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, [†]Correspondence to: raunaqbhirangi@nyu.edu

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.

1 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.

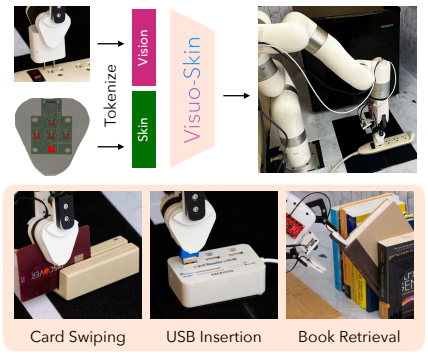

Figure 1: VISK uses AnySkin with a simple transformer-based architecture to solve precise, contact-rich tasks.

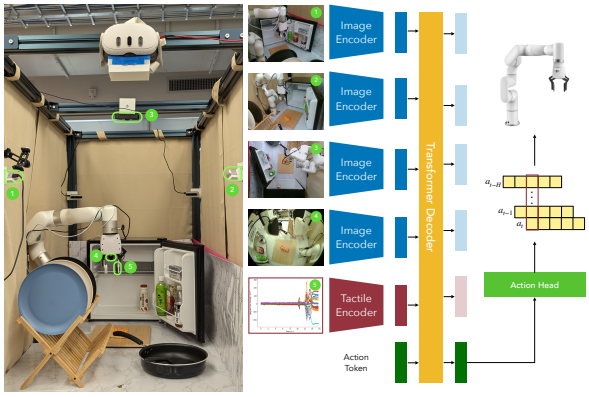

Figure 2: Robot setup used for experiments **(left)** and VISK policy architecture **(right)**.

8th Conference on Robot Learning (CoRL 2024), Munich, Germany.

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. A variety of tactile sensors have been developed to bridge this gap in robotics, with optical tactile sensors like Gelsight [3] and DIGIT [4] 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 [5, 6] and object recognition [7, 8]. In most cases, the use of optical sensors necessitates dimensionality reduction through representation learning [4], explicit state estimation [9, 10] or discretization [11, 12] to make it amenable to 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.

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 [13], 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 [14] 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 [9, 10] of the tactile input or pretraining [4, 12]. 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 3.1).
2. Policies trained with the AnySkin tactile sensor [13] outperform those using optical tactile sensors such as DIGIT [4] by at least 43% on two real-world tasks, highlighting the benefits of skin-based sensors for visuotactile policy learning (Section 3.2).
3. Through an ablation analysis, we study the impact of different modalities on policy learning, particularly the difference between visual and visuotactile policies for precise manipulation (Section C.3).

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/`.

2 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 [13], 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 [14] 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.

2.1 Data Collection

We use a VR-based teleoperation framework [15] 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 [16, 17], 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.

2.2 Policy Architecture

The VISK policy builds on top of BAKU [14], 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 [18] 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 3 use DIGIT sensors and robot proprioception as inputs to the policy. In line with prior works [19, 20], 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 [21]. A deterministic action head is used to predict the action from the action feature. We follow prior work [14, 22, 23] and include action chunking and exponential temporal smooting [22] to counteract the covariate shift often seen in the low-data imitation learning regime. More details about the policy architecture have been provided in Appendix B.1.

3 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 [4]?

Additional analysis about the effect of different input combinations on VISK, the generalization capabilities of VISK, and the environment setup and task descriptions have been provided in Appendix C.

3.1 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 1, 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,

Table 1: 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

followed by grasping and retrieval as shown in Fig. B.1. Further, for the book retrieval task, repeated interaction with the sharp edges of the book caused the AnySkin to tear. All evaluations for this task reported in Table 1 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.

3.2 Comparison between AnySkin and DIGIT

To further demonstrate the effectiveness of AnySkin for precise manipulation tasks, we collect demonstration datasets for two tasks from Section C.2 (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 1 compare ViSk using the skin-based AnySkin sensor with the optical DIGIT [4] 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.

4 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.

Acknowledgments

Special thanks for Krishna Bodduluri, Mike Lambeta, and team from Meta AI Research for providing the DIGIT sensors for comparison. Thanks to Tess Hellebrekers for providing the sensor skins and discussions for the experiments. This work was supported by grants from Honda, Hyundai, NSF award 2339096 and ONR awards N00014-21-1-2758 and N00014-22-1-2773. LP is supported by the Packard Fellowship.

References

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

[2] J. Jenner and J. Stephens. Cutaneous reflex responses and their central nervous pathways studied in man. *The Journal of physiology*, 333(1):405–419, 1982.

[3] W. Yuan, S. Dong, and E. H. Adelson. Gelsight: High-resolution robot tactile sensors for estimating geometry and force. *Sensors*, 17(12):2762, 2017.

[4] 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*, 5(3):3838–3845, 2020.

[5] S. Suresh, M. Bauza, K.-T. Yu, J. G. Mangelson, A. Rodriguez, and M. Kaess. Tactile slam: Real-time inference of shape and pose from planar pushing. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11322–11328. IEEE, 2021.

[6] 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*, pages 319–331. PMLR, 2023.

[7] 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)*, pages 57–63. IEEE, 2019.

[8] 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.

[9] 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*, pages 3988–3993. IEEE, 2014.

[10] 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)*, pages 10241–10247. IEEE, 2022.

[11] H. Qi, B. Yi, S. Suresh, M. Lambeta, Y. Ma, R. Calandra, and J. Malik. General in-hand object rotation with vision and touch. In *Conference on Robot Learning*, pages 2549–2564. PMLR, 2023.

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

[13] 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.

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

[15] 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.

[16] 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)*, pages 11336–11342. IEEE, 2023.

[17] 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.

[18] 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*, pages 770–778, 2016.

[19] 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)*, pages 3644–3650. IEEE, 2019.

[20] 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*, pages 10609–10618, 2019.

[21] 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*, 30, 2017.

[22] 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.

[23] 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.

[24] S. Sundaram, P. Kellnhofer, Y. Li, J.-Y. Zhu, A. Torralba, and W. Matusik. Learning the signatures of the human grasp using a scalable tactile glove. *Nature*, 569(7758):698–702, 2019.

[25] T. Bhattacharjee, A. Jain, S. Vaish, M. D. Killpack, and C. C. Kemp. Tactile sensing over articulated joints with stretchable sensors. In *2013 World Haptics Conference (WHC)*, pages 103–108. IEEE, 2013.

[26] S. Stassi, V. Cauda, G. Canavese, and C. F. Pirri. Flexible tactile sensing based on piezoresistive composites: A review. *Sensors*, 14(3):5296–5332, 2014.

[27] O. Glauser, D. Panozzo, O. Hilliges, and O. Sorkine-Hornung. Deformation capture via soft and stretchable sensor arrays. *ACM Transactions on Graphics (TOG)*, 38(2):1–16, 2019.

[28] T.-Y. Wu, L. Tan, Y. Zhang, T. Seyed, and X.-D. Yang. Capacitivo: Contact-based object recognition on interactive fabrics using capacitive sensing. In *Proceedings of the 33rd annual acm symposium on user interface software and technology*, pages 649–661, 2020.

[29] N. Wettels, V. J. Santos, R. S. Johansson, and G. E. Loeb. Biomimetic tactile sensor array. *Advanced robotics*, 22(8):829–849, 2008.

[30] T. P. Tomo, M. Regoli, A. Schmitz, L. Natale, H. Kristanto, S. Somlor, L. Jamone, G. Metta, and S. Sugano. A new silicone structure for uskin—a soft, distributed, digital 3-axis skin sensor and its integration on the humanoid robot icub. *IEEE Robotics and Automation Letters*, 3(3): 2584–2591, 2018.

[31] R. Bhirangi, T. Hellebrekers, C. Majidi, and A. Gupta. Reskin: versatile, replaceable, lasting tactile skins. In *5th Annual Conference on Robot Learning*, 2021.

[32] S. Suresh, H. Qi, T. Wu, T. Fan, L. Pineda, M. Lambeta, J. Malik, M. Kalakrishnan, R. Calandra, M. Kaess, et al. Neural feels with neural fields: Visuo-tactile perception for in-hand manipulation. *arXiv preprint arXiv:2312.13469*, 2023.

[33] J. Li, S. Dong, and E. Adelson. Slip detection with combined tactile and visual information. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7772–7777. IEEE, 2018.

[34] N. Jamali and C. Sammut. Majority voting: Material classification by tactile sensing using surface texture. *IEEE Transactions on Robotics*, 27(3):508–521, 2011. doi:10.1109/TRO.2011.2127110.

[35] J. Ilonen, J. Bohg, and V. Kyrki. Fusing visual and tactile sensing for 3-d object reconstruction while grasping. In *2013 IEEE International Conference on Robotics and Automation*, pages 3547–3554, 2013. doi:10.1109/ICRA.2013.6631074.

[36] I. Guzey, B. Evans, S. Chintala, and L. Pinto. Dexterity from touch: Self-supervised pre-training of tactile representations with robotic play. *arXiv preprint arXiv:2303.12076*, 2023.

[37] I. Guzey, Y. Dai, B. Evans, S. Chintala, and L. Pinto. See to touch: Learning tactile dexterity through visual incentives. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13825–13832. IEEE, 2024.

[38] T. Lin, Y. Zhang, Q. Li, H. Qi, B. Yi, S. Levine, and J. Malik. Learning visuotactile skills with two multifingered hands. *arXiv preprint arXiv:2404.16823*, 2024.

[39] Y. Yuan, H. Che, Y. Qin, B. Huang, Z.-H. Yin, K.-W. Lee, Y. Wu, S.-C. Lim, and X. Wang. Robot synesthesia: In-hand manipulation with visuotactile sensing. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6558–6565. IEEE, 2024.

[40] T. Kelestemur, R. Platt, and T. Padir. Tactile pose estimation and policy learning for unknown object manipulation. *arXiv preprint arXiv:2203.10685*, 2022.

[41] M. A. Lee, Y. Zhu, P. Zachares, M. Tan, K. Srinivasan, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg. Making sense of vision and touch: Learning multimodal representations for contact-rich tasks. *IEEE Transactions on Robotics*, 36(3):582–596, 2020.

[42] H. Li, Y. Zhang, J. Zhu, S. Wang, M. A. Lee, H. Xu, E. Adelson, L. Fei-Fei, R. Gao, and J. Wu. See, hear, and feel: Smart sensory fusion for robotic manipulation. *arXiv preprint arXiv:2212.03858*, 2022.

A Related Work

A.1 Tactile sensing in Robotics

Most robotic tasks involve physical interaction with the environment. Tactile sensing is critical in its ability to enable robots to reason about the physics of contact directly at the point of contact. Over the years, a number of diverse transduction mechanisms have been explored for tactile sensing. Resistive tactile sensors [24, 25, 26] are inexpensive and relatively easy to fabricate, and provide discrete sensing making them well-suited for a range of applications that involve sensing the presence or absence of contact. Capacitative tactile sensors [27, 28] tend to provide more fine-grained measurements compared to resistive sensors and include proximity sensing in addtion to tactile sensing. Another versatile category of sensors are MEMS-based sensors [29] that often combine multiple sensors such as audio and IMU sensors and can offer multimodal feedback in addition to higher resolution and mm-scale form factor.

Recently, optical tactile sensors like Gelsight [3] and DIGIT [4] have emerged as a popular, high resolution alternative to existing tactile sensors for robotics due to a number of desirable properties such as their ease of replaceability and compatibility with well-understood neural architectures like convolutional neural networks [7]. Similarly, magnetic tactile sensors like Xela [30] and ReSkin [31] have garnered significant interest due to their scalable form factor, low dimensionality and ability to sense shear force in addition to their consistency across sensor instances [31, 32]. In light of these characteristics, the VISK framework presented in this work uses AnySkin [13] a magnetic tactile sensor that strikes the right balance between low dimensionality and continuous contact sensing. Furthermore, its superior cross-instance signal consistency makes it more amenable than optical sensors to policy learning without the need for complex additional fabrication to prevent wear and tear [12].

A.2 Visuotactile learning

The meteoric rise of deep learning has paralleled recent developments in rapid prototyping and additive manufacturing. As a result, a number of recent works have investigated the use of machine learning for a host of tactile prediction tasks such as slip detection [33], material classification [34], object identification [19] and 3D reconstruction [35] across a range of tactile sensors. In this paper, we specifically focus on policy learning – incorporating tactile information into robotic policies to enhance contact-rich manipulation.

Recent works have demonstrated impressive improvements from incorporating tactile data into the policy learning framework for precise dexterity [36, 37] and bimanual manipulation [38]. However, the high dimensional nature of dexterous control limits the task complexity and extent of generalizability enabled by these works. While [11, 39] use sim2real learning to demonstrate significant generalizability across objects for an in-hand rotation task, the task lacks precision, and sim2real transfer necessitates significant dilution of the tactile input to only capture coarse, discrete information. This limits the scalability of this approach to the precise, contact-rich tasks considered in this work.

Yet other works rely on explicit pose estimation [40] and handcrafted feature extraction [9, 10] from optical tactile data for alignment when performing insertion tasks. While interesting, these techniques do not generalize to arbitrary tasks and require significant effort and domain knowledge to adapt to every new task. While some existing works have learned visuotactile policies for precise tasks such as insertion [41, 42], all of these works evaluate performance in restricted settings with little to no spatial variation in the location of the insertion slot. In this paper, we investigate visuotactile policy learning for contact-rich, high-precision tasks requiring spatial generalization, and conclusively show that VISK policies use tactile feedback in conjunction with vision to substantially improve task performance.

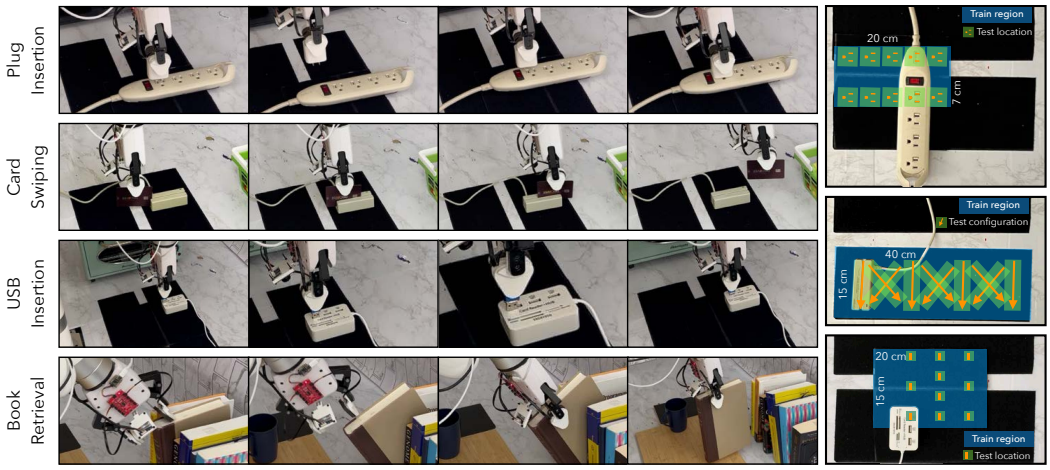

Figure 3: (a) Close-up views of a VISK policy rollout for the four tasks presented. (b) Overhead view depicting variations in target object locations for training and evaluation for plug insertion, USB insertion and card swiping. The blue box denotes the extent of variation in the training data. Test locations for plug insertion and USB insertion are marked on the image. For the card swiping task, arrows denote test locations and orientations of the card machine used for evaluation.

B Visuo-Skin Policy Learning (VISK)

B.1 Policy Architecture

The VISK policy builds on top of BAKU [14], a state-of-the-art transformer-based policy learning architecture that learns visual policies across multiple camera views. Similar to BAKU, our architecture contains three main components:

Sensory Encoders Visual inputs from cameras are encoded using a modified ResNet-18 [18] visual encoder. Low-dimensional tactile inputs from the AnySkin sensor are encoded with a two-layer multilayer perceptron (MLP). Drawing from [31], we subtract a baseline measurement from each tactile reading to account for sensor drift. The encoded representations for each modality are projected to the same dimensionality to facilitate combining modalities in the observation trunk. Some of the ablations and comparisons presented in Section 3 also use DIGIT sensors and robot proprioception as inputs to the policy. In line with prior works using DIGIT sensors for policy learning [19, 20], tactile images from the DIGIT sensor are encoded using the same ResNet-18 encoder as the visual data. The proprioceptive inputs are encoded using a two-layer MLP.

Observation Trunk The encoded inputs from all camera views, robot proprioception, and the tactile signals are treated as separate observation tokens and passed through a transformer decoder network [21]. A learnable action token is appended to the list of observation tokens and is used to obtain action features.

Action Head Finally, an action head takes as input the action features from the observation trunk and predicts the corresponding actions. We found a deterministic action head learned using a mean squared error loss to suffice for our experiments. Considering the temporal correlation in robot movements, we follow prior work [14, 22, 23] and include action chunking to counteract the covariate shift often seen in the low-data imitation learning regime. During inference, we apply exponential temporal smoothing [22] for producing smoother robot motions. Our full policy architecture is depicted in Figure 2.

C Experiments

C.1 Environment Setup

We use a Ufactory xArm 7 robot with its standard two-fingered gripper for all our experiments. To enable tactile sensing, we attach AnySkin sensor tips to the left gripper finger. An identically shaped, plain silicone tip is attached to the right finger. For baseline comparisons with the DIGIT sensor, we use a DIGIT sensor on either fingertip in line with prior work [33]. The camera inputs comprise synchronized RGB images at 128x128 resolution from three static third-person cameras and an egocentric camera mounted on the gripper. The action space is the change in the end-effector pose and gripper state. Our experimental setup is depicted in Figure 2. Learned policies are deployed at a 10Hz frequency.

C.2 Task Descriptions

For all the analysis presented in this paper, we focus on a set of four contact-rich tasks that require high precision as well as spatial generalization. Each task has a target object that the robot must interact with, whose position is varied during demo collection. All evaluations use a fixed set of ten target locations unseen in the training demonstration data.

Plug Insertion This task requires the robot to insert a plug into the first socket on a power strip. The arm starts with the plug grasped and the power strip randomly positioned within a 20cm $\times$ 7cm grid with a fixed orientation. The training dataset consists of 96 demonstrations.

USB Insertion This task has the robot plugging a USB stick into a specific USB port. The arm starts with the USB stick grasped and the USB hub is positioned randomly within a 20cm $\times$ 15cm grid. The training dataset consists of 96 demonstrations.

Card Swiping This task involves swiping a credit card through a card reader. The arm starts with the credit card grasped and the card reader randomly positioned within a 40cm $\times$ 15cm grid, and oriented at a random angle in the range $(-30°, 30°)$ from the direction the robot is facing. The training dataset consists of 96 demonstrations.

Book Retrieval This task requires the robot to retrieve a specific book from a set of eight books placed together, with the order of books randomized each time. The robot must first reach for the target book, pivot it about its edge, and then grasp and pull it out of the bookrack. The training dataset consists of 172 demonstrations.

For the first three tasks, where the robot starts with a grasped object, we do not enforce hard constraints on the grasping location and allow some variability across runs. The extent of variation in target object positions are shown in Fig. B.1. Evaluations are performed on a set of 10 held-out configurations for each task.

C.3 Effect of different input modalities on performance

From Table 1, we find that while the addition of AnySkin inputs to the policy consistently improves performance, the addition of other modalities like the wrist camera and proprioception can have significant impact on policy performance depending on the task. A few consistent patterns emerge across tasks: (1) VISK results in a significant improvement ($\geq 2\times$) in performance over the next best model, indicating its effectiveness on precise, contact-rich manipulation. (2) Adding proprioceptive input almost always results in a drop in performance. This can be attributed to the learned policy overfitting to proprioceptive information which is detrimental to tasks requiring spatial generalizability over target object locations. (3) With the exception of the card swiping task, the addition of a wrist camera improves policy performance. The wrist camera gives the policy a local visual understanding of the scene in the frame of the gripper, and in turn, the same frame as the robot's action space.

This is especially useful for the more fine-grained adjustments required for high-precision tasks. Visualization of demonstration data for the card swiping task indicated that the wrist camera cannot see the card reader due to occlusion from the gripper and therefore simply acts as a noise input to the policy.

While the drops in performance due to proprioception as well as due to the wrist camera in the card swiping task could potentially be addressed by collecting more demonstrations, they highlight the true potential of the VISK framework. The addition of AnySkin and the use of a transformer-based architecture enable the policy to incorporate reliable tactile feedback directly from the interface between the robot and the object being interacted with. The low dimensional nature of AnySkin signal eliminates the need for dimensionality reduction or intermediate representation learning and enables end-to-end learning of visuotactile policies from relatively few (< 200) demonstrations.

C.4 Generalization to Unseen Task Variations

To further probe the strengths of the VISK framework, we investigate performance on unseen task variations for two of the tasks presented above: Plug Insertion and Book Retrieval.

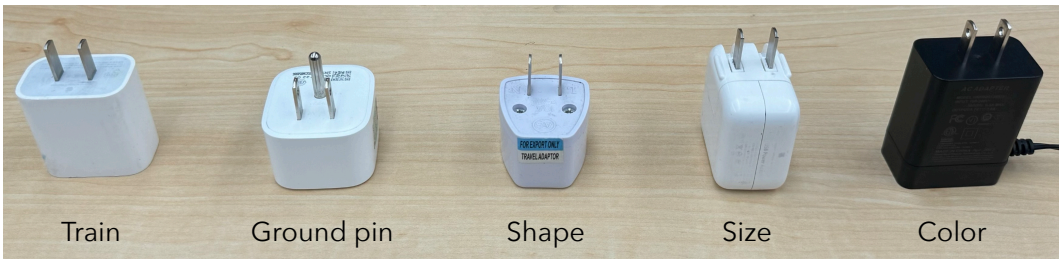

Figure 4: We test the generalization of the best-performing VISK policy to different variations of the plug

C.4.1 Plug Insertion

For plug insertion, we study the efficacy of the best-performing VISK policy on four different variations of the plug as shown in Fig. 4 – addition of ground pin, shape, size and color. We report performance across 10 trials over the same set of target object locations as the previous experiments in Table 2. The VISK policy generalizes surprisingly well to variations in shape, ground pin, and size, despite their pins being in significantly different positions than the plug used for training. This is further evidence of VISK policies effectively leveraging vision and touch even when faced with object variations distinctly different from training. We also find that the policy fails on the color variant which could be attributed to the difficulty of localizing a black plug against a black background.

Table 2: Performance of the best VISK policy on different variations of the plug for the plug insertion task

Train	Ground	Shape	Weight	Color
8/10	6/10	6/10	6/10	1/10

C.4.2 Book Retrieval

Similarly, for the book retrieval task, we study the effectiveness of the best-performing VISK policy on different numbers of books. While our dataset is collected with 8 books, we evaluate performance with 5 and 11 books. For the 5 book variation, we start with the same initial arrangements as used in the original evaluations, and randomly remove 3 books for each trial. For the 11-book variation, we randomize the order of the books for evaluation. Success rates are reported in Table 3. We observe that despite prominent visual differences, the VISK policy is able to generalize to the scenario

with 11 books. This reinstates the effectiveness of the visuotactile representation learned in VISK for generalizing to novel scenarios at inference. However, for the 5-book variation we find that performance drops significantly. This could be attributed to the lower friction from neighboring books resulting from lower inertia.

Table 3: Performance of the best VISK policy on different number of books in the bookrack

Train	5 books	11 books
7/10	3/10	6/10