MANIPULATION

Within arm's reach: A path forward for robot dexterity

Sudharshan Suresh*

Visuotactile pretraining with human data leads to robust manipulation policies trained in simulation.

With record investment, hardware proliferation, and unprecedented data scaling, it is a hopeful time in robotics. We have an abundance of unstructured data capturing the human experience and mature tooling to collect structured robotics data. Slowly yet surely, we are beginning to see robots move out of research laboratories into our daily lives. However, dexterity remains a foundational problem and a critical bottleneck in building general physical intelligence. Moravec's paradox has become a worn cliché among roboticists who have witnessed chess engines beat grandmasters long before robots could pour them a glass of water. What makes physical manipulation such a hard problem for artificial intelligence, which has otherwise been successful in understanding images, languages, and semantic concepts? Several obstacles stand out: a lack of in-domain robot data, incomplete sensing, and the high dimensionality of the control problem.

Although parallel-jaw grippers arguably have untapped potential (*1*), both academia and industry have heavily invested in developing state-of-the-art dexterous hands (*2*, *3*). A seemingly simple task like in-hand rotation needs a finger gait that is precise, perceptive, and reactive. Toward this, reinforcement learning (RL) has exceeded the capabilities of model-based methods for robot control (*4*) while in the process eschewing modular perception and planning.

Often overlooked is the sense of touch, which gives robots direct knowledge of object interaction without occlusion or aliasing. There has been rapid growth in sensing—such as compact, performant touch sensors integrated into a robot's fingertips (*5*). Tactile feedback when combined with proprioception implicitly captures the location of contact and force applied during these interactions.

In their recent article, Ye *et al.* (*6*) present a method to collect, simulate, and learn from noisy tactile signals to achieve robust object manipulation. They propose a policy (Fig. 1) that combines vision and discrete touch to perform in-hand manipulation. Humans fuse these senses effortlessly—for example, we can deftly spin a pen in hand, which has incidentally become an illustrative task in robotics (*7*). With just a simple webcam and piezoresistive sensing, Ye *et al.* (*6*) demonstrate various tasks with the Shadow robot hand (*3*).

The first stage of their pipeline trains an encoder with human data collected in house that captures paired visuotactile events. The pretraining facilitates alignment between the modalities using data collected by participants wearing haptic gloves. A key finding is that the "attention maps" that map image regions contributing most toward the representation overlap with where tactile events occur. After this, through a combination of online imitation learning (IL) and RL, the authors trained a multitask policy conditioned on the encoding.

Previous works in dexterous manipulation (*4*, *8*) typically first trained an expert policy in simulation using "privileged" information (such as pose, velocity, and contact

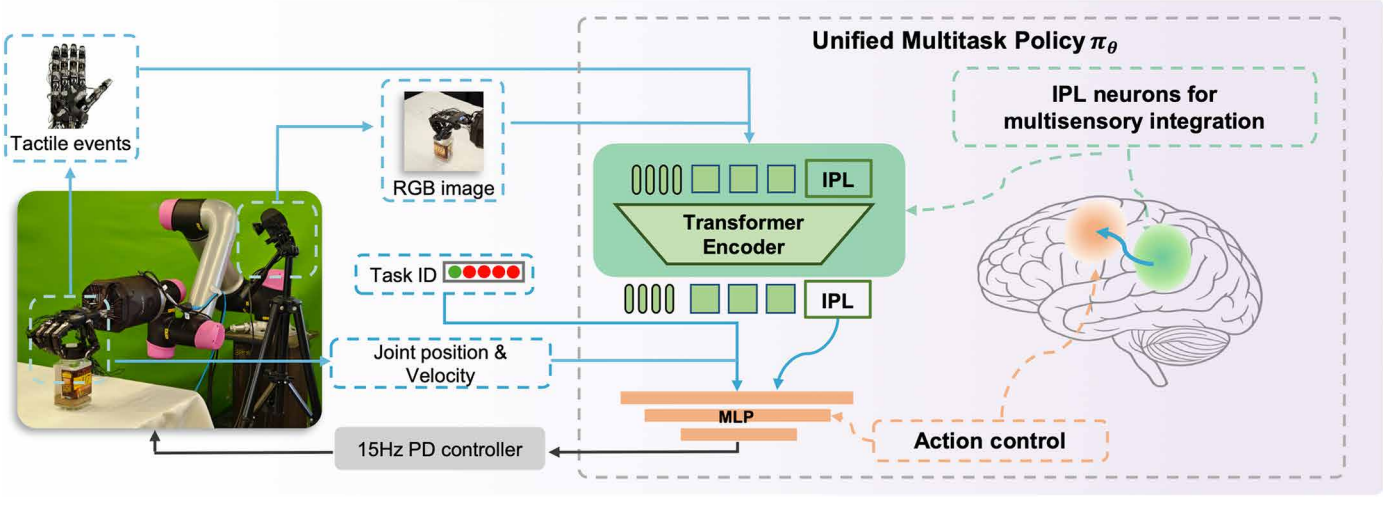

Fig. 1. Vision and tactile inputs are jointly encoded and passed to a multitask policy trained in simulation. RGB refers to red-green-blue channels, and IPL denotes the inferior parietal lobule.

CREDIT: ADAPTED FROM (*6*)

Boston Dynamics Inc., Waltham, MA, USA.
*Corresponding author. Email: suddhus@gmail.com

data). After this, the expert policy is distilled to a less-performant student policy that can work with realistic sensorimotor streams, such as vision, touch, and proprioception. Interestingly, Ye *et al.* (*6*) sidestepped training with privileged information and directly used their visuotactile encoder to train the expert policy. This shared representation across the student and expert empirically led to a boost in the student policy's performance. In practice, this resulted in high success rates for seen and unseen manipulation tasks across a diversity of objects. Examples of these are atomic chores we see in daily life, such as screw unfastening, pencil sharpening, and object reorientation. The "success" metric, common in policy evaluation, was tied to achieving an object goal configuration within a window of time. For all tasks, the fixed-base manipulator remained stationary, and the hand was actuated using a proportional-derivative (PD) controller.

In the article, the authors went one step further and performed A/B tests on different sensing resolutions and technologies, such as pressure-based versus piezoresistive touch sensing. This serves as a segue into several critical questions for practitioners in the field—what sensing technology should we adopt and converge toward? What robot surfaces (distal, proximal, and palm) are worth sensorizing? What resolution of touch is truly needed, and is binary contact sufficient for most tasks?

This work provides a good recipe for algorithmic ideas that will scale: effective sensor fusion, large-scale human data collection, and principled simulation techniques. However, there is much more to be done on its heels; an interesting data point is that although a multitask policy outperformed task-specific baselines, it did so by only a slight margin. For context, academic research in RL and IL has been making strides toward realizing multitask, cross-embodiment generalization (*9*). Another frontier is the ability to learn from unstructured human demonstrations, akin to an inquisitive child learning from a parent's daily activities. Last, to achieve more realistic tasks, robots require language following and semantic reasoning, as realized by modern-day vision-language-action (VLA) models (*10*). We have a long way to go, but the field is well poised to tackle these problems. Over the next decade, we may see a step change in the physical capabilities of these robots and, consequently, their adoption.

REFERENCES AND NOTES

1. Generalist AI Team, "GEN-0: Embodied foundation models that scale with physical interaction" Generalist AI Blog, 4 November 2025; https://generalistai.com/blog/nov-04-2025-GEN-0.
2. K. Shaw, A. Agarwal, D. Pathak, "LEAP Hand: Low-cost, efficient, and anthropomorphic hand for robot learning" in *Proceedings of Robotics: Science and Systems XIX*, K. Bekris, K. Hauser, S. Herbert, J. Yu, Eds. (RSS Foundation, 2023); 10.15607/RSS.2023.XIX.089.
3. D. Sharma, K. Tokas, A. Puri, K. Sharda, Shadow Hand. *J. Adv. Res. Appl. Sci.* **1**, 04–07 (2014).
4. T. Chen, M. Tippur, S. Wu, V. Kumar, E. Adelson, P. Agrawal, Visual dexterity: In-hand reorientation of novel and complex object shapes. *Sci. Robot.* **8**, eadc9244 (2023).
5. M. Lambeta, T. Wu, A. Sengul, V. R. Most, N. Black, K. Sawyer, R. Mercado, H. Qi, A. Sohn, B. Taylor, N. Tydingco, G. Kammerer, D. Stroud, J. Khatha, K. Jenkins, K. Most, N. Stein, R. Chavira, T. Craven-Bartle, E. Sanchez, Y. Ding, J. Malik, R. Calandra, Digitizing touch with an artificial multimodal fingertip. arXiv:2411.02479 [cs.RO] (2024).
6. Q. Ye, Q. Liu, S. Wang, J. Chen, Y. Cui, K. Jin, H. Chen, X. Cai, G. Li, J. Chen, Visual-tactile pretraining and online multitask learning for human-like manipulation dexterity. *Sci. Robot.* **11**, eady2869 (2026).
7. J. Wang, Y. Yuan, H. Che, H. Qi, Y. Ma, J. Malik, X. Wang, "Lessons from learning to spin 'pens'" in *Proceedings of the 8th Conference on Robot Learning*, P. Agrawal, O. Kroemer, W. Burgard, Eds. (PMLR, 2024), vol. 270, pp. 3124–3138.
8. H. Qi, B. Yi, S. Suresh, M. Lambeta, Y. Ma, R. Calandra, J. Malik, "General in-hand object rotation with vision and touch" in *Proceedings of the 7th Conference on Robot Learning*, J. Tan, M. Toussaint, K. Darvish, Eds. (PMLR, 2023), vol. 229, pp. 2549–2564.
9. M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, Q. Vuong, T. Kollar, B. Burchfiel, R. Tedrake, D. Sadigh, S. Levine, P. Liang, C. Finn, "OpenVLA: An open-source vision-language-action model" in *Proceedings of the 8th Conference on Robot Learning*, P. Agrawal, O. Kroemer, W. Burgard, Eds. (PMLR, 2024), vol. 270, pp. 2679–2713.
10. K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, M. Y. Galliker, D. Ghosh, L. Groom, K. Hausman, B. Ichter, S. Jakubczak, T. Jones, L. Ke, D. LeBlanc, S. Levine, A. Li-Bell, M. Mothukuri, S. Nair, K. Pertsch, A. Z. Ren, L. X. Shi, L. Smith, J. T. Springenberg, K. Stachowicz, J. Tanner, Q. Vuong, H. Walke, A. Walling, H. Wang, L. Yu, U. Zhilinsky, "$\pi_{0.5}$: A vision-language-action model with open-world generalization" in *Proceedings of the 9th Conference on Robot Learning*, J. Lim, S. Song, H. Park, Eds. (PMLR, 2025), vol. 305, pp. 17–40.

Acknowledgments

Competing interests: S.S. acknowledges employment at Boston Dynamics Inc.; this article was written in a personal capacity. Any opinions expressed here are those of the author and do not reflect their employer's.

10.1126/scirobotics.aee5782

Science Robotics

Within arm's reach: A path forward for robot dexterity

Sudharshan Suresh

Sci. Robot. **11** (110), eaee5782. DOI: 10.1126/scirobotics.aee5782

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.