Learning Precise, Contact-Rich Manipulation through Uncalibrated Tactile Skins

Venkatesh Pattabiraman[1,*] Yifeng Cao[2] Siddhant Haldar[1] Lerrel Pinto[1] Raunaq Bhirangi[1,3,*,†]

[1] New York University [2] Columbia University [3] Carnegie Mellon University

* equal contribution

https://visuoskin.github.io/

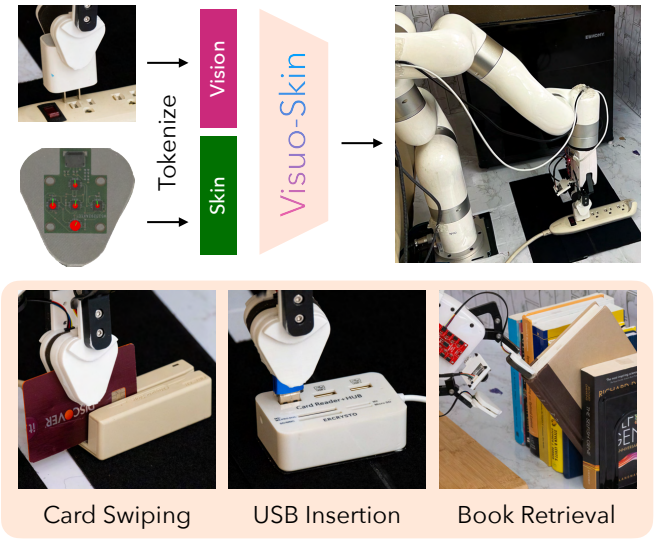

Fig. 1: ViSk uses AnySkin with a simple transformer-based architecture to solve precise, contact-rich tasks.

I. Introduction

Humans effortlessly perform precise manipulation tasks in their everyday lives, such as plugging in charger cords, or swiping credit cards – activities that demand exact alignment and involve constrained motion. While the role of tactile feedback for robust execution of precise skills in humans is widely acknowledged [1, 2], analogous capabilities in robotic policies have lagged behind their vision-based counterparts.

In this work, we present Visuo-Skin (ViSk), a simple framework for training precise robot policies using skin-based tactile sensing. ViSk uses a simple visuotactile policy architecture that incorporates tactile signals from AnySkin [3], an affordable magnetic tactile sensor demonstrated to provide spatially continuous, low-dimensional (15-dimensional) sensing while being replaceable, making it well-suited for policy learning applications. The ViSk policy builds upon the BAKU [4] architecture, which enables policy learning across multiple camera views and tasks. Through ViSk, we demonstrate that simply incorporating a tactile token obtained from a tactile encoder into state-of-the-art visual policy learning architectures enables effective visuotactile policy learning for precise real-world manipulation tasks that require visual as well as tactile inputs for localization. Furthermore, using a low-dimensional sensor like AnySkin allows policies to be learned end-to-end without requiring any task-specific preprocessing [5, 6] of the tactile input or pretraining [7, 8]. To the best of our knowledge, this work presents the first visuotactile framework enabling robots to perform precise contact-rich manipulation skills with policies that generalize across spatial variations while requiring a small number of robot demonstrations (< 200).

To demonstrate the effectiveness of ViSk, we run extensive experiments on four precise manipulation tasks using a real-world xArm robot - *plug insertion*, *card swiping*, *USB insertion*, and *bookshelf retrieval*. Our main findings are:

1) Policies trained with ViSk using skin-based tactile sensing exhibit an overall 27.5% absolute improvement in performance compared to vision-only models across 4 precise manipulation tasks (Section III-A).
2) Policies trained with the AnySkin tactile sensor [3] outperform those using optical tactile sensors such as DIGIT [7] by at least 43% on two real-world tasks, highlighting the benefits of skin-based sensors for visuotactile policy learning (Section III-B).

Videos and code for training and evaluation are available at https://visuoskin.github.io/.

II. Visuo-Skin Policy Learning (ViSk)

ViSk employs AnySkin [3], a skin-based magnetic tactile sensor shown to yield consistent tactile measurements reliably under various conditions. It builds upon state-of-the-art approaches to visual policy learning [4] by incorporating a tactile encoding stream, allowing the network to profitably learn from multimodal visuotactile data. Below, we describe the components of our method.

A. Data Collection

We use a VR-based teleoperation framework [9] employing the Meta Quest 3 headset to collect xArm data for our real world experiments. Drawing from prior work demonstrating the benefits of adding noise to demonstrations

† Correspondence to: raunaqbhirangi@nyu.edu

TABLE I: Success rates (out of 10) averaged over three seeds for policies trained on four tasks: Plug Insertion, USB Insertion, Card Swiping and Book Retrieval. ViSk policies are highlighted in grey.

Tactile Sensor	Input Modalities			Policy performance			
	3rd Person Camera	Wrist Cameras	Robot Proprio	Plug Insertion	USB Insertion	Card Swiping	Book Retrieval
None	✓	✗	✗	0.0 ± 0.0	0.7 ± 0.6	3.3 ± 1.6	2.0 ± 1.0
	✓	✗	✓	0.0 ± 0.0	0.0 ± 0.0	3.0 ± 1.0	0.6 ± 0.5
	✓	✓	✗	3.6 ± 0.5	2.3 ± 2.0	1.3 ± 0.5	3.3 ± 1.1
	✓	✓	✓	1.0 ± 1.0	2.0 ± 1.0	3.0 ± 1.7	2.3 ± 1.5
AnySkin (ViSk)	✓	✗	✗	2.3 ± 1.1	2.0 ± 1.0	$\mathbf{7.0 \pm 1.7}$	3.6 ± 2.5
	✓	✗	✓	1.3 ± 0.5	1.0 ± 1.0	2.6 ± 1.5	2.6 ± 0.5
	✓	✓	✗	$\mathbf{6.6 \pm 1.5}$	$\mathbf{5.6 \pm 1.5}$	1.0 ± 1.0	$\mathbf{5.3 \pm 2.0}$
	✓	✓	✓	3.6 ± 1.5	2.0 ± 1.0	3.0 ± 1.7	4.6 ± 2.0
DIGIT	✓	✗	✗	2.3 ± 0.5	0.0 ± 0.0	N/A	N/A
	✓	✓	✗	1.6 ± 1.5	0.3 ± 0.5	N/A	N/A

for policy learning [10, 11], we add a uniformly sampled angular perturbation to the direction of the commanded robot velocity during teleoperation, increasing the diversity of contact-rich signals in the dataset by rendering the tasks slightly more challenging for the human operator.

B. Policy Architecture

The ViSk policy builds on top of BAKU [4], a state-of-the-art transformer-based policy learning architecture that learns visual policies across multiple camera views. We encode the visual inputs using a modified ResNet-18 [12] visual encoder. Low-dimensional tactile inputs from the AnySkin sensor are encoded with a two-layer multilayer perceptron (MLP). The encoded representations for each modality are projected to the same dimensionality to facilitate combining modalities in the observation trunk. Some of the comparisons in Section III use DIGIT sensors and robot proprioception as inputs to the policy. In line with prior works [13, 14], tactile images from the DIGIT sensor are encoded using the same ResNet-18 encoder as the visual data. The encoded inputs from all modalities along with a learnable action token are passed through a transformer decoder network [15]. A deterministic action head is used to predict the action from the action feature. We follow prior work [4, 16, 17] and include action chunking and exponential temporal smoothing [16] to counteract the covariate shift often seen in the low-data imitation learning regime.

III. Experiments

We study the effectiveness of the ViSk framework in a policy learning setting using behavior cloning. Our experiments are designed to answer the following questions:

- How does ViSk perform on precise manipulation tasks?
- Does ViSk's use of AnySkin improve over DIGIT [7]?

A. Performance of ViSk policies

We evaluate the performance of ViSk policies on the aforementioned precise manipulation tasks in real world. For each evaluation, we train policies across 3 random seeds and conduct 10 trials per seed for 30 trials. We report aggregated success rate across seeds in Table I, and find that ViSk policies consistently outperform variations across tasks.

Additionally, we observe that ViSk policies exhibit emergent seeking behavior. For plug and USB insertion, and card swiping, we find that the policy first gets close to the location of the target, makes contact, and proceeds to move around as it tries to find the target. Similarly, for the book retrieval task, ViSk policies apply a controlled downward force that enables them to pivot the book to an appropriate tilt, followed by grasping and retrieval.

B. Comparison between AnySkin and DIGIT

To further demonstrate the effectiveness of AnySkin for precise manipulation tasks, we collect demonstration datasets for two tasks (Plug Insertion and USB Insertion) using DIGIT sensors instead of AnySkin sensors. We keep the same policy architecture, except for the tactile encoder, where we replace the MLP with a modified ResNet-18 encoder. We ensure the DIGIT and AnySkin datasets are closely aligned, maintaining the same test positions. The results in Table I compare ViSk using the skin-based AnySkin sensor with the optical DIGIT [7] sensor. Our findings show that policies trained with AnySkin significantly outperform those trained with DIGIT. This difference arises from DIGIT's lower sensitivity, which hinders detection of small tactile signals from contact with objects.

IV. Conclusions

In this work, we presented Visuo-Skin (ViSk), a simple yet effective framework that leverages low-dimensional skin-based tactile sensing for visuotactile policy learning in the real world. Our results demonstrate the efficacy of ViSk across a diverse range of precise, contact-rich manipulation tasks. The overall performance (65%) suggests potential for further enhancement through fine-tuning the ViSk policy using reinforcement learning techniques. Further, the unexpected result of robot proprioception not improving performance warrants further investigation and presents an interesting direction for future research. We believe that ViSk presents a signficant step in the right direction for advancing visuotactile policy learning in robotics.

References

[1] R. S. Johansson, "Sensory control of dexterous manipulation in humans," in *Hand and brain*. Elsevier, 1996, pp. 381–414.

[2] J. Jenner and J. Stephens, "Cutaneous reflex responses and their central nervous pathways studied in man," *The Journal of physiology*, vol. 333, no. 1, pp. 405–419, 1982.

[3] R. Bhirangi, V. Pattabiraman, E. Erciyes, Y. Cao, T. Hellebrekers, and L. Pinto, "Anyskin: Plug-and-play skin sensing for robotic touch," *arXiv preprint arXiv:2409.08276*, 2024.

[4] S. Haldar, Z. Peng, and L. Pinto, "Baku: An efficient transformer for multi-task policy learning," 2024. [Online]. Available: https://arxiv.org/abs/2406.07539

[5] R. Li, R. Platt, W. Yuan, A. Ten Pas, N. Roscup, M. A. Srinivasan, and E. Adelson, "Localization and manipulation of small parts using gelsight tactile sensing," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2014, pp. 3988–3993.

[6] S. Kim and A. Rodriguez, "Active extrinsic contact sensing: Application to general peg-in-hole insertion," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 10 241–10 247.

[7] M. Lambeta, P.-W. Chou, S. Tian, B. Yang, B. Maloon, V. R. Most, D. Stroud, R. Santos, A. Byagowi, G. Kammerer, *et al.*, "Digit: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 3838–3845, 2020.

[8] A. George, S. Gano, P. Katragadda, and A. B. Farimani, "Visuo-tactile pretraining for cable plugging," *arXiv preprint arXiv:2403.11898*, 2024.

[9] A. Iyer, Z. Peng, Y. Dai, I. Guzey, S. Haldar, S. Chintala, and L. Pinto, "Open teach: A versatile teleoperation system for robotic manipulation," *arXiv preprint arXiv:2403.07870*, 2024.

[10] D. Brandfonbrener, S. Tu, A. Singh, S. Welker, C. Boodoo, N. Matni, and J. Varley, "Visual backtracking teleoperation: A data collection protocol for offline image-based reinforcement learning," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 11 336–11 342.

[11] S. Dasari, J. Wang, J. Hong, S. Bahl, Y. Lin, A. Wang, A. Thankaraj, K. Chahal, B. Calli, S. Gupta, *et al.*, "Rb2: Robotic manipulation benchmarking with a twist," *arXiv preprint arXiv:2203.08098*, 2022.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[13] J. Lin, R. Calandra, and S. Levine, "Learning to identify object instances by touch: Tactile recognition via multimodal matching," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 3644–3650.

[14] Y. Li, J.-Y. Zhu, R. Tedrake, and A. Torralba, "Connecting touch and vision via cross-modal prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 609–10 618.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[16] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," *arXiv preprint arXiv:2304.13705*, 2023.

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