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/

***Abstract*— While visuomotor policy learning has advanced robotic manipulation, precisely executing contact-rich tasks remains challenging due to the limitations of vision in reasoning about physical interactions. To address this, recent work has sought to integrate tactile sensing into policy learning. However, many existing approaches rely on optical tactile sensors that are either restricted to recognition tasks or require complex dimensionality reduction steps for policy learning. In this work, we explore learning policies with magnetic skin sensors, which are inherently low-dimensional, highly sensitive, and inexpensive to integrate with robotic platforms. To leverage these sensors effectively, we present the VISK framework, a simple approach that uses a transformer-based policy and treats skin sensor data as additional tokens alongside visual information. Evaluated on four complex real-world tasks involving credit card swiping, plug insertion, USB insertion, and bookshelf retrieval, VISK significantly outperforms both vision-only and optical tactile sensing based policies. Further analysis reveals that combining tactile and visual modalities enhances policy performance and spatial generalization, achieving an average improvement of 27.5% across 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. These tasks are so commonplace that we often overlook the complexity involved in executing them with the necessary accuracy. In contrast, much of the existing robot learning literature remains focused on simple, low-precision primitives such as pick-and-place, slide, push-pull, and lift that does not require such fine-grained spatial accuracy. As we strive to create robots capable of everyday tasks like handling cables and opening jars, it is crucial to develop frameworks that enable precise, contact-rich manipulation.

While the role of tactile feedback for robust execution of precise skills in humans is widely acknowledged [1], analogous capabilities in robotic policies have lagged behind their vision-based counterparts. A variety of tactile sensors have been developed to bridge this gap in robotics, with optical tactile sensors like Gelsight [2] and DIGIT [3] becoming popular choices in robot learning due to their high resolution. This increased resolution has facilitated several impressive works in areas like 3D reconstruction and localization [4] and object recognition [5, 6]. However, the high dimensionality of tactile data from such sensors introduces additional complexity to the already challenging problem of policy learning. This observation prompts an investigation into using alternative tactile sensing modalities that naturally offer lower-dimensional representations while still effectively capturing the essential characteristics of physical contact.

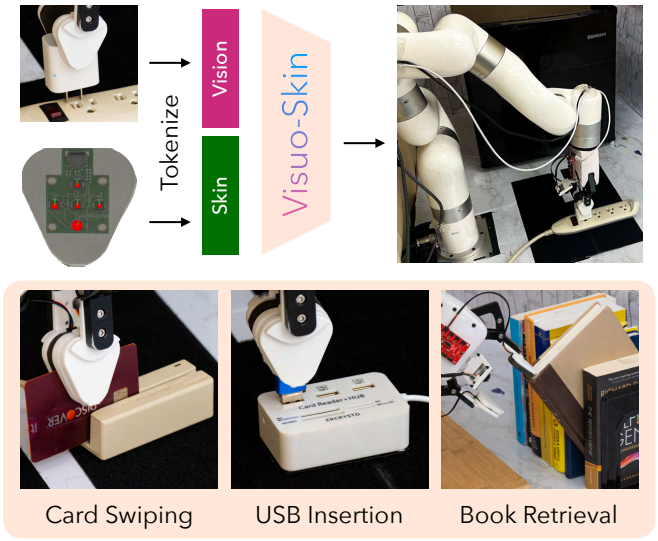

Fig. 1: VISK uses AnySkin with a simple transformer-based architecture to solve precise, contact-rich tasks.

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 [7], 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 [8] 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 visuotac-

† Correspondence to: raunaqbhirangi@nyu.edu

tile policy learning for precise real-world manipulation tasks that require visual as well as tactile inputs for localization. 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*, *credit card swiping*, *USB insertion*, and *bookshelf retrieval*. Our main findings are summarized below:

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 [7] outperform those using optical tactile sensors such as DIGIT [3] by at least 43% on two real-world tasks, highlighting the benefits of skin-based sensors for visuotactile policy learning (Section III-B).

All of our datasets, code for training, and robot evaluation will be made publicly available. Robot videos are best viewed at https://visuoskin.github.io/.

II. Visuo-Skin Policy Learning (ViSk)

Two key considerations in designing a framework for visuotactile policy learning include the choice of a tactile sensor capable of providing reliable tactile data across diverse environments and tasks, and designing a neural architecture able to effectively leverage multimodal visual and tactile information. Our proposed approach, ViSk, addresses these in two ways. First, it employs AnySkin [7], a skin-based magnetic tactile sensor shown to yield consistent tactile measurements reliably under various conditions. Second, it builds upon state-of-the-art approaches to visual policy learning [8] by incorporating a tactile encoding stream, allowing the network to profitably learn from multimodal visuotactile data. Below, we describe each component of our method in detail.

A. Data Collection

We use a VR-based teleoperation framework [10] employing the Meta Quest 3 headset to collect data for our real-world xArm robot experiments. Visual data from 4 camera views, including an egocentric camera attached to the robot gripper, is recorded at 30 Hz. Tactile data for the AnySkin experiments is recorded as magnetometer signals at 100 Hz, while data from the DIGIT sensors in comparative tests are recorded at 30 Hz, identical to the cameras. Drawing from prior work demonstrating the benefits of adding noise to demonstrations for policy learning [11], we add a uniformly sampled angular perturbation to the direction of the commanded robot velocity during teleoperation. This proves especially useful for 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 [8], a state-of-the-art transformer-based policy learning architecture that learns visual policies across multiple camera views (Figure 2). We encode the visual inputs from cameras using using a modified ResNet-18 [9] 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 work [12], 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 [13]. A deterministic action head is used to predict the action from the action feature. We follow prior work [8, 14, 15] and include action chunking and exponential temporal smoothing [14] 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 [3]?

A. Performance of ViSk policies

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

Additionally, we observe that ViSk policies exhibit emergent seeking behavior. For instance, with the plug insertion and USB insertion tasks, we find that the policy first gets close to the location of the target (socket or port respectively), makes contact, and proceeds to move around as it tries to find the target. This behavior is strong evidence of ViSk policies effectively leveraging tactile information from AnySkin. Further, it is distinctly different from the behavior of vision-only policies that simply attempt to push downwards once close to the insertion location regardless of alignment.

Similarly, for the book retrieval task, policies without AnySkin either apply too little force causing the book to flip back into the bookrack, or too much force causing the book to topple over entirely. ViSk policies apply a controlled downward force that enables them to pivot the book to an appropriate tilt, followed by grasping and retrieval. Further, for the book retrieval task, repeated interaction with the sharp edges of the book caused the AnySkin to tear. All evaluations

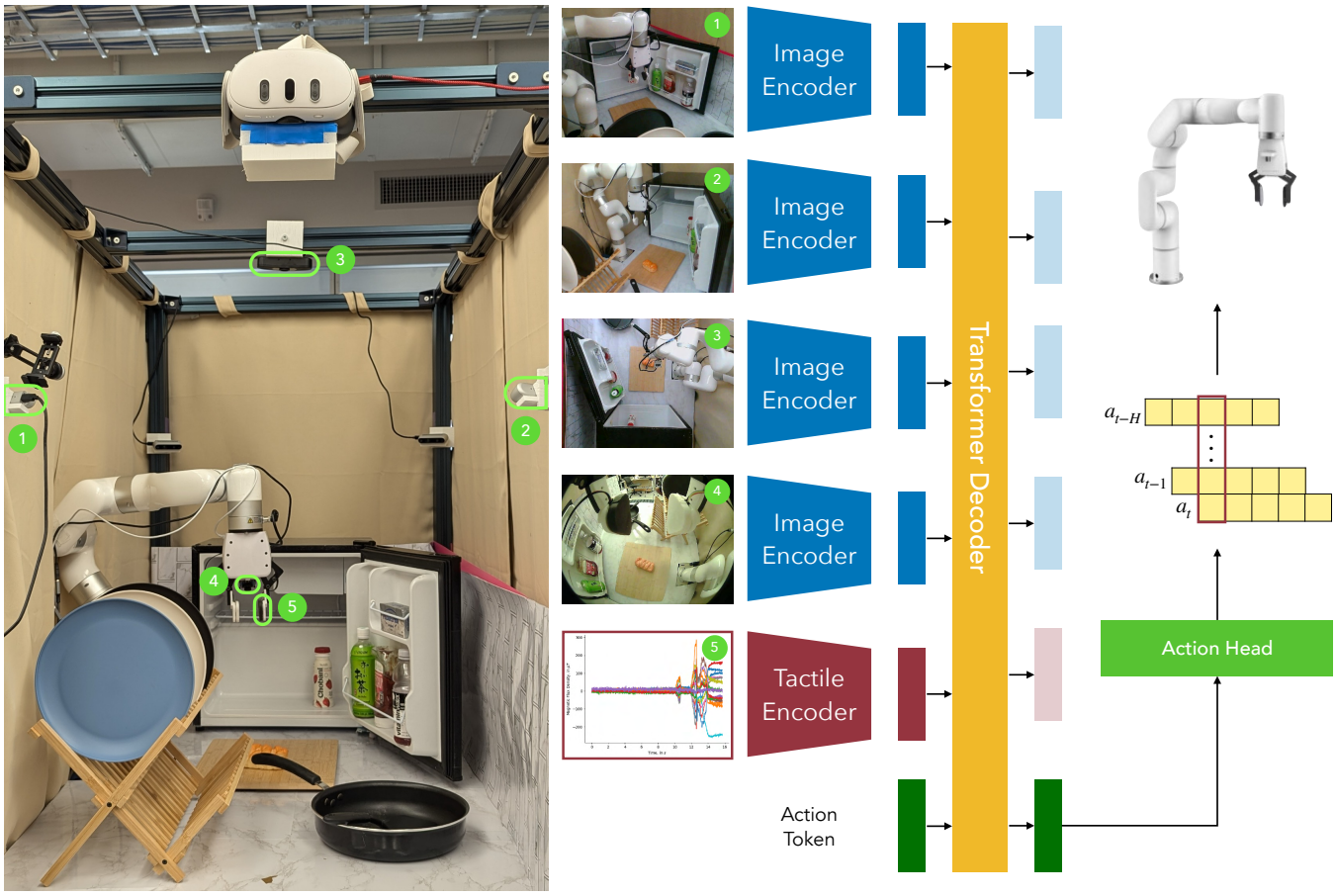

Fig. 2: (left) Robot setup used for experiments in Section III; (right) VISK policy architecture uses ResNet-18 [9] encoders for camera inputs and an MLP encoder for AnySkin input. An action token is appended to the encoded inputs before passing them through a transformer decoder, and the corresponding feature is used for action prediction by the action head.

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 this task reported in Table I use a new instance of AnySkin. The sustained improvement of ViSk policies with new skins underscores the importance of using AnySkin to the ViSk framework.

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 [3] 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 the object. Additionally, the higher dimensionality of DIGIT observations may complicate learning a sensory encoder without overfitting. These experiments underscore the superiority of AnySkin over optical sensors for visuotactile policy learning in precise tasks.

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. We address a few limitations in this work: (a) While ViSk shows significant improvements over vision-only policies, the policy's performance remains at approximately 60% across all tasks. This suggests potential for further enhancement through fine-tuning the ViSk policy using reinforcement learning techniques. (b) Contrary to findings in prior studies, we observe that robot proprioception did not contribute to improved policy learning performance in precise manipulation tasks. This unexpected result warrants further investigation and presents an interesting direction for future research. These limitations notwithstanding, 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] W. Yuan, S. Dong, and E. H. Adelson, "Gelsight: High-resolution robot tactile sensors for estimating geometry and force," *Sensors*, vol. 17, no. 12, p. 2762, 2017.

[3] 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.

[4] S. Suresh, Z. Si, S. Anderson, M. Kaess, and M. Mukadam, "Midastouch: Monte-carlo inference over distributions across sliding touch," in *Conference on Robot Learning*. PMLR, 2023, pp. 319–331.

[5] S. Funabashi, G. Yan, A. Geier, A. Schmitz, T. Ogata, and S. Sugano, "Morphology-specific convolutional neural networks for tactile object recognition with a multi-fingered hand," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 57–63.

[6] R. Bhirangi, A. DeFranco, J. Adkins, C. Majidi, A. Gupta, T. Hellebrekers, and V. Kumar, "All the feels: A dexterous hand with large-area tactile sensing," *IEEE Robotics and Automation Letters*, 2023.

[7] 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.

[8] 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

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.