ARTIFICIAL INTELLIGENCE

The robot will see you now: Foundation models are the path forward for autonomous robotic surgery

Michael Yip*

Foundation models in robotics are here to stay, but can surgical robotics keep up with their data-intense requirement?

Foundation models have provided machines a seismic shift in language and visual understanding. Large language models and vision transformers—such as ChatGPT, Stable Diffusion, and CLIP—can not only discriminate text and scenes with remarkable understanding of context but also generate contextually informed text and imagery. Thus, where in the past, the relational and graphical modeling of context was handcrafted and labeled in a domain-specific manner, much of it now latently resides in the parameters of large models learned over massive and diverse datasets. This jump toward artificial general intelligence is something that many fields of robotics have capitalized on.

Surgical robotics is at a unique disadvantage when it comes to incorporating foundation models in advancing its autonomous capabilities. Because all foundation models are learned on the basis of accessible data on the internet or in massive repositories of private data, they will have captured the nuances and contexts of only those domains (however large and diverse). Despite thousands of robotic surgeries being performed daily around the world, there are little to no available data to train on. Why is this? Labeling of images and videos and recording of robotic behaviors are difficult to scale beyond a few surgeries, and it is currently impossible to envision a scenario in which researchers could collect sufficient data to create a generalized foundation model for surgical robot autonomy. Privacy laws and corporations ultimately hold the keys to the data, but practically speaking, there are also not enough expert labelers (ultimately medical practitioners) to label data and little incentive to perform and open source data labeling, and there may never be. Most surgical robotics artificial intelligence (AI) papers are using the same handful of datasets. Cholec80 (*1*), for example, is still a gold standard for visual context understanding (segmentation, phase recognition, etc.), and, at only a few videos as its total dataset, it is a nonrobotic dataset and is microscopic in comparison with those used to train foundation models.

Despite the unavoidable data and practical limitations, surgical robotics can and should move toward a future where foundational models are the backbone of its intelligence. It is, presently, the only way to follow along with the generational leaps in scene understanding and contextual decision-making and reach a level of high autonomy [different levels of autonomy mapped to robotic surgery: Yang *et al.* (*2*)]. Consider a surgical robot that is assisting in surgery as if it were a trained surgical technician or fellow: It should make most low-level decisions on its own, requiring high-level assignment of tasks via voice only when context cannot be gleaned visually from the scene. Incorporating a visual-language model with an action model in an end-to-end manner would be the simplest approach. In scene understanding, SurgicalGPT (*3*) is a visual question-answer (VQA) model that has been shown to use prompt engineering with endoscopic video sequences to identify the procedure, step, and instruments, trained from scratch on a narrow set of procedures. GSViT (*4*) fine-tunes video prediction foundation models of 26.1 million parameters, 13.7 million of which are tunable, for next-frame image prediction and surgical phase annotation.

Unlike these examples, fine-tuning is not an absolute necessity, which can lead to rapid overfitting and poor extrapolation to unseen environments. Another approach is to consider foundation models at the top of a hierarchy of increasingly specific models of context building and decision-making that incorporate domain-specific priors. In this way, foundation models provide warm starts to context, where then a collection of middle-layer models provides additional, domain-specific context to surgical environments. These middle-layer models—which include analytical models, physical simulators, and learned embeddings—offer refinement in a tangible and more modular way compared with foundation model fine-tuning and can be designed to offer guarantees of performance and safety.

In our own laboratory, we see the benefits of leveraging foundation models such as Segment Anything (*5*) for global image segmentation and CoTracker (*6*) for full scene visual tracking; this higher-level information is taken raw, without fine-tuning, and instead "refined" by geometric and physical middle-layer models—splines, volumes, fluids, and surfaces—that have a lower number of parameters to represent many different surgical scenes, tools, and objects. In (*7*), the segmentation models are not sufficient alone to reconstruct suture thread for multistep thread and needle regrasping. However, by pairing segmentation results with spline models and stereo camera models, we can translate the raw foundation model output to an output that also models the reconstruction variance and simultaneously offers reliability-based candidate grasp poses, completing an image-to-grasp pipeline for both needles and suture threads for autonomous suturing. Physical models and low-parameter decision-making tools, such as finite-state machines and behavior trees, can provide structure and constraints based on realistic world assumptions to foundation model outputs. This ensures that they adhere to the physical and safety limits of tissues and instruments during surgery.

One category of foundation models that surgical robotics ought to seek integration with is those aimed at general-purpose manipulation. Robot Transformer X, or RT-X (*8*), is one of the original foundation models for robot manipulation, trained on 22 different robots across 21 institutions, learning 527 skills over 160,266 tasks. Driven by industry interest

Department of Electrical and Computer Engineering, University of California San Diego, La Jolla, CA 92130, USA.
*Corresponding author. Email: yip@ucsd.edu

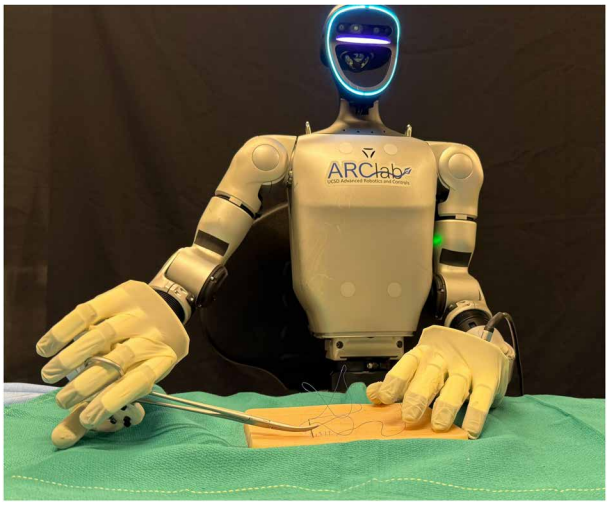

Fig. 1. A surgical humanoid robot holds a forcep in its hands to perform suturing. A new class of surgical robots that focuses on proficient multifingered manipulation would open the door to using the nearly endless array of surgical instrumentation available to doctors while bootstrapping on the rapid gains and foundation models being generated toward multifingered manipulation.

CREDIT: IMAGE FROM (*11*)

and lower barrier to entry over surgical robotics, foundation models for general-purpose robot manipulators such as RT-X serve as excellent "warm starts" to manipulation policies of rigid-body and rigid-multibody objects and could be bridged to deformable object and environment manipulation in surgery. The challenge will be that, morphologically speaking, surgical end effectors have some of the more unique designs (*9*) compared with those on the ends of industrial robot manipulators. Most of the current foundation models in robotics have been trained on fingered grippers, aiming for general ubiquity in grasping arbitrary objects. Given the kinematic and dynamical differences and constraints for surgical robots, there will inevitably be a kinodynamic domain gap to jump that will continually plague transfer of these robotic foundation models to surgical robotic applications.

In a somewhat unconventional view, I believe that surgical robot platforms could benefit substantially by aligning their design more closely with industrial robotic arms and end effectors. In this paradigm, surgical robots would be equipped with multifingered hands or grippers capable of manipulating standard medical instruments. Many surgical procedures, whether open or endoscopic, could be performed largely or entirely using gripper-held medical instruments. By adopting multifingered hands and grippers found to be more prevalent in industrial manipulators, surgical robots could reduce the domain gap between surgical and general-purpose robotics (Fig. 1). This would enhance the applicability of existing robotics datasets to surgical tasks and allow direct use of robot foundation models trained on industrial systems, minimizing or even eliminating the need for retraining. Notably, a leading theory in behavioral science posits that a key leap in human cognitive evolution stemmed from our ability to manipulate tools with our hands (*10*). A similar leap in surgical robotics may emerge once these systems gain the dexterity and intelligence to wield instruments with sufficient competence to perform general surgery.

REFERENCES AND NOTES

1. A. P. Twinanda, S. Shehata, D. Mutter, J. Marescaux, M. de Mathelin, N. Padoy, Endonet: A deep architecture for recognition tasks on laparoscopic videos. *IEEE Trans. Med. Imaging* **36**, 86–97 (2016).
2. G.-Z. Yang, J. Cambias, K. Cleary, E. Daimler, J. Drake, P. E. Dupont, N. Hata, P. Kazanzides, S. Martel, R. V. Patel, V. J. Santos, R. H. Taylor, Medical robotics—Regulatory, ethical, and legal considerations for increasing levels of autonomy. *Sci. Robot.* **2**, eaam8638 (2017).
3. L. Seenivasan, M. Islam, G. Kannan, H. Ren, "SurgicalGPT: End-to-end language-vision GPT for visual question answering in surgery" in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Springer, 2023), pp. 281–290.
4. S. Schmidgall, J. W. Kim, J. Jopling, A. Krieger, General surgery vision transformer: A video pre-trained foundation model for general surgery. arXiv:2403.05949 [cs.CV] (12 April 2024).
5. A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W. Y. Lo, P. Dollár, "Segment anything" in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (IEEE, 2023), pp. 4015–4026.
6. N. Karaev, I. Rocco, B. Graham, N. Neverova, A. Vedaldi, C. Rupprecht, "Cotracker: It is better to track together" in *Computer Vision – ECCV 2024,* A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, G. Varol, Eds., vol. 15120 of *Lecture Notes in Computer Science* (Springer, 2025), pp. 18–35.
7. N. Joglekar, F. Liu, F. Richter, M. C. Yip, Autonomous image-to-grasp robotic suturing using reliability-driven suture thread reconstruction. *IEEE Robot. Autom. Lett.* **10**, 3676–3683 (2024).
8. A. O'Neill, A. Rehman, A. Maddukuri, A. Gupta, A. Padalkar, A. Lee, A. Pooley, A. Gupta, A. Mandlekar, A. Jain, A. Tung, A. Bewley, A. Herzog, A. Irpan, A. Khazatsky, A. Rai, A. Gupta, A. Wang, A. Singh, A. Garg, A. Kembhavi, A. Xie, A. Brohan, A. Raffin, A. Sharma, A. Yavary, A. Jain, A. Balakrishna, A. Wahid, B. Burgess-Limerick, B. Kim, B. Schölkopf, B. Wulfe, B. Ichter, C. Lu, C. Xu, C. Le, C. Finn, C. Wang, C. Xu, C. Chi, C. Huang, C. Chan, C. Agia, C. Pan, C. Fu, C. Devin, D. Xu, D. Morton, D. Driess, D. Chen, D. Pathak, D. Shah, D. Büchler, D. Jayaraman, D. Kalashnikov, D. Sadigh, E. Johns, E. Foster, F. Liu, F. Ceola, F. Xia, F. Zhao, F. Stulp, G. Zhou, G. S. Sukhatme, G. Salhotra, G. Yan, G. Feng, G. Schiavi, G. Berseth, G. Kahn, G. Wang, H. Su, H.-S. Fang, H. Shi, H. Bao, H. B. Amor, H. I. Christensen, H. Furuta, H. Walke, H. Fang, H. Ha, I. Mordatch, I. Radosavovic, I. Leal, J. Liang, J. Abou-Chakra, J. Kim, J. Drake, J. Peters, J. Schneider, J. Hsu, J. Bohg, J. Bingham, J. Wu, J. Gao, J. Hu, J. Wu, J. Wu, J. Sun, J. Luo, J. Gu, J. Tan, J. Oh, J. Wu, J. Lu, J. Yang, J. Malik, J. Silvério, J. Hejna, J. Booher, J. Tompson, J. Yang, J. Salvador, J. J. Lim, J. Han, K. Wang, K. Rao, K. Pertsch, K. Hausman, K. Go, K. Gopalakrishnan, K. Goldberg, K. Byrne, K. Oslund, K. Kawaharazuka, K. Black, K. Lin, K. Zhang, K. Ehsani, K. Lekkala, K. Ellis, K. Rana, K. Srinivasan, K. Fang, K. P. Singh, K.-H. Zeng, K. Hatch, K. Hsu, L. Itti, L. Y. Chen, L. Pinto, L. Fei-Fei, L. Tan, L. J. Fan, L. Ott, L. Lee, L. Weihs, M. Chen, M. Lepert, M. Memmel, M. Tomizuka, M. Itkina, M. G. Castro, M. Spero, M. Du, M. Ahn, M. C. Yip, M. Zhang, M. Ding, M. Heo, M. K. Srirama, M. Sharma, M. J. Kim, N. Kanazawa, N. Hansen, N. Heess, N. J. Joshi, N. Suenderhauf, N. Liu, N. D. Palo, N. M. M. Shafiullah, O. Mees, O. Kroemer, O. Bastani, P. R. Sanketi, P. T. Miller, P. Yin, P. Wohlhart, P. Xu, P. D. Fagan, P. Mitrano, P. Sermanet, P. Abbeel, P. Sundaresan, Q. Chen, Q. Vuong, R. Rafailov, R. Tian, R. Doshi, R. Martín-Martín, R. Baijal, R. Scalise, R. Hendrix, R. Lin, R. Qian, R. Zhang, R. Mendonca, R. Shah, R. Hoque, R. Julian, S. Bustamante, S. Kirmani, S. Levine, S. Lin, S. Moore, S. Bahl, S. Dass, S. Sonawani, S. Song, S. Xu, S. Haldar, S. Karamcheti, S. Adebola, S. Guist, S. Nasiriany, S. Schaal, S. Welker, S. Tian, S. Ramamoorthy, S. Dasari, S. Belkhale, S. Park, S. Nair, S. Mirchandani, T. Osa, T. Gupta, T. Harada, T. Matsushima, T. Xiao, T. Kollar, T. Yu, T. Ding, T. Davchev, T. Z. Zhao, T. Armstrong, T. Darrell, T. Chung, V. Jain, V. Vanhoucke, W. Zhan, W. Zhou, W. Burgard, X. Chen, X. Wang, X. Zhu, X. Geng, X. Liu, X. Liangwei, X. Li, Y. Lu, Y. J. Ma, Y. Kim, Y. Chebotar, Y. Zhou, Y. Zhu, Y. Wu, Y. Xu, Y. Wang, Y. Bisk, Y. Cho, Y. Lee, Y. Cui, Y. Cao, Y.-H. Wu, Y. Tang, Y. Zhu, Y. Zhang, Y. Jiang, Y. Li, Y. Li, Y. Iwasawa, Y. Matsuo, Z. Ma, Z. Xu, Z. J. Cui, Z. Zhang, Z. Lin, "Open X-Embodiment: Robotic learning datasets and RT-X models" in *Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA)* (IEEE, 2023), pp. 6892–6903.
9. P. E. Dupont, N. Simaan, H. Choset, C. Rucker, Continuum robots for medical interventions. *Proc. IEEE* **110**, 847–870 (2022).
10. K. Vaesen, The cognitive bases of human tool use. *Behav. Brain Sci.* **35**, 203–218 (2012).
11. S. Atar, X. Liang, C. Joyce, F. Richter, W. Ricardo, C. Goldberg, P. Suresh, M. Yip, Humanoids in hospitals: A technical study of humanoid surrogates for dexterous medical interventions. arXiv: 2503.12725 (2025).

Acknowledgments

Funding: This work was supported by the National Science Foundation grant 2045803. **Competing interests:** The author declares that he has no competing interests.

10.1126/scirobotics.adt0684

Science Robotics

The robot will see you now: Foundation models are the path forward for autonomous robotic surgery

Michael Yip

Sci. Robot. **10** (104), eadt0684. DOI: 10.1126/scirobotics.adt0684

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