Lincoln, USA), and Maestro (Moon Surgical, Paris, France), alongside research platforms such as the dVRK, KUKA LBR iiwa, Franka Panda, UR5e, and the USTC Torin (Figure 2a). No prior surgical dataset spans more than a single platform family, making Open-H the first corpus capable of supporting cross-embodiment transfer experiments in this domain.

Task and Procedure Coverage: The dataset spans the full granularity spectrum relevant to surgical autonomy, from complete multi-hour clinical procedures down to isolated manipulation primitives. At the procedure level, four clinical workflows each exceed 100 hours: prostatectomy (169 hours), cholecystectomy (172 hours), hysterectomy (122 hours), and hernia repair (119 hours). At the subtask level, the corpus includes suturing and knot tying (57 hours across 30 datasets), tissue manipulation (19 hours, 18 datasets), and skills bench-

Table 1: **Open-H-Embodiment compared with prior surgical robotics datasets containing kinematic data.** Open-H is the first dataset to span multiple embodiments while exceeding prior work by more than an order of magnitude in duration.

	Hours	**Episodes**	**Platforms**
JIGSAWS (*28*)	∼3	103	1
Exp. CRCD (*29*)	∼7	80	1
SutureBot (*25*)	∼6	1,890	1
ImitateCholec (*30*)	∼20	∼18,000	1
Open-H (ours)	**780**	**125,815**	**20**

marks such as peg transfer and needle threading (25 hours, 14 datasets). Beyond surgery, 22 datasets (54 hours) cover robotic ultrasound scanning, tracked ultrasound, and ultrasound-guided intervention. 8 datasets (30 hours) address flexible endoscopy and colonoscopy navigation (Figure 2c). This coverage across complete procedures and isolated primitives provides training signal for high-level procedural planning and low-level motor control alike.

Sensing Modality Richness: Beyond the universal pairing of video and kinematics present in every dataset, individual contributions extend the observation space along multiple sensing axes. 72 datasets (61%) provide two or more simultaneous camera views, and 52 datasets (44%) include stereo endoscopic video, providing native depth cues absent from monocular feeds. Wrist-mounted cameras appear in 32 datasets (27%), offering close-range views of instrument-tissue interaction that complement the endoscopic field of view. Depth sensing is available in 28 datasets (24%), spanning RGB-D cameras (Intel RealSense, 11 datasets), computed stereo depth, and simulation-derived ground-truth depth maps. The 22 ultrasound datasets pair time-synchronized B-mode imaging with RGB cameras and, in several contributions from CUHK, ImFusion, and TUM, simultaneous wrist-camera and third-person views, creating observation spaces with three or more synchronized visual streams. One ultrasound contribution (TUM CAMP SonATA) additionally provides synchronized force kinematics. The simulation subset (84 hours across 18 datasets) provides dense output modalities infeasible to capture on real tissue, including surface normals, optical flow, and per-pixel tool segmentation masks. Five datasets from UBC include synchronized eye-gaze tracking at 120 Hz, capturing operator visual attention during endoscopic procedures.

Environment and Realism Spectrum: The corpus spans five realism tiers: digital simulation (84 hours, 18 datasets), benchtop and phantom (119 hours, 57 datasets), ex vivo (75 hours, 27 datasets), in vivo (3 hours, 6 datasets), and clinical human procedures (499 hours, 11 datasets). Clinical data, predominantly contributed by CMR Surgical (489 hours of Versius procedures), accounts for 64% of the total corpus by duration (Figure 2b). This distribution, spanning from fully controlled simulation through cadaveric tissue to live surgery, supports progressive model validation across increasing levels of clinical realism.

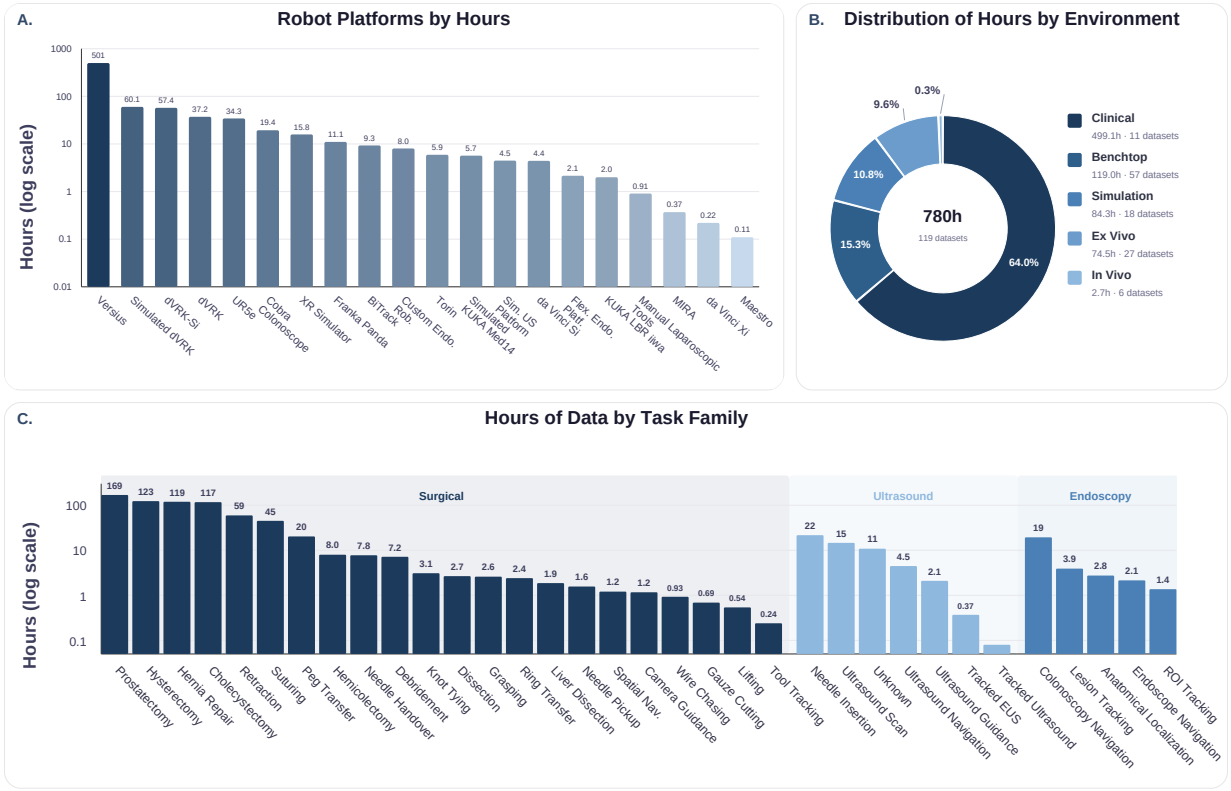


Figure 2: **Composition of the Open-H-Embodiment dataset.** (**A**) Dataset hours by robot platform. (**B**) Distribution of dataset hours by environment type. (**C**) Distribution of dataset hours across task families. Together, these panels summarize the current distribution of contributed data across embodiments, collection environments, and task families in Open-H-Embodiment.

GR00T-H: A Foundational Vision-Language-Action Model for Surgery

To demonstrate the utility of the Open-H dataset in enhancing robot learning performance for surgical tasks, we developed GR00T-H, the first open foundational surgical Vision-Language-Action (VLA) model. GR00T-H is based on NVIDIA's pretrained GR00T-N1.6 policy (*21*), which was then post-trained on the Open-H dataset. We evaluated GR00T-H by comparing its performance against the base GR00T-N1.6, the ACT policy, an architecture previously established as a leading baseline for surgical robotics (*25*), and LingBot-VA, a recent world action model that we post-train using a 50-hour dVRK-focused subset of Open-H (Supplementary Text). All policies were subsequently fine-tuned on small, embodiment- and task-specific datasets to assess their performance. For the end-to-end suturing experiment, we report cumulative task survival across sequential subtasks. For all other experiments, we report individual sub-task success rates and an averaged sub-task success rate with 95% Clopper-Pearson confidence intervals and p-values from Fisher's exact test with Holm-Bonferroni correction for multiple comparisons.

End-To-End Suturing: To evaluate long-horizon policy performance, we utilize the SutureBot benchmark (*25*), which employs a dVRK-based suturing dataset with a phantom silicone pad and suture needle as shown in Figure S1. To ensure a fair comparison, we implemented a per-setup evaluation protocol: for each specific environment configuration, robot configuration, pad position, and needle placement, GR00T-N1.6, GR00T-H, and ACT each performed a single suturing attempt. This setup was held constant across the three models and only changed once all models had completed the trial. LingBot-VA was evaluated in a separate evaluation session, with a best effort made to mirror the diversity of setup seen by the other three models. The suturing procedure is decomposed into five sub-tasks: needle-pickup, handover, throw, extraction, and knot-tying. Each model was initialized at the first sub-task, and performance was measured by the total progress achieved through the task sequence prior to failure.

As shown in Figure 3, all four models perform comparably through the early manipulation stages, with GR00T-H retaining all 20 trials through pickup and handover, while by the handover stage GR00T-N1.6 drops to 16, and LingBot-VA falls to 12. The performance

gap widens substantially at the throw stage, where GR00T-H retains 12/20 trials compared to 4/20 for both ACT and GR00T-N1.6, and LingBot-VA's 1/20. Only GR00T-H achieves full end-to-end task completion, finishing 5/20 trials (25%), while ACT, GR00T-N1.6, and LingBot-VA complete 0/20 trials. This experiment emphasizes the importance of policies being robust to compounding errors over the task horizon. A successful end-to-end suturing rollout is shown in Movie S1.

Generalization: To assess whether healthcare-specific post-training improves robustness to distribution shift, we evaluated GR00T-H, ACT, and GR00T-N1.6 on a wound configuration unseen during training and under different lighting conditions ($n = 10$ trials per subtask, $n = 30$ total) as shown in Figure S1. Due to relatively low end-to-end performance across policies for the in-distribution setup, we evaluate the policy generalization by looking at average sub-task performance instead of end-to-end performance for a stronger evaluation signal. ACT only achieved 15% on the pick up and handover task and failed all the other tasks. Averaging across individual sub-tasks, GR00T-H achieves a success rate of 54%, compared to 30% for GR00T-N1.6. Notably, GR00T-H leads on Pick Up and Handover (70% vs. 40%), and also Throw and Extract (42.5% vs. 5%). However, GR00T-N1.6 achieved similar performance to GR00T-H on the knot-tying task (50% vs. 45%). These results confirm that large pretrained VLAs improve robustness and visual generalization, while indicating that GR00T-H post-training may improve robustness in the surgical domain.

Data Efficiency: To test whether Open-H post-training reduces the amount of task-specific fine-tuning data required, we trained GR00T-H, ACT, and GR00T-N1.6 with either 33% or 100% of the full fine-tuning dataset ($\sim$2 hours vs. $\sim$6 hours of demonstrations), using proportional sampling to preserve task distribution and the ratio of nominal to recovery demonstrations. Similar to generalization experiments, we use sub-task success rates instead of end-to-end success rates for a stronger evaluation signal. Results are shown in Figure 4.

At 33% data, GR00T-H already matches ACT on the 3-task average (both $\approx$47%), while GR00T-N1.6 lags at $\approx$20%. With full data, GR00T-H improves substantially to $\approx$73%, compared to $\approx$50% for ACT and $\approx$37% for GR00T-N1.6. This suggests that Open-H post-training

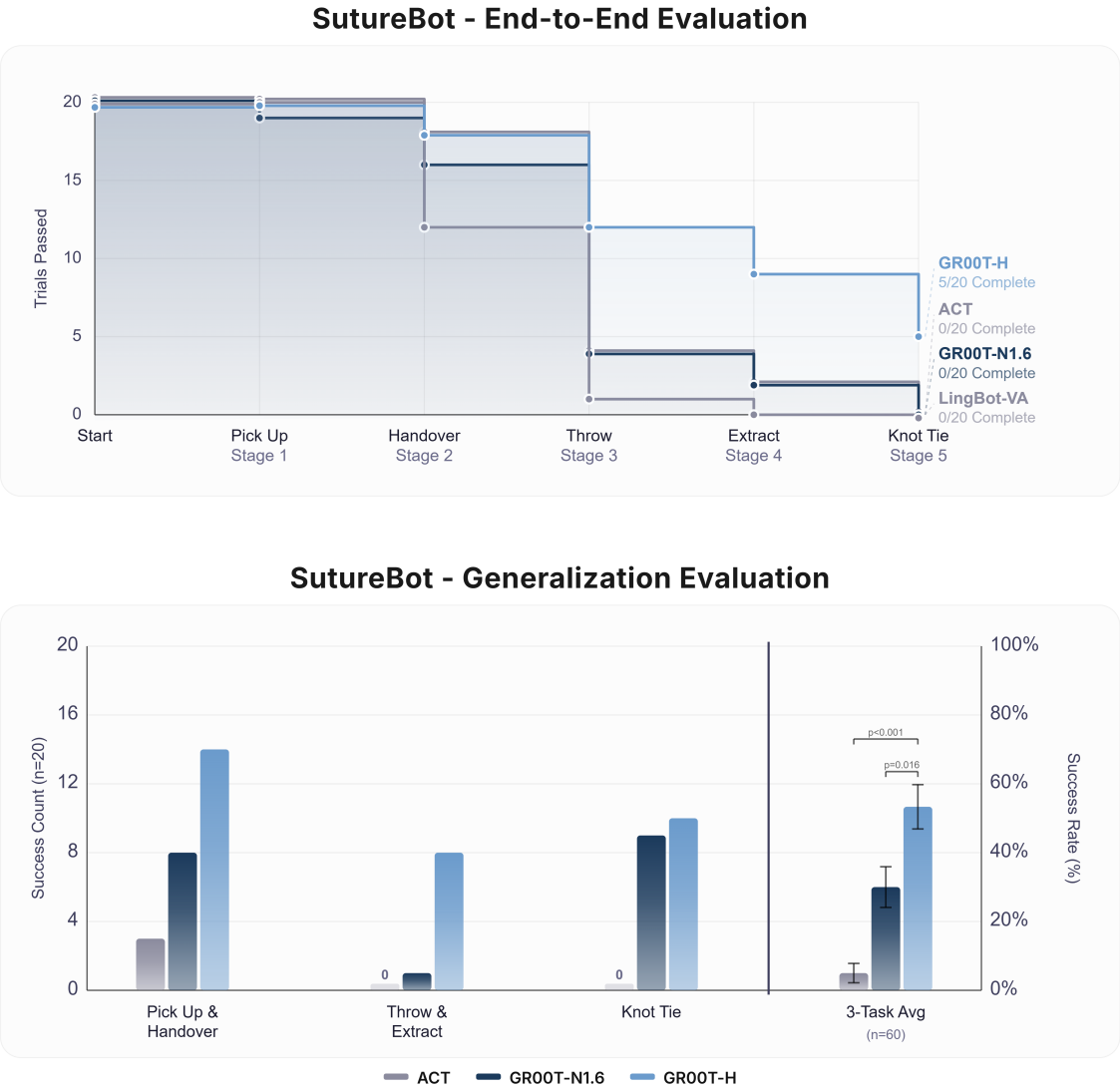


Figure 3: **SutureBot end-to-end and out-of-distribution evaluation.** Top: Task survival for GR00T-H on the SutureBot end-to-end suturing task compared to GR00T-N1.6, ACT, and LingBot-VA. GR00T-H improves end-to-end performance with a long-horizon success rate of 25%, showing better handling of compounding errors. Bottom: Out-of-distribution evaluation on SutureBot with an unseen wound configuration under varied lighting ($n = 20$ per subtask). GR00T-H achieves a 3-task average of 54%, outperforming GR00T-N1.6 (30%) and ACT (5%). Clopper-Pearson 95% confidence intervals are represented as error bars.

provides a stronger initialization for surgical fine-tuning: GR00T-H achieves competitive performance with limited data and may scale more effectively than GR00T-N1.6 when more data is available.

Multi-Embodiment: A primary advantage of training VLAs on diverse multi-embodiment data is the resulting performance gain across distinct platforms within the data distribution. To validate GR00T-H's cross-embodiment capabilities, we compared it against the GR00T-N1.6 base model across three systems: the CMR Versius, Virtual Incision MIRA, and the da Vinci Research Kit Si (dVRK-Si). The Versius system was evaluated on Peg Transfer, a standard surgical training task, using 5.2 hours of training data, with the procedure decomposed into block pickup from the peg, block handover, and placement on a peg. For the MIRA robot, we evaluated the needle pickup sub-task using only 22 minutes of data. Finally, the dVRK-Si was used for the SutureBot sub-tasks described previously, trained on 6 hours of data. Images of all system setups can be found in Figure S1. As shown in Figure 5, GR00T-H demonstrates a significant performance boost over the base model across all platforms ($p < 0.001$ for the overall average), confirming that Open-H post-training effectively enhances surgical capabilities across diverse robotic embodiments.

Ex Vivo Experiment: To evaluate GR00T-H in a clinically proximate setting, we conducted an ex vivo suturing evaluation on skin-on pork belly, assessing performance across the full 29-subtask sequence required to complete a suture covering needle manipulation, needle throwing, wound opening, suture manipulation, knot tying, and suture cutting. The ex vivo suturing setup can be seen in Figure S1. Each subtask was evaluated over 10 trials ($n = 290$ total), with final results shown in Figure 6 and video of successful subtask rollouts shown in Movie S2.

GR00T-H achieves an overall average success rate of 64% across all 29 subtasks. Performance is strongest on structured manipulation primitives: needle pickup (10/10), all handover stages (9/10, 10/10, 9/10), set-down needle (10/10), and all three knot-tying steps (10/10 each) all reach near-perfect or perfect success. Performance is lower on subtasks requiring fine instrument coordination or tissue contact, including readjust (4/10), open wound

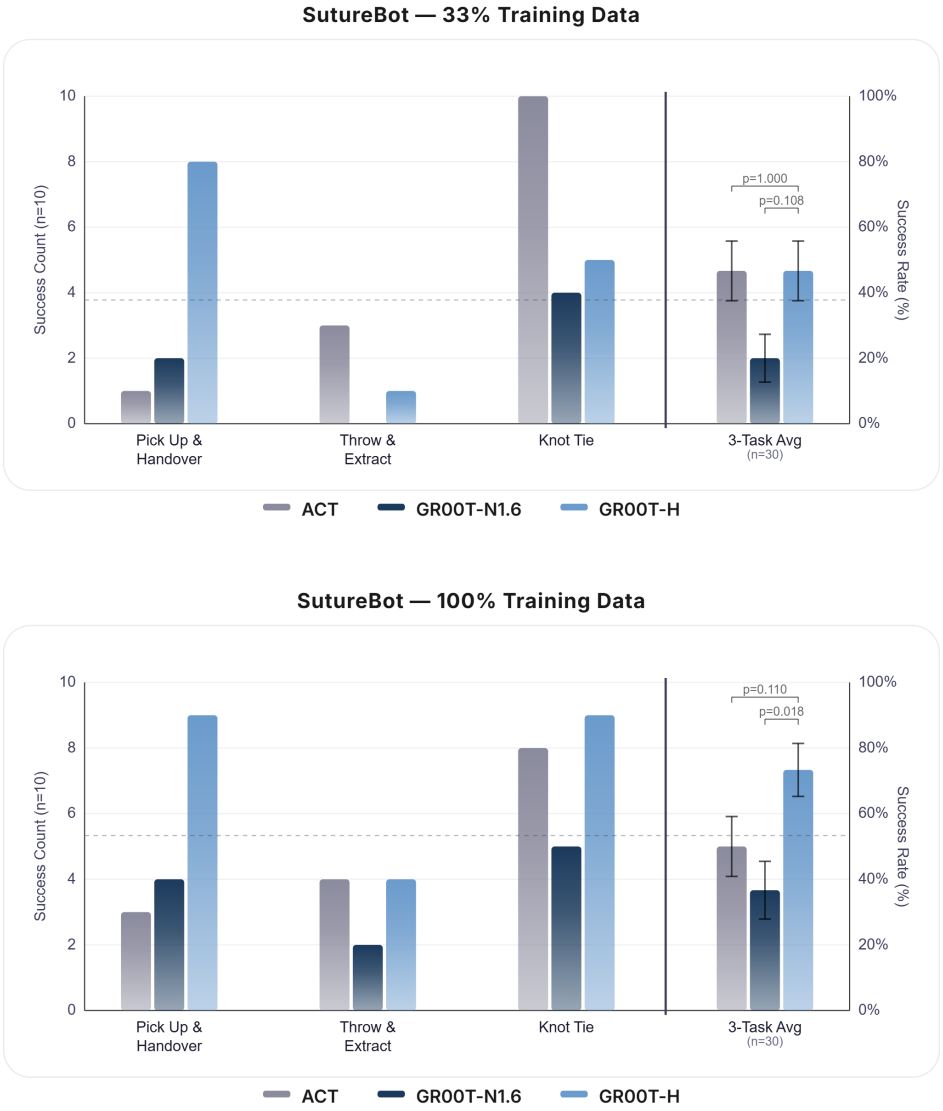


Figure 4: **SutureBot data efficiency evaluation.** Task success rate at 33% and 100% fine-tuning data on SutureBot ($n = 10$ per subtask). At 33% data, GR00T-H matches ACT while GR00T-N1.6 underperforms both. At 100% data, GR00T-H outperforms all baselines, indicating that Open-H post-training enables both data-efficient learning and stronger scaling. Clopper-Pearson 95% confidence intervals are represented as error bars.

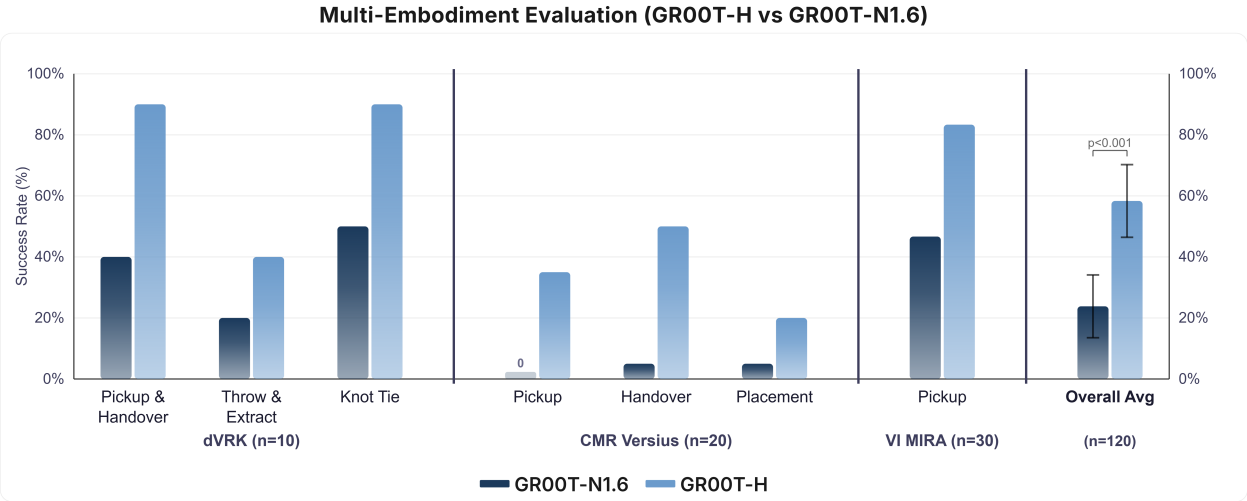


Figure 5: **Multi-Embodiment Performance Comparison.** Evaluation of the GR00T-H foundational VLA (post-trained on Open-H) versus the GR00T-N1.6 base policy across three surgical platforms: the da Vinci Research Kit Si (dVRK-Si), CMR Versius, and Virtual Incision MIRA. GR00T-H demonstrates significant performance gains across all robot embodiments and sub-tasks, with the overall average success rate showing a statistically significant improvement ($p < 0.001$). Error bars represent the Clopper-Pearson 95% confidence intervals across trials.

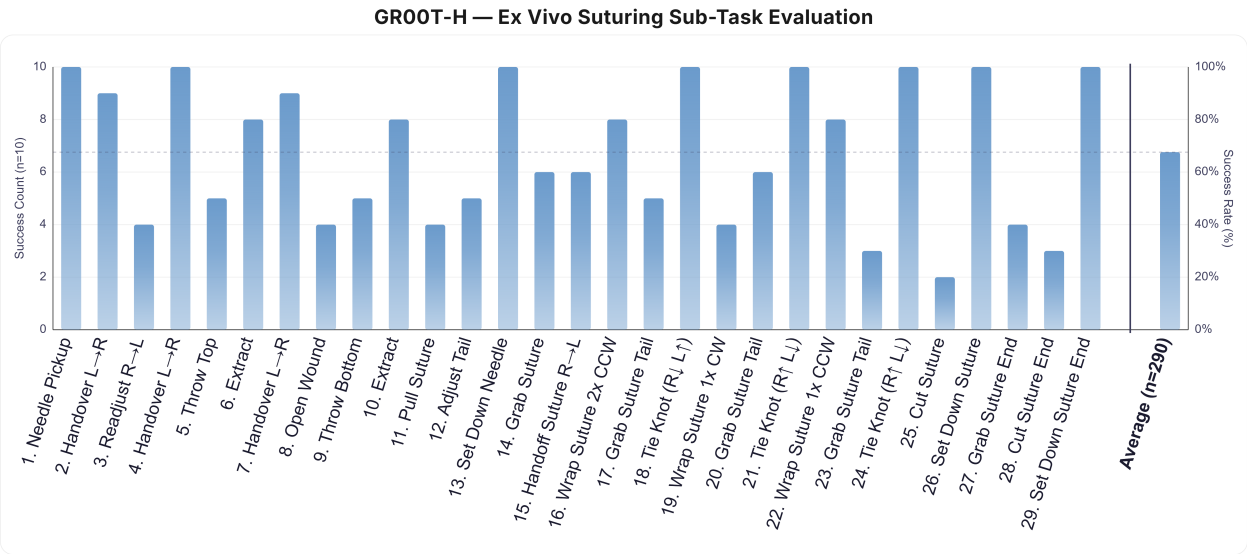


Figure 6: **Ex vivo suturing evaluation across 29 subtasks.** ($n = 10$ per subtask, $n = 290$ total). Tasks span needle manipulation, wound opening, suture passing, knot tying, and suture cutting. GR00T-H achieves an average success rate of ≈64%, with near-perfect performance on structured manipulation primitives and lower success on fine-contact and cutting steps. The rightmost bar shows the overall average across all subtasks. Clopper-Pearson 95% confidence intervals are represented as error bars.

(4/10), wrapping steps (8/10, 4/10, and 8/10), grab the suture tail (5/10, 6/10, and 3/10), and the two cut suture steps (2/10 and 3/10). The cutting steps represent the most consistent failure mode, potentially due to the precise and quick nature of the cutting motion.

These results demonstrate that GR00T-H can execute the sub-tasks of a clinically relevant manipulation sequence on real biological tissue, with reliable performance on the majority of sub-tasks. The pattern of failures, concentrated in fine-contact and cutting steps rather than distributed uniformly, suggests that targeted data collection and fine-tuning on these specific failure modes represent a tractable path to achieve end-to-end performance.

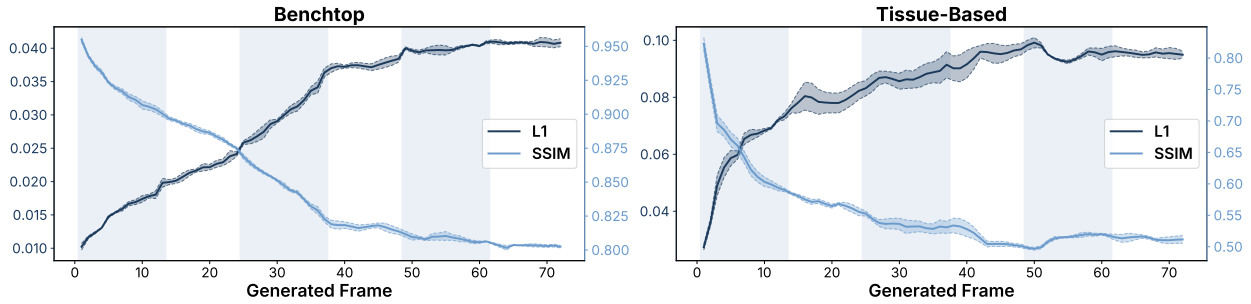


Figure 7: **Quantitative evaluation of Cosmos-H-Surgical-Simulator.** Per-frame L1 and SSIM for benchtop vs. tissue-based datasets. Mean L1 error and SSIM as a function of generated frame index across 72 autoregressively generated frames (6 chunks $\times$ 12 frames each). Mean over 3 generation seeds, each averaged across all evaluated episodes within the category; shaded bands indicate 1 standard deviation across seeds. Left: benchtop datasets (phantom and bench procedures). Right: tissue-based datasets (clinical, cadaver, and ex vivo tissue).

Cosmos-H-Surgical-Simulator: A World Model for Surgical Robotics Simulation

Cosmos-H-Surgical-Simulator is the first multi-embodiment, kinematic action-conditioned world foundation model for surgical robotics simulation. Built by fine-tuning Cosmos-Predict 2.5 (*31*), a 2B-parameter latent video diffusion transformer, on the Open-H surgical data mixture spanning nine robotic platforms and 32 datasets (Table S4), C-H-S-S accepts a single video frame and a sequence of kinematic action vectors, and autoregressively generates future frames that predict the visual outcome of those actions. Through iterative rollout, it can produce complete surgical trajectory videos for any embodiment in its training distribution, enabling in silico policy evaluation and synthetic data generation from a single publicly available, commercially usable checkpoint. Movie S3 shows representative generated rollouts across multiple embodiments. To our knowledge, no prior surgical world model has supported more than a single robotic platform.

Quantitative evaluation of surgical world models remains an open problem. Standard unconditional video generation metrics such as FID and FVD measure distributional similarity to a reference set but do not assess whether a generated video faithfully tracks the spe-

cific visual consequences of a given action sequence. Because C-H-S-S is action-conditioned, replaying the recorded action sequence from a held-out test episode produces a generated rollout that can be compared frame-by-frame against the original ground-truth video. We adopt two complementary pixel-level metrics: L1 (mean absolute error), which measures overall pixel fidelity, and SSIM (structural similarity index), which captures perceptual structural correspondence. Both are computed per frame across 72 autoregressively generated frames (6 chunks of 12 frames each) on held-out test episodes from 25 Open-H datasets, using 3 generation seeds per episode to quantify variability. Results are stratified by dataset category; benchtop (datasets covering phantom and bench procedures) and tissue-based (datasets covering clinical, cadaver, and ex vivo tissue), to characterize how scene complexity affects generation quality.

Figure 7 reports the per-frame L1 and SSIM trajectories for both categories. Benchtop scenes, which feature controlled lighting, static backgrounds, and clearly visible instruments, maintain low L1 and high SSIM throughout the 72-frame horizon, with only gradual degradation. Tissue-based scenes exhibit higher L1 and lower SSIM overall, consistent with the greater visual complexity of in-body environments: variable lighting from the endoscope, frequent instrument occlusion, specular reflections on wet tissue, and deformable anatomy that is inherently harder to predict. The error bands across seeds are narrow for benchtop scenes, indicating stable generation; tissue-based scenes show wider seed-to-seed variability, particularly in L1, reflecting the greater stochasticity involved in predicting visually complex in-body environments. The expected sawtooth pattern at chunk boundaries, where each new 12-frame chunk re-conditions on the last generated frame, is visible but modest, suggesting that the model maintains coherent scene state across autoregressive transitions.

Discussion

Our results show that training on a large, diverse surgical dataset improves both task performance and robustness of foundation VLA policies. GR00T-H was the only model to achieve full end-to-end task completion on SutureBot, and it consistently outperformed ACT and the base GR00T-N1.6 model under out-of-distribution conditions and with limited fine-tuning

data. These findings suggest that the scale and diversity of Open-H provide a surgical prior that transfers across tasks, embodiments, and environments. The data bottleneck constraining surgical robot learning is not fully resolved, but it can be meaningfully reduced through large-scale data collection paired with healthcare-specific foundation model post-training.

Our evaluation also revealed an interesting finding about policy robustness to hardware variation. Over the course of the evaluation period, the experimental setup underwent changes typical of long-running robot deployments: surgical instruments accumulated mechanical wear that shifted their cable-driven kinematics, and wrist cameras required replacement with units of slightly different optical properties. ACT was sensitive to these changes and performed below its originally reported results (*25*). GR00T-H, on the other hand, still achieved 25% end-to-end task completion under the same conditions. This suggests that post-training on the diverse Open-H dataset makes the policy more resilient to the kind of gradual hardware drift that is common in real surgical environments, where instrument wear and camera replacements are routine.

Post-training on Open-H also appears to improve data efficiency. GR00T-H matches ACT with only 33% of the full fine-tuning dataset and outperforms the GR00T-N1.6 baseline at 100%. This suggests that post-training on Open-H provides a stronger starting point for learning new surgical tasks, which is useful for institutions that want to adapt the model to new platforms or procedures without collecting large amounts of local data. The multi-embodiment results further support this interpretation: GR00T-H shows consistent gains across different robotic platforms, suggesting the learned representations are not specific to a single robot's kinematics.

Open-H also enables a new class of world foundation model for surgical simulation. Cosmos-H-Surgical-Simulator is, to our knowledge, the first kinematic action-conditioned world model that supports multi-embodiment surgical video generation from a single checkpoint, spanning nine robotic platforms. Prior surgical world models (*32*) have been limited to a single platform; C-H-S-S extends this to any embodiment in its training distribution without platform-specific retraining. The per-frame L1 and SSIM evaluation confirms that the model maintains visual fidelity over 72-frame autoregressive rollouts, with the performance gap between benchtop and tissue-based scenes tracking the inherent difficulty of the under-

lying environments. Prior work on the dVRK platform (*32*) has further demonstrated that action-conditioned world models of this kind can serve as simulation environments for closed-loop policy evaluation. Open-H can therefore support not just policy training but also world models, visual encoders for instrument and tissue understanding, and language-conditioned planning systems, making it useful as shared infrastructure for the broader surgical robotics community.

Several limitations of the current work must be addressed before these methods could be applied in animal or human surgical settings. In ex vivo evaluation, GR00T-H performs well on structured steps like needle pickup, handover, and knot tying, reaching near-perfect success on many of these. However, performance drops on subtasks requiring fine instrument coordination under tissue contact, with cutting steps reaching only 20–30% success. The overall average of 64% across the full 29-step ex vivo suturing sequence, and 25% end-to-end completion on the SutureBot benchmark, are encouraging but well below what would be needed for reliable autonomous execution. All evaluations are conducted on tissue phantoms and benchtop ex vivo tissue in controlled lab settings. How the policy performs on live tissue, which varies in stiffness, bleeds, and moves, is not yet known. The policy also has no way to detect or respond to unexpected events such as tissue tearing, instrument failures, or patient movement, all of which are important safety considerations for any real deployment. Errors also tend to compound across long task sequences, which limits end-to-end completion rates even when individual tasks are performed reliably. More data from a wider range of institutions, procedures, and tissue types will be needed for the model to generalize to the variety of cases seen in clinical practice. While initial procedural annotations covering approximately 205 hours of Versius-500 clinical data are released alongside the dataset, richer labels including instrument state, tissue contact events, and failure modes will be needed to address the fine-contact steps where the model currently struggles the most. Pre-clinical evaluation in animal models, with appropriate regulatory oversight, is a necessary next step before any component of this pipeline could be considered for use in humans. LingBot-VA, which leverages a world foundation model backbone as opposed to a VLM backbone, represents a promising but distinct architectural class for robot learning. Its comparatively rapid performance degradation across subsequent end-to-end suturing subtasks does not necessarily

reflect the potential of the architecture, as several factors remain unexplored. The model was post-trained on a 50-hour dVRK-focused subset of Open-H, roughly an order of magnitude less data than GR00T-H's 601-hour surgical mixture. World action models also introduce extensive training and inference parameters, including action and video denoising schedules, KV cache update frequency, video frame rate, and context history length, all of which require further exploration to properly adapt these models to surgical tasks. Determining the appropriate data recipe and inference configuration for world action models on dexterous surgical data is an important direction for future work. Regarding C-H-S-S, the L1 and SSIM evaluation relies on open-loop replay of recorded action sequences. In closed-loop deployment, where actions come from a policy or a surgeon's console reacting to the model's own generated frames, no such ground truth exists, and pixel-level metrics are no longer applicable. More broadly, L1 and SSIM are scene-level measures that do not capture surgery-specific aspects of generation quality, such as whether instruments are rendered in the correct position or whether tool-tissue interactions are physically plausible. Developing domain-specific metrics, such as spatial consistency of generated instruments against ground-truth tool trajectories or fidelity of tool-tip localization across frames, is an important open problem. Early experiments with segmentation-based metrics such as generated-vs-actual tool consistency and tool centroid distance showed promise on individual platforms but proved insufficiently robust across the diversity of embodiments and visual conditions. Additionally, the Open-H dataset predominantly comprises successful task demonstrations. Prior work has shown that failure episodes are a critical factor for surgical world model training (*32*). Future data collection efforts should explicitly include such failure trajectories to improve the realism and coverage of surgical world models. Extending the automated closed-loop evaluation pipeline demonstrated for a single platform (*32*) to the multi-embodiment setting of C-H-S-S is an important next step.

Materials and Methods

The Open-H Dataset

While large-scale, multi-embodiment datasets have driven rapid progress in general-purpose robotics, assembling a comparable corpus for healthcare robotics introduces challenges that are largely absent from that setting.

In general-purpose robotics, foundational policy research is predominantly conducted on purpose-built platforms such as the Franka Panda (*18*), Unitree G1 (*21*), and ALOHA (*26*). These platforms provide integrated hardware and software that support teleoperation and data collection with minimal additional engineering. In healthcare robotics, by contrast, research is largely conducted on devices engineered for clinical use rather than for robotics research. Most contributing institutions have therefore developed their own suites of proprietary hardware, software, and interfaces to repurpose these medical devices for modern robotics workflows.

Creating a multi-institution, multi-embodiment healthcare-robotics dataset that is *usable* by downstream researchers is constrained by two principal challenges:

1. **Inconsistent data formatting.** Training downstream models on dozens of independently formatted datasets requires detailed knowledge of each collection's schema to convert the complete corpus into a unified representation.

2. **Opaque platform nuances.** The diversity of hardware and task families in healthcare robotics introduces considerations that are not self-evident from the data alone, such as surgical clutching mechanisms that invalidate portions of recorded kinematics, multi-arm systems in which only a subset of visible instruments are actively controlled, and variation in operator skill level (e.g., expert surgeon versus machine-learning researcher).

To address these concerns, Open-H adopts two complementary strategies: all constituent datasets are converted into a unified LeRobot v2.1 format, and each dataset is accompanied by a consistently structured README that discloses platform-specific nuances to downstream users.

Data Formatting: We adopt the LeRobot v2.1 format as the standard representation for all Open-H data on the basis of its storage efficiency and broad community adoption (*33*). The format stores low-dimensional kinematics in columnar Parquet files and visual observations as hardware-accelerated, lossy MP4 video, reducing the overall storage footprint by up to 98% compared to formats such as RLDS or HDF5 (*34*). Healthcare-robotics tasks, however, frequently require contextual metadata that falls outside the standard LeRobot schema for video, kinematics, and task prompts, including modality-specific parameters (e.g., ultrasound acquisition settings), task-level annotations (e.g., surgical phase and active instrument identity), and other domain-relevant fields. To accommodate these requirements without breaking compatibility with the core LeRobot library, we developed the Open-H data-collection repository (*35*), which provides standardized guidelines and examples for extending the LeRobot schema with healthcare-specific fields.

Documentation: Even with a unified storage format, the practical utility of an aggregate dataset at this scale depends on whether downstream users can understand the collection-specific details of each constituent contribution. To this end, every Open-H dataset includes a structured README that follows a common template. The template captures the robot type(s) used, the data-collection method (e.g., teleoperation or autonomous execution), the operator skill level (e.g., expert surgeon or machine-learning researcher), the data synchronization strategy employed, the dimensions of data diversity (e.g., camera positioning, environment, and lighting variation), the kinematic representation (e.g., absolute Cartesian or relative joint space), and additional domain-relevant fields. The complete template with all recorded fields is available in (*35*).

Clinical Procedure Annotation: The clinical procedures in the Versius-500 contribution present a distinct challenge for downstream learning. Unlike tabletop manipulation tasks, where the correct next action is largely determined by a single observation frame, clinical surgical procedures are temporally extended and non-Markovian: from a given endoscopic image, the next action is not immediately clear, as it depends on temporal context, clinical and patient history, as well as surgeon preference. To support future research that leverages

this structure, we release initial annotations alongside the dataset. For Versius-500-inguinal-hernia, 13 episodes covering approximately 32 minutes of procedure video were manually annotated at the pixel and frame level (5 FPS) with instrument and anatomy segmentation masks and action classifications. For Versius-500-cholecystectomy, we manually annotated 500 minutes of phase-level and gesture-level labels, and for Versius-500-hysterectomy, 1000 minutes phase-level labels. We fine-tuned LemonFM (*36*), a surgical video foundation model pretrained on 938 hours of endoscopic video, on this manually annotated data, and generated labels for the full 80 hours of Versius-500-cholecystectomy and 124 hours of Versius-500-hysterectomy. Evaluated against a held-out test subset of Versius-500 data, the model achieves phase accuracy of approximately 0.65 and gesture accuracy of 0.56 on cholecystectomy, and phase accuracy of 0.75 on hysterectomy.

GR00T-H

The Open-H dataset supports a range of downstream applications, from monocular depth estimation models to world models, though its most immediate use is as a pre-training corpus for healthcare-focused manipulation policies. To test whether domain-specific data at this scale can close the transfer gap documented above, we train GR00T-H, a surgical foundation policy built on the GR00T-N1.6-3B vision-language-action model (VLA).

We build GR00T-H on GR00T-N1.6-3B (*21*) due to the strong foundation it establishes for training an imitation learning policy on the multimodal and multi-embodiment Open-H corpus. The model's extensive pre-training on diverse real and synthetic manipulation data provides a meaningful initialization for the dexterous, long-horizon tasks common in surgical robotics. Its Cosmos-2B VLM backbone encodes images at flexible resolution, accommodating the heterogeneous resolutions and aspect ratios across healthcare-specific robotic embodiments. Furthermore, its embodiment-specific action heads allow each robot configuration to learn its own output projection, an important feature when training across diverse robot embodiments with varying kinematic output distributions.

GR00T-H is adapted from the upstream foundation model using an Open-H post-training phase. For this post-training, we select a 601-hour surgical subset of the full 780-hour Open-H corpus, isolating the largest and most coherent subgroup: real-world surgical tasks. Thus,

only real-world surgical datasets were used in the training of GR00T-H; we leave training and evaluating policies that generalize across ultrasound, endoscopy, and simulation data for future work. For the sampled real-world surgical subset, the Versius-500 contribution dominates by volume. Therefore, we cap its sampling frequency at 20% of training steps to prevent any single embodiment, environment, or task family from dominating the loss signal. The remaining datasets are sampled proportionally to their size in the corpus. The full training mixture is provided in Table S3.

To establish a common kinematic learning space across embodiments, we normalize all datasets to use relative end-effector (EEF) control with 6D rotation matrices (*37*) for the orientation component. Relative actions remove the requirement for the model to learn each unique embodiment's forward kinematics, a necessity under joint-space control, and improve robustness to variation in workspace position and scale, following the relative action formulation adopted in prior surgical policy work (*6*).

To accommodate the heterogeneous kinematics across the Open-H corpus, we leverage GR00T-N1.6's embodiment-specific action heads, where each unique robot configuration learns its own MLP projection from the DiT action expert outputs to the normalized action space. We assign a distinct action head to each robot configuration rather than sharing heads across datasets collected on the same platform. This accommodates a practical constraint of cable-driven surgical robots: cable stretch, hysteresis, and wear characteristics vary across individual machines, and different instrument types introduce further variation in tool-tip kinematic mapping. A shared projection would conflate these instance-specific differences, degrading action prediction accuracy.

To ensure well-normalized kinematic values across robot configurations, we collect per-dataset normalization statistics and merge them into per-configuration statistics at train time via weighted moment matching, where the weights correspond to training-mixture sampling ratios. Statistics are computed per action dimension and per temporal step within the action chunk to preserve resolution for near-future actions (*38*). We apply z-score normalization with $[-5, 5]$ clipping to faithfully model each configuration's kinematic distribution while excluding extreme outliers.

The design decisions described above, including the data mixture, action normalization

scheme, embodiment-specific action heads, and kinematic normalization strategy, were iteratively refined using a combination of real-world evaluation and Cosmos-Surg-dVRK (*32*), an automated evaluation pipeline. Validation loss on held-out demonstrations is a poor proxy for closed-loop task success in imitation learning, making standard early-stopping criteria unreliable for checkpoint selection. Cosmos-Surg-dVRK addresses this by using an action-conditioned world model (architecturally identical to C-H-S-S, trained on monocular SutureBot data) to autoregressively render full task-length rollouts from the policy under evaluation, after which a finetuned V-JEPA 2 (*39*) labels the generated video with per-subtask success rates. Prior work established a strong Pearson correlation between this pipeline's automated scores and real-world success rates on the dVRK (*32*). During GR00T-H development, Cosmos-Surg-dVRK served as a practical screening tool to narrow the space of checkpoint iterations and training hyperparameters before committing physical robot time.

Using this iteratively validated recipe, the final Open-H post-training run takes 65,000 steps of full-weight training with a global batch size of 1,024 on the 601-hour surgical subset. The complete post-training configuration, including optimizer, learning rate schedule, and augmentation parameters, is provided in Table S2. For downstream use, including all evaluations in this work, we recommend an additional fine-tuning stage in which the VLM backbone is frozen and only the action prediction components are fine-tuned on the target task, following the protocol established in GR00T-N1.6 (*21*). This preserves the healthcare-domain representations acquired during post-training while efficiently specializing the action expert DiT to the target task distribution.

Cosmos-H-Surgical-Simulator

Beyond policy learning, the Open-H dataset enables fine-tuning of world foundation models for surgical robotics simulation. Kinematic action-conditioned video generation models that predict future visual observations given a current frame and a sequence of kinematic actions are a promising frontier for in silico surgical policy evaluation and synthetic data generation (SDG). By learning the visual dynamics of instrument-tissue interaction, such models can provide a low-cost evaluation environment for surgical policies before physical deployment, as previously demonstrated for the dVRK platform (*32*) and in the development of

GR00T-H. We extend this paradigm to the multi-embodiment setting by fine-tuning Cosmos-Predict 2.5 (*31*), a 2B-parameter latent video diffusion transformer, on the surgical subset of Open-H, producing the Cosmos-H-Surgical-Simulator checkpoint. Cosmos-Predict 2.5 is a diffusion transformer (DiT) that generates video in the latent space of a pre-trained variational autoencoder. An MLP conditions the denoiser on per-timestep kinematic actions, enabling action-conditioned next-frame prediction. Given a single context frame and a sequence of 12 action vectors, the model generates 12 subsequent frames. Through autoregressive rollout, the model produces videos covering complete surgical trajectories. We adopt the publicly released Cosmos-Predict 2.5-2B-Video2World weights as initialization and fine-tune on Open-H surgical data with a unified 44-dimensional action space that accommodates all embodiments via zero-padding.

The C-H-S-S training mixture comprises 32 datasets from 9 robot embodiments and 10+ institutions. CMR Surgical Versius-500 accounts for 50% of the training compute across four clinical procedures (cholecystectomy, prostatectomy, inguinal hernia, hysterectomy), while the remaining 50% is distributed proportionally by frame count across dVRK variants (JHU, Stanford, Hamlyn, UCSD, UCB), Turin MITIC, USTC Torin, and Moon Surgical platforms. All action representations are converted to a hybrid-relative format (relative translation + 6D rotation) and z-score normalized per-embodiment. The full dataset specification and mixture ratios are detailed in Table S4.

Fine-tuning uses fused AdamW with a learning rate of 1.6×10^{-4}, linear decay schedule with warm-up, and a global batch size of 1,024, running for 42,000 steps on 64 A100 80 GB GPUs. Video resolution is 512×288 (16:9). The model processes 13 frames per sample (1 context + 12 prediction) with 12 corresponding action timesteps. Each embodiment's native kinematic frequency is downsampled to a consistent $\sim$10 fps effective rate via embodiment-specific timestep strides.

To quantitatively assess generation fidelity, we replay recorded action sequences from held-out test episodes through C-H-S-S and compare the generated video frame-by-frame against the ground-truth recording. Each episode is rolled out for 6 autoregressive chunks (72 generated frames). We report two pixel-level metrics: L1 (mean absolute error in $[0, 1]$ pixel space), which measures overall pixel fidelity, and SSIM (structural similarity index),

which captures perceptual structural correspondence. Evaluation covers 25 of the 32 training datasets (7 are excluded because their test-split episodes are shorter than the 72-frame evaluation horizon). For each dataset, 2 episodes are drawn from the 5% held-out test split using a fixed selection seed, and each episode is generated with 3 independent random seeds, yielding up to 150 episode evaluations in total. To isolate generation variability from cross-dataset differences, we aggregate by first computing the per-frame mean across all episodes within each seed, then reporting the mean and standard deviation across the 3 seed-level means. Results are stratified into benchtop (18 datasets: bench procedures) and tissue-based (7 datasets: clinical, cadaver, and ex vivo tissue) categories.

References

1. G. Plc, *The Complexities of Physician Supply and Demand: Projections From 2021 to 2036*, Tech. rep., AAMC, Washington, DC (2024).

2. S. Moffatt-Bruce, J. Crestanello, D. P. Way, T. E. Williams Jr, Providing cardiothoracic services in 2035: signs of trouble ahead. *The Journal of thoracic and cardiovascular surgery* **155** (2), 824–829 (2018).

3. T. Haidegger, Autonomy for surgical robots: Concepts and paradigms. *IEEE Transactions on Medical Robotics and Bionics* **1** (2), 65–76 (2019).

4. S. Schmidgall, J. D. Opfermann, J. W. Kim, A. Krieger, Will your next surgeon be a robot? Autonomy and AI in robotic surgery. *Science robotics* **10** (104), eadt0187 (2025).

5. P. Kazanzides, *et al.*, An open-source research kit for the da Vinci® Surgical System, in *2014 IEEE international conference on robotics and automation (ICRA)* (IEEE) (2014), pp. 6434–6439.

6. J. W. Kim, *et al.*, Surgical Robot Transformer (SRT): Imitation Learning for Surgical Tasks, in *Proceedings of the 8th Conference on Robot Learning (CoRL 2024)* (2024).

7. Y. Long, *et al.*, Surgical embodied intelligence for generalized task autonomy in laparoscopic robot-assisted surgery. *Science Robotics* **10** (104), eadt3093 (2025).

8. T. Brown, *et al.*, Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems* **33**, 1877–1901 (2020).

9. J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics* (2019), pp. 4171–4186.

10. A. Dosovitskiy, *et al.*, An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, in *International Conference on Learning Representations* (2021).

11. A. Radford, *et al.*, Learning Transferable Visual Models From Natural Language Supervision, in *Proceedings of the 38th International Conference on Machine Learning* (PMLR) (2021), pp. 8748–8763.

12. K. He, *et al.*, Masked Autoencoders Are Scalable Vision Learners, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 16000–16009.

13. Open X-Embodiment Collaboration, Open X-Embodiment: Robotic Learning Datasets and RT-X Models, in *IEEE International Conference on Robotics and Automation (ICRA)* (2024).

14. A. Brohan, *et al.*, RT-1: Robotics Transformer for Real-World Control at Scale, in *Robotics: Science and Systems (RSS)* (2023).

15. B. Zitkovich, *et al.*, Rt-2: Vision-language-action models transfer web knowledge to robotic control, in *Conference on Robot Learning* (PMLR) (2023), pp. 2165–2183.

16. Octo Model Team, *et al.*, Octo: An Open-Source Generalist Robot Policy. *arXiv preprint arXiv:2405.12213* (2024).

17. R. Doshi, H. Walke, O. Mees, S. Dasari, S. Levine, Scaling Cross-Embodied Learning: One Policy for Manipulation, Navigation, Locomotion and Aviation. *arXiv preprint arXiv:2408.11812* (2024).

18. M. J. Kim, *et al.*, Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246* (2024).

19. AgiBot-World-Contributors, *et al.*, AgiBot World Colosseo: A Large-Scale Manipulation Platform for Scalable and Intelligent Embodied Systems. *arXiv preprint arXiv:2503.06669* (2025).

20. G. R. Team, *et al.*, Gemini robotics 1.5: Pushing the frontier of generalist robots with advanced embodied reasoning, thinking, and motion transfer. *arXiv preprint arXiv:2510.03342* (2025).

21. J. Bjorck, *et al.*, Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734* (2025).

22. M. J. Kim, *et al.*, Cosmos Policy: Fine-Tuning Video Models for Visuomotor Control and Planning, in *The Fourteenth International Conference on Learning Representations* (2026), `https://openreview.net/forum?id=wPEIStHxYH`.

23. S. Ye, *et al.*, World Action Models are Zero-shot Policies (2026), `https://arxiv.org/abs/2602.15922`.

24. L. Li, *et al.*, Causal World Modeling for Robot Control (2026), `https://arxiv.org/abs/2601.21998`.

25. J. Haworth, *et al.*, SutureBot: A Precision Framework & Benchmark for Autonomous End-to-End Suturing, in *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track* (2025).

26. K. Black, *et al.*, π_0: A Vision-Language-Action Flow Model for General Robot Control, in *Robotics: Science and Systems (RSS)* (2025).

27. T. Z. Zhao, V. Kumar, S. Levine, C. Finn, Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705* (2023).

28. Y. Gao, *et al.*, JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS): A Surgical Activity Dataset for Human Motion Modeling, in *Modeling and Monitoring of Computer Assisted Interventions (M2CAI) – MICCAI Workshop* (2014).

29. K.-H. Oh, *et al.*, Expanded Comprehensive Robotic Cholecystectomy Dataset (CRCD). *Journal of Medical Robotics Research* (2025).

30. P. Hansen, *et al.*, ImitateCholec: A Multimodal Dataset for Long-Horizon Imitation Learning in Robotic Cholecystectomy. *Scientific Data* **13** (1), 210 (2026).

31. A. Ali, *et al.*, World simulation with video foundation models for physical ai. *arXiv preprint arXiv:2511.00062* (2025).

32. L. Zbinden, *et al.*, Cosmos-Surg-DVRK: World Foundation Model-Based Automated Online Evaluation of Surgical Robot Policy Learning. *IEEE Robotics and Automation Letters* **11** (5), 5978–5985 (2026), doi:10.1109/LRA.2026.3675962.

33. R. Cadene, *et al.*, LeRobot: An Open-Source Library for End-to-End Robot Learning, in *The Fourteenth International Conference on Learning Representations (ICLR)* (2026).

34. K. Chen, *et al.*, Robo-DM: Data Management For Large Robot Datasets, in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)* (2025).

35. Open-H Initiative, Open-H-Embodiment: Data Contribution How-To Guide and Scripts, `https://github.com/open-h/open-h-embodiment` (2026), accessed: 2026-04-02.

36. C. Che, C. Wang, T. Vercauteren, S. Tsoka, L. C. Garcia-Peraza-Herrera, LEMON: A Large Endoscopic MONocular Dataset and Foundation Model for Perception in Surgical Settings, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2026), arXiv:2503.19740.

37. Y. Zhou, C. Barnes, J. Lu, J. Yang, H. Li, On the Continuity of Rotation Representations in Neural Networks, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 5745–5753.

38. T. L. Team, *et al.*, A Careful Examination of Large Behavior Models for Multitask Dexterous Manipulation. *arXiv preprint arXiv:2507.05331* (2025).

39. M. Assran, *et al.*, V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning. *arXiv preprint arXiv:2506.09985* (2025).

40. L. Li, *et al.*, Causal World Modeling for Robot Control. *arXiv preprint arXiv:2601.21998* (2026).

Acknowledgments

We would like to thank the many institutions, companies, researchers, and students who participated in this effort; without their contributions, this dataset would not have been

possible.

Funding: This material is based on work supported by the following funding sources: NIH R56EB033807 (A.K.), ARPA-H award 75N91023C00048 (J.C., Xin.C., A.K.), ARPA-H award AY1AX000023 (J.G., E.K., A.K.), ARPA-H award D24AC00415-00 (N.Y., A.K.), NSF CAREER Award 2144348 (A.K.), and NSF Graduate Research Fellowship 2023354859 (Je.H.). Pe.K. and Hao.Z. were partially supported by NSF AccelNet awards 1927354 (JHU) and 1927275 (WPI). A.M.O.'s contribution was supported by the Stanford Institute for Human-Centered Artificial Intelligence and the Stanford Robotics Center. A.E.A.'s contribution was supported by the Stanford Institute for Human-Centered Artificial Intelligence. T.H., E.L. and K.T. were partially supported by Project 2024-1.2.3-HU-RIZONT-00069, implemented with the support provided by the Ministry of Culture and Innovation of Hungary from the National Research, Development and Innovation Fund, financed under the 2024-1.2.3-HU-RIZONT funding scheme. T.H. is a consolidator researcher, supported by the Distinguished Researcher Program of Óbuda University. Zh.C., Yam.Z., T.Y., and Yun.T. were funded by the Multi-scale Medical Robotics Center, AIR@InnoHK. Xia.C. was funded by the Multi-scale Medical Robotics Center, AIR@InnoHK and Direct Grants (The Chinese University of Hong Kong) under Grant 4055245. P.V., D.J., B.C., and Ju.H. were supported by the Engineering and Physical Sciences Research Council (EPSRC) under Grants EP/V047914/1 and UKRI180 and by the National Institute for Health and Care Research (NIHR) Leeds Biomedical Research Centre (BRC) (NIHR203331). J.L., M.S., G.W., Zi.H., E.W., and H.R. were supported by the NSFC Young Scientists Fund - Scheme A T252500134, the Ministry of Science and Technology (MOST) of China Key Project 2025YFE0122500, and the Hong Kong Research Grants Council (RGC) General Research Fund (GRF 14204524, 14206125). E.W., Ru.J., Z.L., N.Z., Xi.Z., and H.R. received funding from NSFC/RGC Joint Research Scheme (N_CUHK420/22). X.G. and H.C. were supported by the Hong Kong Research Grants Council Early Career Scheme grant 22203525. Mat.W., Pr.K., Sab.M., and M.N. received funding from the European Union's Horizon 2020 research and agreement No 857533. The contribution was made within the project of the Minister of Science and Higher Education "Support for the activity of Centers of Excellence established in Poland

under Horizon 2020" on the basis of the contract number MEiN/2023/DIR/3796. Ra.Y., O.H., Mar.W., S.B., and S.S. were supported by the German Research Foundation (DFG, Deutsche Forschungsgemeinschaft) as part of Germany's Excellence Strategy - EXC 2050/1 - Project ID 390696704 - Cluster of Excellence "Centre for Tactile Internet with Human-in-the-Loop" (CeTI). A.R.J. was supported by the Federal Ministry of Research, Technology, and Space (BMFTR) as part of the research program Communication Systems "Souverän. Digital. Vernetzt.". Joint project 6G-life, project identification number: 16KIS2413K. Lo.M. was supported by the project "Next Generation AI Computing (gAIn)," funded by the Bavarian Ministry of Science and the Arts and the Saxon Ministry for Science, Culture, and Tourism. F.A., M.R.J., Si.K., and So.K. were supported by the National Cancer Institute of the National Institutes of Health under Award Number R21CA280747, the IC2 Institute, and Texas Robotics Seed Collaborative Funding at the University of Texas at Austin. Al.A., Giu.D., Fe.B., Ke.H., Gio.D., Fed.L., Lu.M., and C.A.A. were supported by the European Research Council (ERC) under the Horizon Europe programme (Grant EndoTheranostics, Grant Agreement No. 101118626, https://doi.org/10.3030/101118626), Funded by the European Union. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union, the European Research Council Executive Agency, or the Health and Digital Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Author contributions: Data Contribution: Everyone, Conceptualization: N.N., J.C., D.H., Z.J., A.K., S.D.H., M.A.; Methodology: N.N., J.C., A.K., S.D.H., L.Z.; Evaluation: N.N., J.C., Je.H., Xin.C., D.G.M., P.T.; Software: N.N., J.C., L.Z.; Visualization: N.N., J.C., Je.H., L.Z.; Data Curation: N.N., J.C., Xin.C., D.H.; Formal analysis: N.N., Je.H.; Funding acquisition: A.K., S.D.H., M.A., Na.N.; Supervision: A.K., S.D.H., M.A., Na.N.; Writing - original draft: Je.H., N.N., J.C., Xin.C., L.Z.; Writing - review and editing: Je.H., N.N., J.C., Xin.C., L.Z., A.K., S.D.H., Al.K., A.M.O., K.G., T.H., Pe.K.;

Competing interests: The following authors declare competing interests. All other authors have no competing interests to declare.

- A.K. is a co-founder of Semaphor Surgical.
- F.F. receives consultant fees from Boston Scientific and Johnson & Johnson. F.F. has received research grants from Intuitive Surgical.
- A.M.O. has received research grants from Intuitive Surgical.

Data and materials availability: The Open-H-Embodiment dataset is publicly available on Hugging Face under a CC-BY-4.0 license at `https://huggingface.co/datasets/nvidia/PhysicalAI-Robotics-Open-H-Embodiment`. Data collection guidelines, conversion scripts, and validation tools are maintained in the Open-H data collection repository at `https://github.com/open-h/open-h-embodiment`. GR00T-H model weights are available at `https://huggingface.co/nvidia/GR00T-H`, with training and inference code at `https://github.com/NVIDIA-Medtech/GR00T-H`. LingBot-VA-Suture model weights are available at `https://huggingface.co/semaphorsurgical/lingbot-va-suture`, with code at `https://github.com/semaphor-surgical/lingbot-va-suture`. Cosmos-H-Surgical-Simulator model weights are available at `https://huggingface.co/nvidia/Cosmos-H-Surgical-Simulator`, with training, inference, and evaluation code at `https://github.com/NVIDIA-Medtech/Cosmos-H-Surgical-Simulator`. All materials needed to reproduce the results in this paper are accessible through these repositories without restriction beyond the stated licenses.

Supplementary materials

Supplementary Text
Figure S1
Tables S1 to S4
Movies S1 to S3

Supplementary Materials for Open-H-Embodiment: A Large-Scale Dataset for Enabling Foundation Models in Medical Robotics

Open-H-Embodiment Consortium

Nigel Nelson, et al.

This PDF file includes:

Supplementary Text

World Action Model Baseline: LingBot-VA

To assess whether video-based world-action modeling offers a viable alternative for accurate action prediction in surgical autonomy, we adapt LingBot-VA (*40*) to the SutureBot suturing task, producing LingBot-VA-Suture. LingBot-VA-Suture is initialized from the Wan2.2-5B video diffusion model: at each autoregressive step it predicts a chunk of future latent visual states via conditional flow matching and simultaneously decodes the corresponding end-effector actions through inverse dynamics. This formulation decouples visual dynamics modeling from action control, in contrast to GR00T-H and ACT, which couple both within a single observation-to-action mapping.

Surgical Adaptation. We follow a two-stage procedure. The base LingBot-VA checkpoint is first further pretrained on 50 hours of JHU IMERSE dVRK-Si datasets from Open-H-Embodiment to acquire surgical visual and kinematic priors, and then post-trained on the SutureBot dataset at a reduced learning rate. LingBot-VA-Suture processes three synchronized camera views (endoscope, left wrist, and right wrist) and outputs 16-dimensional actions (7-DoF end-effector pose plus gripper jaw angle per arm) at a chunk size of 48 actions at 30 Hz with quantile-based action normalization, together with video predictions at 10 Hz.

Table S1: **Complete Open-H-Embodiment dataset inventory.** Each row corresponds to one contributed dataset in the final release. The Scale column lists episodes / frames / hours. A dagger (†) after the dataset name indicates that ground-truth kinematics are not included in the released dataset. Abbreviations: Ster. = stereo endoscope, Endo. = endoscope, US = ultrasound, Wr. = wrist camera, D = depth, TPV = third-person view, Seg. = segmentation, HO = handover, PU = pickup, Sim. = simulated/simulation, Bench = benchtop/phantom, Clin. = clinical, Ex V. = ex vivo, In V. = in vivo.

Group	Dataset	Domain / Task [Env.]	Scale	Robot	Views
Balgrist	sonogym_open_h_US_ guidance_L1	US – Robotic US (US Guidance) [Sim.]	1,024 ep 183,296 fr 0.42 h	Sim. KUKA Med14	US
Balgrist	sonogym_open_h_US_ guidance_L2	US – Robotic US (US Guidance) [Sim.]	1,024 ep 183,296 fr 0.42 h	Sim. KUKA Med14	US

Continued on next page

Table S1 continued

Group	Dataset	Domain / Task [Env.]	Scale	Robot	Views
Balgrist	sonogym_open_h_US_guidance_L3	US – Robotic US (US Guidance) [Sim.]	1,024 ep 183,296 fr 0.42 h	Sim. KUKA Med14	US
Balgrist	sonogym_open_h_US_guidance_L4	US – Robotic US (US Guidance) [Sim.]	1,024 ep 183,296 fr 0.42 h	Sim. KUKA Med14	US
Balgrist	sonogym_open_h_US_guidance_L5	US – Robotic US (US Guidance) [Sim.]	1,024 ep 183,296 fr 0.42 h	Sim. KUKA Med14	US
Balgrist	ultrabones_lerobot_dataset_full	US – Robotic US (US Scan) [Bench]	60 ep 90,371 fr 1.20 h	Sim. KUKA Med14	US
Balgrist	ultrabones_lerobot_dataset_full_2	US – Robotic US (US Scan) [Bench]	12 ep 20,668 fr 0.27 h	Sim. KUKA Med14	US
Balgrist	ultrabones_lerobot_dataset_full_2_synthetic_robot_2	US – Robotic US (US Scan) [Sim.]	12 ep 20,668 fr 0.27 h	Sim. KUKA Med14	US
Balgrist	ultrabones_lerobot_dataset_full_3	US – Robotic US (US Scan) [Bench]	18 ep 24,104 fr 0.32 h	Sim. KUKA Med14	US
Balgrist	ultrabones_lerobot_dataset_full_3_synthetic_robot_2	US – Robotic US (US Scan) [Sim.]	18 ep 24,104 fr 0.32 h	Sim. KUKA Med14	US
Balgrist	ultrabones_lerobot_dataset_full_synthetic_robot	US – Robotic US (US Scan) [Sim.]	60 ep 90,371 fr 1.20 h	Sim. KUKA Med14	US
CMR Surg.	cholecystectomy	Surg. – Cholecystectomy (Chole.) [Clin.]	4,792 ep 16,999,777 fr 78.70 h	Versius	Endo.
CMR Surg.	drybox	Surg. – Skills Benchmark (Peg Transfer) [Bench]	480 ep 1,677,540 fr 7.77 h	Versius	Endo.
CMR Surg.	hysterectomy	Surg. – Hysterectomy (Hysterect.) [Clin.]	7,432 ep 26,374,851 fr 122.11 h	Versius	Endo.
CMR Surg.	inguinal_hernia	Surg. – Hernia Repair (Hernia Rep.) [Clin.]	7,286 ep 25,807,467 fr 119.48 h	Versius	Endo.
CMR Surg.	peg_transfer_extra	Surg. – Skills Benchmark (Peg Transfer) [Bench]	216 ep 762,828 fr 3.53 h	Versius	Endo.
CMR Surg.	prostatectomy	Surg. – Prostatectomy (Prostatect.) [Clin.]	10,329 ep 36,516,224 fr 169.06 h	Versius	Endo.
CUHK	bmt_insertion_dataset	US – US-Guided Interv. (Needle Ins.) [Ex V.]	847 ep 814,451 fr 7.54 h	UR5e	US, TPV
CUHK	bmt_insertion_tumor_fishball_dataset	US – US-Guided Interv. (Needle Ins.) [Ex V.]	414 ep 401,977 fr 3.72 h	UR5e	US, TPV
CUHK	bmt_insertion_vessel_dataset	US – US-Guided Interv. (Needle Ins.) [Ex V.]	665 ep 637,649 fr 5.90 h	UR5e	US, TPV
CUHK	bmt_needle_insertion_dataset	US – US-Guided Interv. (Needle Ins.) [Ex V.]	386 ep 376,835 fr 3.49 h	UR5e	US, TPV
CUHK	bmt_tumor(grape)_insertion_dataset	US – US-Guided Interv. (Needle Ins.) [Ex V.]	99 ep 107,054 fr 0.99 h	UR5e	US, TPV

Continued on next page

Table S1 continued

Group	Dataset	Domain / Task [Env.]	Scale	Robot	Views
CUHK	US_PPNR	US – Probe Placement and Needle Retrieval [Ex V.]	2,546 ep 1,172,070 fr 10.85 h	UR5e	US, Wr., D, TPV
CUHK	Find_greater_curvature	Endo. – Flex. Endoscopy (Anat. Loc.) [Bench]	462 ep 107,488 fr 1.49 h	Custom Endo. Rob.	Endo.
CUHK	Find_lesser_curvature	Endo. – Flex. Endoscopy (Anat. Loc.) [Bench]	200 ep 61,997 fr 0.86 h	Custom Endo. Rob.	Endo.
CUHK	Find_pyloric_antrum	Endo. – Flex. Endoscopy (Anat. Loc.) [Bench]	204 ep 28,731 fr 0.40 h	Custom Endo. Rob.	Endo.
CUHK	Track_orange_lesion	Endo. – Flex. Endoscopy (Lesion Track.) [Bench]	156 ep 33,345 fr 0.46 h	Custom Endo. Rob.	Endo.
CUHK	Track_small_round_nodule	Endo. – Flex. Endoscopy (Lesion Track.) [Bench]	722 ep 249,461 fr 3.46 h	Custom Endo. Rob.	Endo.
CUHK	Track_white_oval_ROI	Endo. – Flex. Endoscopy (ROI Track.) [Bench]	414 ep 98,908 fr 1.37 h	Custom Endo. Rob.	Endo.
CUHK	Tracked_EUS	US – Endo. US (Tracked EUS) [In V.]	34 ep 33,425 fr 0.37 h	dVRK	Endo., US
CUHK	Tracked_US	US – Robotic US (Tracked US) [Clin.]	30 ep 7,007 fr 0.08 h	dVRK	US, TPV
Hamlyn/ Imperial	knot_tying	Surg. – Suture/Knot (Knot Tying) [Ex V.]	49 ep 30,222 fr 0.28 h	dVRK	Endo., Wr., D
Hamlyn/ Imperial	needle_grasp_and_handover	Surg. – Suture/Knot (Needle HO) [Bench]	137 ep 46,582 fr 0.43 h	dVRK	Endo., Wr., D
Hamlyn/ Imperial	peg_transfer	Surg. – Skills Benchmark (Peg Transfer) [Bench]	317 ep 102,187 fr 0.95 h	dVRK	Endo., Wr., D
Hamlyn/ Imperial	Suturing-1	Surg. – Suture/Knot (Suturing) [Ex V.]	180 ep 100,067 fr 0.93 h	dVRK	Endo., Wr., D
Hamlyn/ Imperial	Suturing-2	Surg. – Suture/Knot (Suturing) [Ex V.]	186 ep 251,355 fr 2.33 h	dVRK	Endo., Wr., D
Hamlyn/ Imperial	tissue_lifting	Surg. – Tissue Manip. (Lifting) [Ex V.]	75 ep 8,180 fr 0.15 h	dVRK	Endo., Wr., D
Hamlyn/ Imperial	Tissue_Retraction	Surg. – Tissue Manip. (Retraction) [Ex V.]	75 ep 14,160 fr 0.13 h	dVRK	Endo., Wr., D
HKBU	46_datasets_summary	US – Robotic US (US Nav.) [Sim.]	6,072 ep 322,000 fr 4.47 h	Sim. US Platform	US
ImFusion	ImFusion_dataset_corrected	US – Robotic US (US Scan) [Bench]	1,644 ep 559,897 fr 5.18 h	Franka Panda	US, Wr., TPV
JHU	cao_cautery_combined	Surg. – Tissue Manip. (Debride.) [Ex V.]	783 ep 52,748 fr 0.49 h	dVRK-Si	Ster., Wr.
JHU	cautery	Surg. – Tissue Manip. (Debride.) [Ex V.]	22 ep 5,288 fr 0.05 h	dVRK-Si	Ster., Wr.

Continued on next page

Table S1 continued

Group	Dataset	Domain / Task [Env.]	Scale	Robot	Views
JHU	Cholecystectomy	Surg. – Cholecystectomy (Chole.) [Ex V.]	750 ep 181,021 fr 1.68 h	dVRK-Si	Ster., Wr.
JHU	endosrt	Endo. – Flex. Endoscopy (Endo. Nav.) [Bench]	2,047 ep 193,203 fr 2.15 h	Flex. Endo. Platf.	Other
JHU	hf_suturebot	Surg. – Suture/Knot (Suturing) [Bench]	1,452 ep 516,334 fr 4.78 h	dVRK-Si	Ster., Wr.
JHU	nephfat	Surg. – Tissue Manip. (Debride.) [Bench]	2,112 ep 494,525 fr 4.58 h	dVRK-Si	Ster., Wr.
JHU	SurgSync-stitch-coldcut-P1	Surg. – Suture/Knot (Suturing) [Ex V.]	578 ep 53,114 fr 1.48 h	dVRK-Si	Endo.
JHU	SurgSync-stitch-coldcut-P2	Surg. – Suture/Knot (Suturing) [Ex V.]	313 ep 32,426 fr 0.90 h	dVRK-Si	Endo.
JHU	SurgSync-stitch-coldcut-P3	Surg. – Suture/Knot (Suturing) [Ex V.]	196 ep 17,485 fr 0.49 h	dVRK-Si	Endo.
JHU	Prepare to Pierce	Surg. – Suture/Knot (Needle Pos.) [Bench]	2 ep 582 fr 0.01 h	dVRK-Si	Ster.
JHU	srt_needle_pickup+handover	Surg. – Suture/Knot (Needle PU) [Bench]	240 ep 58,305 fr 0.54 h	dVRK	Ster., Wr.
JHU	srt_tissue_lift	Surg. – Tissue Manip. (Lifting) [Bench]	225 ep 27,487 fr 0.25 h	dVRK	Ster., Wr.
JHU	srth_porcine_chole_fix	Surg. – Cholecystectomy (Chole.) [Ex V.]	14,836 ep 1,878,393 fr 17.39 h	dVRK-Si	Ster., Wr.
JHU	star_IL	Surg. – Suture/Knot (Suturing) [Ex V.]	512 ep 216,140 fr 2.00 h	KUKA LBR iiwa	Endo., Wr.
JHU	wound_closure	Surg. – Suture/Knot (Suturing) [Bench]	216 ep 1,883,971 fr 17.44 h	dVRK-Si	Ster., Wr.
Moon Surg.	moon	Surg. – Camera/View Mgmt. (Cam. Guidance) [Clin.]	65 ep 12,020 fr 0.11 h	Maestro	Endo., TPV
Obuda	FRS_Dome_1	Surg. – Suture/Knot (Suturing) [Bench]	102 ep 141,078 fr 1.31 h	dVRK	Ster., Wr., D
Obuda	NeedleThreading_1	Surg. – Suture/Knot (Suturing) [Bench]	196 ep 103,067 fr 0.95 h	dVRK	Ster., Wr., D
Obuda	NeedleThreading_2	Surg. – Suture/Knot (Suturing) [Bench]	204 ep 102,221 fr 0.95 h	dVRK	Ster., Wr., D
Obuda	PegTransfer_1	Surg. – Skills Benchmark (Peg Transfer) [Bench]	216 ep 134,832 fr 1.25 h	dVRK	Ster., Wr., D
Obuda	PegTransfer_2	Surg. – Skills Benchmark (Peg Transfer) [Bench]	184 ep 78,140 fr 0.72 h	dVRK	Ster., Wr., D
Obuda	Pork_1	Surg. – Tissue Manip. (Grasping) [Ex V.]	318 ep 165,486 fr 1.53 h	dVRK	Ster., Wr., D
Obuda	Rollercoaster_1	Surg. – Skills Benchmark (Spatial Nav.) [Bench]	95 ep 130,268 fr 1.21 h	dVRK	Ster., Wr., D

Continued on next page

Table S1 continued

Group	Dataset	Domain / Task [Env.]	Scale	Robot	Views
Obuda	Seaspike_1	Surg. – Skills Benchmark (Ring Transfer) [Bench]	207 ep 89,269 fr 0.83 h	dVRK	Ster., Wr., D
Obuda	Seaspike_2	Surg. – Skills Benchmark (Ring Transfer) [Bench]	153 ep 67,658 fr 0.63 h	dVRK	Ster., Wr., D
Obuda	Seaspike_3	Surg. – Skills Benchmark (Ring Transfer) [Bench]	219 ep 102,948 fr 0.95 h	dVRK	Ster., Wr., D
Obuda	Skinphantom_1	Surg. – Suture/Knot (Suturing) [Bench]	106 ep 41,979 fr 0.39 h	dVRK	Ster., Wr., D
HK PolyU	OpenH_Dataset_full	Surg. – Cholecystectomy (Retraction) [Sim.]	11,520 ep 5,760,000 fr 53.33 h	Sim. dVRK	Endo.
Rob Surg.	all_merged_data (hemicolectomy)	Surg. – Clinical Proc. (Hemicolect.) [Clin.]	5 ep 1,003,887 fr 7.98 h	BiTrack	Endo.
Rob Surg.	all_merged_data (hysterectomy)	Surg. – Clinical Proc. (Hysterect.) [Clin.]	1 ep 1,003,887 fr 1.32 h	BiTrack	Endo.
SanoScience	Expert_demonstrations	Surg. – Cholecystectomy (Chole.) [Sim.]	5,454 ep 603,054 fr 6.70 h	XR Simulator	Endo., D, Seg.
SanoScience	NonExpert_failure	Surg. – Cholecystectomy (Chole.) [Sim.]	126 ep 17,622 fr 0.16 h	XR Simulator	Endo., D, Seg.
SanoScience	NonExpert_full_ modalities	Surg. – Cholecystectomy (Chole.) [Sim.]	6,156 ep 713,070 fr 6.60 h	XR Simulator	Endo., D, Seg.
SanoScience	NonExpert_partial_ modalities	Surg. – Cholecystectomy (Chole.) [Sim.]	666 ep 58,752 fr 0.54 h	XR Simulator	Endo.
SanoScience	NonExpert_recovery	Surg. – Cholecystectomy (Chole.) [Sim.]	126 ep 20,376 fr 0.19 h	XR Simulator	Endo., D, Seg.
SanoScience	NonExpert_stereo	Surg. – Cholecystectomy (Chole.) [Sim.]	1,602 ep 179,640 fr 1.66 h	XR Simulator	Ster., D, Seg.
Semaphor	open_h_semaphor	Surg. – Suture/Knot (Suturing) [Ex V.]	535 ep 47,050 fr 0.44 h	Manual Lap. Tools	TPV
Semaphor	open_h_semaphor_1.18	Surg. – Suture/Knot (Suturing) [Ex V.]	470 ep 50,664 fr 0.47 h	Manual Lap. Tools	TPV
Stanford	block_transfer_sim_ lerobot_1_28	Surg. – Skills Benchmark (Peg Transfer) [Sim.]	500 ep 236,968 fr 3.46 h	Sim. dVRK	Ster.
Stanford	Needle Transfer	Surg. – Suture/Knot (Needle HO) [Bench]	700 ep 313,882 fr 2.91 h	dVRK-Si	Ster.
Stanford	needle_transfer_sim_ lerobot_1_28	Surg. – Suture/Knot (Needle HO) [Sim.]	500 ep 229,232 fr 3.35 h	Sim. dVRK	Ster.
Stanford	Peg Transfer	Surg. – Skills Benchmark (Peg Transfer) [Bench]	598 ep 268,729 fr 2.49 h	dVRK-Si	Ster.
Stanford	Tissue Retraction	Surg. – Tissue Manip. (Retraction) [Bench]	698 ep 291,826 fr 2.70 h	dVRK-Si	Ster.
TU Dresden	endoscope_guidance	Surg. – Camera/View Mgmt. (Cam. Guidance) [In V.]	50 ep 58,605 fr 0.54 h	UR5e	Endo.

Continued on next page

Table S1 continued

Group	Dataset	Domain / Task [Env.]	Scale	Robot	Views
TU Dresden	endoscope_guidance	Surg. – Camera/View Mgmt. (Cam. Guidance) [In V.]	50 ep 56,484 fr 0.52 h	UR5e	Endo.
TU Dresden	grasping_retraction	Surg. – Tissue Manip. (Retraction) [In V.]	116 ep 35,406 fr 0.33 h	UR5e	Endo.
TU Dresden	grasping_retraction	Surg. – Tissue Manip. (Retraction) [In V.]	146 ep 47,753 fr 0.44 h	UR5e	Endo.
TUM CAMP	SonATA_all	US – Robotic US (US Scan) [Bench]	2,397 ep 633,604 fr 5.87 h	Franka Panda	US, Wr., TPV
Turin	mitic_lerobot_ex_vivo	Surg. – Suture/Knot (Suturing) [Ex V.]	799 ep 388,690 fr 3.60 h	dVRK	Ster.
Turin	mitic_lerobot_plastic_pad	Surg. – Suture/Knot (Suturing) [Bench]	550 ep 149,846 fr 1.39 h	dVRK	Ster.
Turin	mitic_lerobot_plastic_pad_3DMED	Surg. – Suture/Knot (Suturing) [Bench]	370 ep 243,229 fr 2.25 h	dVRK	Ster.
Turin	mitic_lerobot_plastic_tube	Surg. – Suture/Knot (Suturing) [Bench]	480 ep 216,070 fr 2.00 h	dVRK	Ster.
UBC	GauzeCutting_merged†	Surg. – Benchtop Tasks (Gauze Cut.) [Bench]	234 ep 75,277 fr 0.69 h	da Vinci Si	Ster.
UBC	KnotTying_merged†	Surg. – Suture/Knot (Knot Tying) [Bench]	523 ep 75,840 fr 0.70 h	da Vinci Si	Ster.
UBC	NeedlePassing_merged†	Surg. – Suture/Knot (Needle HO) [Bench]	685 ep 70,915 fr 0.67 h	da Vinci Si	Ster.
UBC	PegTransfering_merged†	Surg. – Skills Benchmark (Peg Transfer) [Bench]	550 ep 14,305 fr 0.13 h	da Vinci Si	Ster.
UBC	PickAndPlace_merged†	Surg. – Tissue Manip. (Grasping) [Bench]	702 ep 35,681 fr 0.33 h	da Vinci Si	Ster.
UBC	Suturing_merged†	Surg. – Suture/Knot (Suturing) [Bench]	606 ep 103,917 fr 0.97 h	da Vinci Si	Ster.
UBC	WireChasing3D_merged†	Surg. – Skills Benchmark (Wire Chasing) [Bench]	523 ep 41,102 fr 0.38 h	da Vinci Si	Ster.
UBC	WireChasing_merged†	Surg. – Skills Benchmark (Wire Chasing) [Bench]	615 ep 60,042 fr 0.55 h	da Vinci Si	Ster.
UC Berkeley	debridement_lerobot	Surg. – Tissue Manip. (Debride.) [Ex V.]	589 ep 221,950 fr 2.06 h	dVRK	Ster.
UCSD	surgical_learning_dataset	Surg. – Tissue Manip. (Dissection) [Bench]	912 ep 288,604 fr 2.67 h	dVRK	Ster.
UCSD	surgical_learning_dataset2	Surg. – Tissue Manip. (Retraction) [Bench]	200 ep 26,313 fr 0.24 h	dVRK	Ster.
UCSD	surgical_learning_retraction_dataset3	Surg. – Tissue Manip. (Retraction) [Bench]	598 ep 183,622 fr 1.70 h	dVRK	Ster.
UCSD	surgical_learning_retraction_failurecase	Surg. – Tissue Manip. (Retraction) [Bench]	299 ep 66,839 fr 0.62 h	dVRK	Ster.

Continued on next page

Table S1 continued

Group	Dataset	Domain / Task [Env.]	Scale	Robot	Views
UIC	UIC_CRCD_LeRobot	Surg. – Cholecystectomy (Chole.) [Ex V.]	18 ep 755,891 fr 3.50 h	dVRK	Endo.
USTC/ Tuodao	exvivo_liver_sep	Surg. – Resect./Dissect. (Liver Dissect.) [Ex V.]	666 ep 121,922 fr 1.41 h	Torin	Ster.
USTC/ Tuodao	grasp_on_liver	Surg. – Tissue Manip. (Grasping) [Ex V.]	817 ep 63,538 fr 0.74 h	Torin	Ster.
USTC/ Tuodao	invivo_liver_sep	Surg. – Resect./Dissect. (Liver Dissect.) [In V.]	199 ep 39,899 fr 0.46 h	Torin	Ster.
USTC/ Tuodao	knot_tying	Surg. – Suture/Knot (Knot Tying) [Bench]	1,098 ep 182,836 fr 2.12 h	Torin	Ster.
USTC/ Tuodao	Needle_handover	Surg. – Suture/Knot (Needle HO) [Bench]	260 ep 34,990 fr 0.40 h	Torin	Ster.
USTC/ Tuodao	needle_pickup	Surg. – Suture/Knot (Needle PU) [Bench]	616 ep 57,172 fr 0.66 h	Torin	Ster.
USTC/ Tuodao	tissue_lifting	Surg. – Tissue Manip. (Lifting) [Bench]	110 ep 11,673 fr 0.14 h	Torin	Ster.
UT Austin	colonoscope-lerobot	Endo. – Colonoscopy (Colon. Nav.) [Bench]	1,894 ep 2,095,587 fr 19.40 h	Cobra Colonoscope	Endo.
UTenn	benchtop_datasets_ round2_with_part_ seg†	Surg. – Benchtop Tasks (Tool Track.) [Bench]	2 ep 223 fr 0.01 h	dVRK	Endo., D
UTenn	surgical_video_ datasets†	Surg. – Surg. Video (Annot.) [Clin.]	73 ep 8,183 fr 0.08 h	da Vinci Xi	Endo., D, Seg.
UTenn	surgical_video_ datasets_round2_with_ part_seg†	Surg. – Surg. Video (Annot.) [Clin.]	101 ep 8,760 fr 0.08 h	da Vinci Xi	Endo., D, Seg.
UTenn	surgical_video_ datasets_with_tip_ pose†	Surg. – Surg. Video (Annot.) [Clin.]	62 ep 6,690 fr 0.06 h	da Vinci Xi	Endo., D
Virtual Inc.	Mira Needle Pickup	Surg. – Benchtop Tasks (Needle PU) [Bench]	150 ep 39,960 fr 0.37 h	Mira	Endo.

Table S3: Open-H dataset mixture used in GR00T-H training

Dataset / Embodiment Group	Mixture ratio
CMR Versius	**0.1920**
Cholecystectomy	
Hysterectomy	
Inguinal hernia	
Prostatectomy	
Drybox peg transfer	
JHU dVRK-Si	**0.3498**
ARCADE Cholecystectomy	

Dataset / Embodiment Group	Mixture ratio
ARCADE Cautery	
IMERSE porcine cholecystectomy	
IMERSE CAO cautery	
IMERSE needle pickup and handover	
IMERSE tissue lift	
IMERSE needle pickup	
IMERSE wound closure	
IMERSE SutureBot	
IMERSE NephFat	
JHU IMERSE monocular	**0.0372**
SutureBot	
JHU SMARTS	**0.0074**
SurgSync stitch coldcut P1	
SurgSync stitch coldcut P2	
SurgSync stitch coldcut P3	
Stanford dVRK	**0.0630**
Needle transfer	
Tissue retraction	
Peg transfer	
Obuda dVRK	**0.0834**
FRS Dome 1	
NeedleThreading 1	
PegTransfer 1	
Rollercoaster 1	
Seaspike 1	
NeedleThreading 2	
PegTransfer 2	
Pork 1	
Seaspike 2	
Seaspike 3	
Skinphantom 1	
Rob Surgical	**0.0724**
All merged data	
JHU KUKA IMERSE	**0.0085**
star_IL	
USTC / Torin	**0.0369**
Ex vivo liver separation	
Grasp on liver	
In vivo liver separation	
Knot tying	
Needle handover	

Dataset / Embodiment Group	Mixture ratio
Needle pickup	
Tissue lifting	
Hamlyn dVRK	**0.0393**
Suturing-2	
Peg transfer	
Suturing-1	
Needle grasp and handover	
Knot tying	
Tissue retraction	
UCSD surgical learning	**0.0227**
Surgical learning dataset	
Surgical learning dataset 2	
UC Berkeley dVRK	**0.0160**
Debridement	
Turin MITIC	**0.0680**
MITIC ex vivo	
Plastic pad	
Plastic pad 3DMED	
Plastic tube	
TUD TUNDRA	**0.0034**
Grasping retraction	

Table S4: Open-H dataset mixture used in C-H-S-S training

Dataset / Embodiment Group	Mixture ratio
CMR Versius	**4.000**
Cholecystectomy	
Hysterectomy	
Inguinal hernia	
Prostatectomy	
JHU IMERSE	**2.015**
SRT-H porcine cholecystectomy	
Electrocautery tumor resection	
SutureBot	
Needle pickup and handover	
Suturing	
Tissue lift	
Pickup only	
JHU ARCADE	**0.134**
Cholecystectomy	

Dataset / Embodiment Group	Mixture ratio
Cautery	
Stanford dVRK	**0.638**
Needle transfer	
Tissue retraction	
Peg transfer	
Hamlyn dVRK	**0.396**
Suturing-2	
Peg transfer	
Suturing-1	
Needle grasp and handover	
Knot tying	
Tissue retraction	
Turin MITIC	**0.283**
Ex vivo	
UCSD dVRK	**0.229**
Surgical learning dataset	
Surgical learning dataset 2	
UC Berkeley dVRK	**0.162**
Debridement	
USTC Torin	**0.135**
Knot tying	
Needle handover	
Needle pickup	
Moon Surgical	**0.009**
Moon	

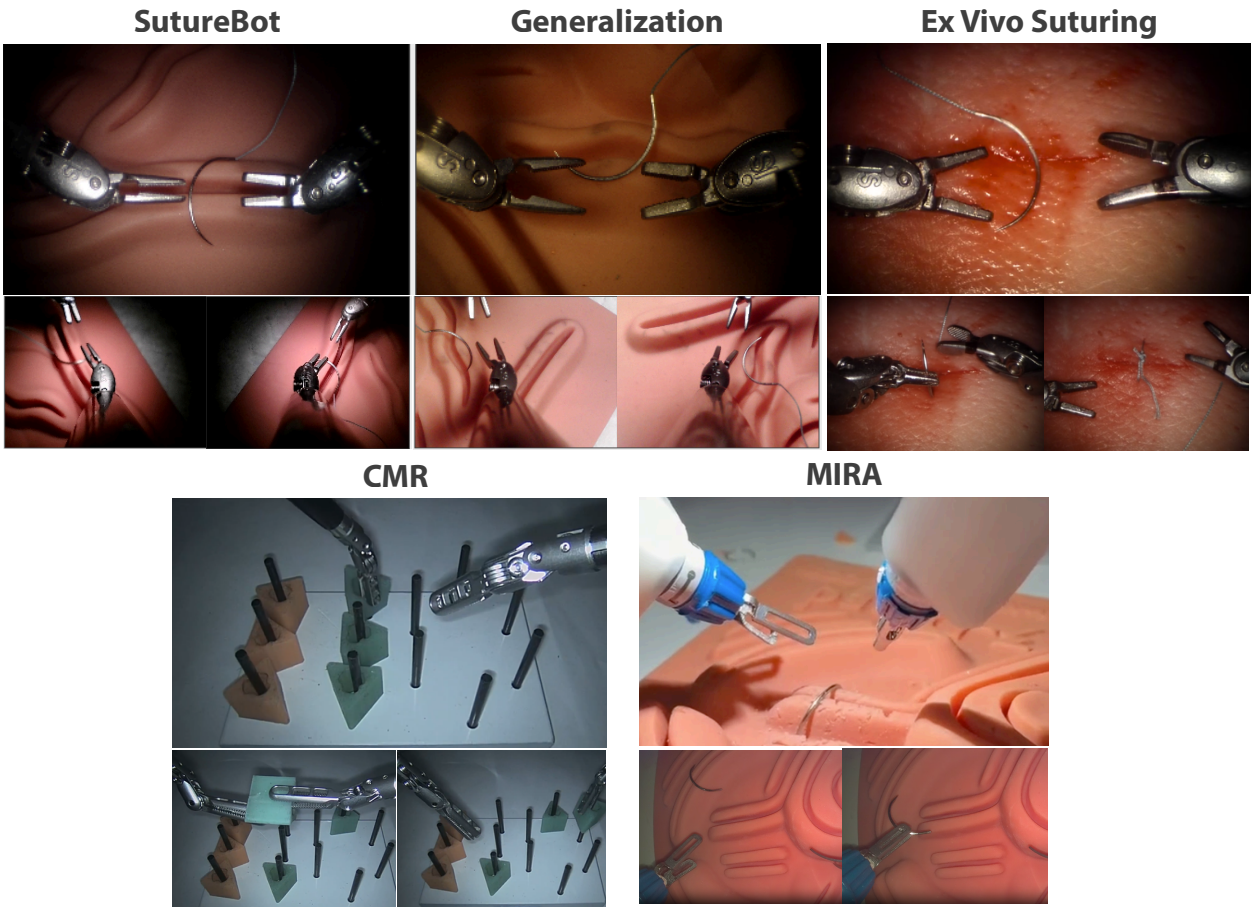
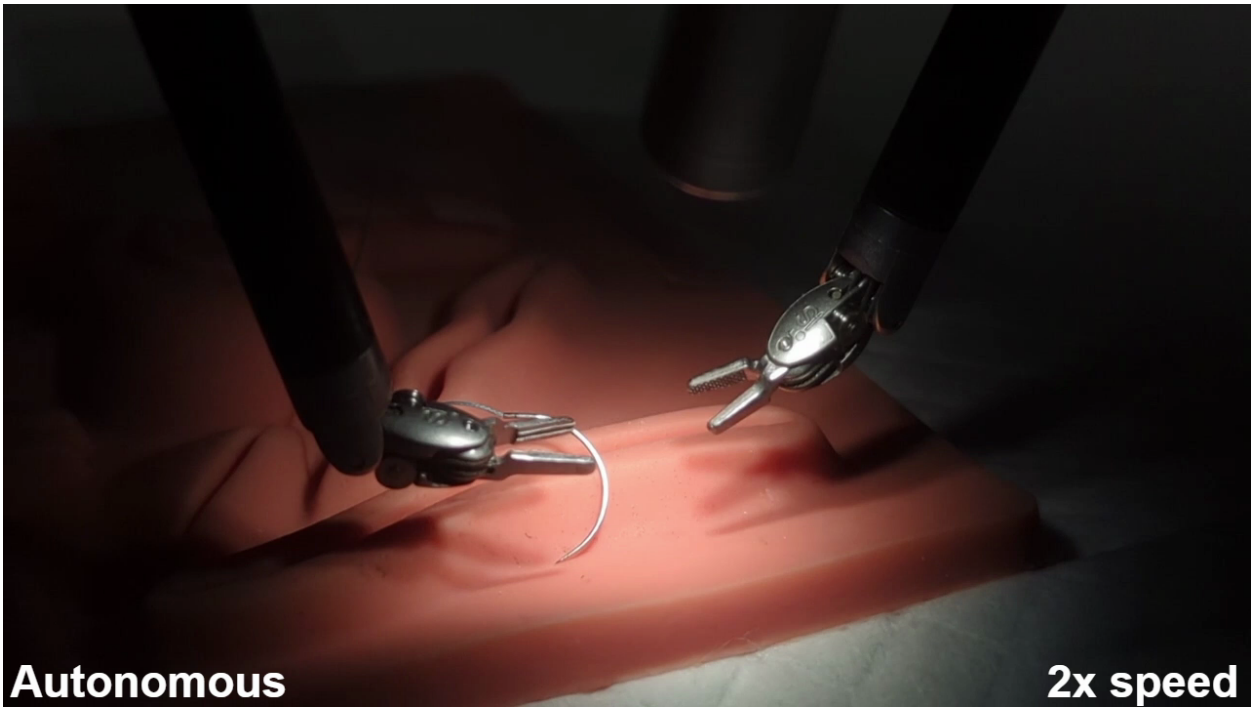
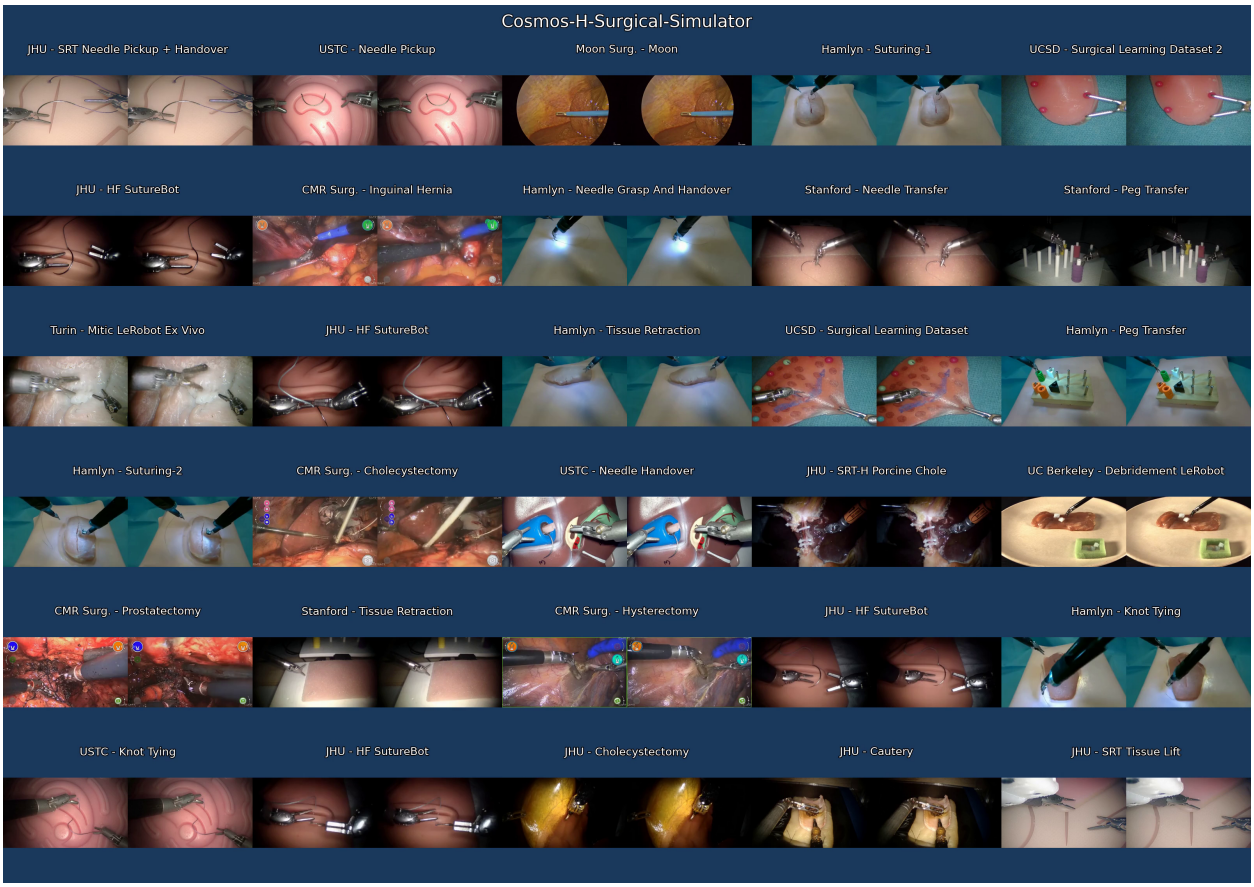


Figure S1: **Experimental setups for GR00T-H policy evaluation.** SutureBot: The primary evaluation environment, utilizing a da Vinci Research Kit Si (dVRK-Si), silicone phantom pad, and suture needle. Generalization: An out-of-distribution testbed featuring wound geometries absent from the training data and external lighting sources to replace standard endoscope illumination. MIRA and CMR: Robotic platforms used to validate multi-embodiment policy performance. Ex vivo suturing: Performed with skin-on pork belly to assess performance on real tissue.



Movie S1: Successful end-to-end suturing rollout by GR00T-H on the SutureBot benchmark. GR00T-H controls the dVRK-Si to pick up and hand over the needle, perform needle throw and extraction, and complete the task with a knot tie.

Movie S2: Successful sub-task rollouts by GR00T-H for ex vivo suturing. The policy is conditioned on current observations and sub-task prompts, with the prompts shown in the top left of the video.



Movie S3: Qualitative results from Cosmos-H-Surgical-Simulator across 30 Open-H datasets, 9 institutions, and 9 embodiments. Each panel shows ground-truth observations (left) alongside model-predicted frames (right), conditioned on recorded kinematic action trajectories.

Table S2: GR00T-H training parameters.

Parameter	Value
Initialization	`nvidia/GR00T-N1.6-3B`
Frozen backbone	false
Action horizon	50
Action normalization	Temporal z-score
Image size	`224 x 392`
Random crop	95%
Random rotation	5 deg
Coarse pixel dropout	0.5
Color jitter	Brightness 0.12, contrast 0.15, saturation 0.15, hue 0.02
Multi-view dropout	0.25
State dropout	100%
Optimizer	AdamW
Learning rate	`3e-5`
Weight decay	`1e-5`
LR schedule	Cosine decay
Warmup	5%
Gradient clipping	1.0
Global batch size	1024
Training steps	65,000
Training compute	32 A100 80GB x 1.5 days