Enhancing Visual-Motor Policies with Surface Normal Estimation

Zijin Hu
New York University
zijin.hu@nyu.edu

Zifan Zhao
New York University
zz4330@nyu.edu

Team #3

Abstract

Behavior cloning based on vision has been shown to be an effective and powerful approach for learning visuomotor policies. Its key advantage lies in the ease of data collection—requiring only a camera—and the simplicity of training through direct supervised learning. However, RGB images alone provide limited geometric and spatial understanding, whereas human visual-motor decision-making is inherently guided by richer spatial concepts. In this work, we propose incorporating surface normals as a supplementary feature to enhance spatial awareness in learned policies. We show that surface normals introduce invariance to spatial transformations—since surface orientation remains constant under drastic viewpoint or scene rearrangements—leading to more robust performance when the environment changes. We explore the effectiveness of this approach and investigate various ways to integrate priors from single-image surface normal estimation models into the training of robot visual policies.

1 Introduction

Learning visuomotor policies from demonstration allows robots to acquire complex skills directly from expert data with minimal engineering overhead [1, 2, 3]. Yet, RGB inputs alone often fail to capture essential 3D structure, making learned policies brittle to spatial variations such as changes in camera pose, object placement, or lighting. A key insight of our work is that surface normals encode local geometry—the orientation of surfaces—which remains invariant under many common spatial or color variance. By supplying estimated normals alongside RGB images, we provide the policy with features that emphasize task-relevant geometry while abstracting away nuisance variations. We hypothesize this geometric invariance will yield policies that generalize better and remain robust when the scene undergoes drastic spatial changes.

In this work, we utilize this geometric invariance by incorporating surface normal estimation with common RGB information for precise robot manipulation. By explicitly integrating geometric priors into visuomotor policy learning, we enable more robust policy deployment across diverse real-world conditions and simulation environments. Our main contributions include:

- Introducing surface normals as supplementary input for enhanced geometric invariance in visual BC.
- Demonstrating improved policy robustness and generalization to environmental changes.
- Exploring real-world robotic manipulation tasks, validating practicality and effectiveness.
- Exploring SSL visuo-motor robotic policies in simulation.

2 Related Work

2.1 Behavior Cloning for Visuomotor Control

Behavior cloning (BC) treats policy learning as supervised regression from observations to expert actions [4]. Modern approaches factor the policy π into a vision encoder f and an action head g, training both to minimize

$$\mathcal{L}_{\mathrm{BC}} = \mathbb{E}_{(o,a)\sim\mathcal{D}}\big[\ell\big(a,\, g(f(o))\big)\big].$$

Despite its simplicity and effectiveness, common behavior cloning relying purely on RGB images often encounters significant limitations. RGB inputs primarily capture appearance-based features such as color, texture, and basic contours, but inherently lack explicit geometric and spatial information. As a result, visuomotor policies trained solely with RGB data can become highly sensitive to environmental changes such as variations in lighting, viewpoint shifts, or background clutter, leading to poor generalization performance.

Furthermore, reliance on RGB images increases the risk of policy overfitting to specific visual cues rather than learning robust generalizable features. For instance, policies may become overly dependent on particular colors or textures that are irrelevant or coincidental to the task, making them fragile under real-world variations.

2.2 Surface Normal Estimation

Monocular surface normal estimation predicts a 3-dimensional unit normal vector for each pixel, providing rich local orientation information [5, 6, 7]. Surface normals bridge the gap between traditional 2D visual data and explicit 3D geometry. Unlike depth data, which can suffer from inaccuracies due to sensor noise and scale ambiguities [8], or pointclouds,which overly rely on simple spatial cues such as normal directions or point heights rather than capturing higher-level semantic information [9], surface normals provide consistent geometric representations directly from monocular imagery without requiring calibration or additional sensor modalities. This simplicity and direct applicability make them highly advantageous for real-world robotic applications, where practical constraints often limit the deployment of complex sensor setups.

Moreover, surface normals inherently encode critical geometric cues about surface curvature and local shape. As illustrated in Figure 1, the geometric information captured by surface normals can significantly enhance visual understanding beyond mere texture or color features present in standard RGB images. Consequently, incorporating surface normals into perception pipelines helps disambiguate visually similar textures or patterns that might otherwise confuse purely appearance-based models.

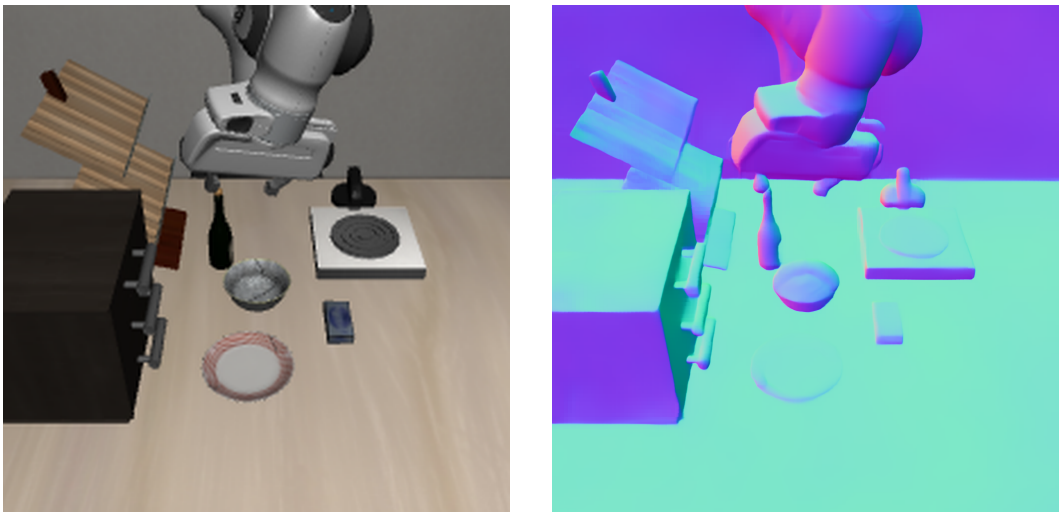
Recent work highlights the utility of surface normals across various computer vision tasks. For instance, normals have proven beneficial for semantic segmentation, enabling models to distinguish objects more effectively by explicitly considering local geometric contexts [10]. In pose estimation tasks, surface normals facilitate improved spatial reasoning, allowing for more accurate and stable predictions under viewpoint variations [11].

Given these strengths, integrating surface normal estimation into visuomotor policies offers significant potential to enhance robotic manipulation capabilities. By providing explicit geometric priors that are invariant to common environmental changes, surface normals have the potential to improve policy robustness and generalization, critical for practical deployment in dynamic environments.

3 Method

Our goal is to investigate how surface normal information can be effectively incorporated into vision encoders for visuomotor policy learning. We hypothesize that combining RGB inputs with geometric surface orientation can enhance spatial generalization by providing invariant cues. To test this, we explore two fusion strategies (illustrated in Figure 2):

Early Fusion. In the early fusion (Figure 2a) setup, we concatenate the RGB image $I \in \mathbb{R}^{H\times W\times 3}$ with the corresponding surface normal map $N \in \mathbb{R}^{H\times W\times 3}$ to form a 6-channel input tensor $[I \,\|\, N]$. We modify the first convolutional layer of a ResNet-18 encoder to accept 6-channel inputs and train



(a) Original image (b) Estimated surface normals

Figure 1: Example of surface normal estimation from scene in Libero Goal [12].

the encoder using a self-supervised learning (SSL) method. This allows the model to learn unified representations of appearance and geometry from the earliest stages of the visual pipeline. The fused input is encoded as:

$$s = f_\theta([I \,\|\, N]),$$

where f_θ is a ResNet-18 encoder with parameters θ, and s is the resulting joint visual-geometric representation.

Late Fusion. In the late fusion setup (Figure 2b), we use two separate ResNet-18 encoders: one for RGB input $f_{\mathrm{RGB}}(I)$, and another for surface normal input $f_{\mathrm{N}}(N)$. Features from both encoders are fused with a 2-layer MLP:

$$s = \mathrm{MLP}\big(f_{\mathrm{RGB}}(I),\ f_{\mathrm{N}}(N)\big).$$

The RGB encoder is trained using DynaMo [13], while the surface normal encoder is trained using either DynaMo or a BYOL-style SSL method (Figure 2c). For the latter, we apply only spatial augmentations—Gaussian blur, random horizontal flip, and random resized crops—since color jitter and grayscale are not meaningful for unit normal vectors.

The resulting representation is passed to an action prediction head, and an imitation policy is trained on top of it. Further details are provided in the Experiments section.

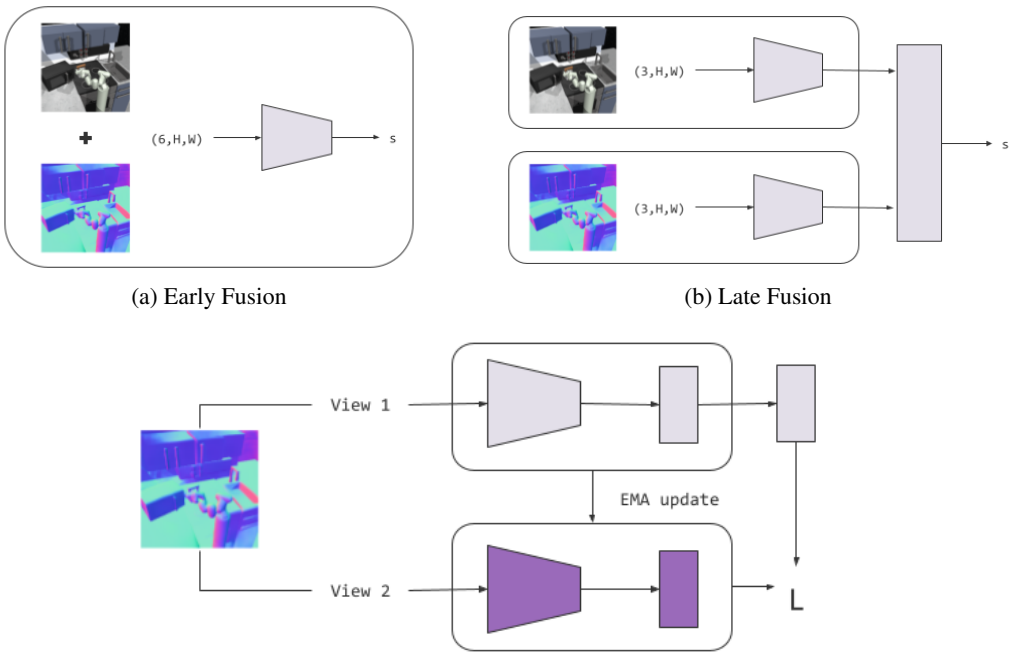
4 Experiments

Our experiments aim to evaluate the effectiveness of incorporating surface normal estimation into visuomotor policy learning. Specifically, we investigate the following:

- Does adding surface normal information improve generalization to spatial and visual scene variations?
- How do early fusion and late fusion strategies compare in terms of representation quality and downstream policy performance?

4.1 Environments and Datasets

We evaluate our method in both simulated and real-world settings to assess generalization and scalability of surface-normal-enhanced policies. Our core hypothesis, particularly relevant for real-world tasks, is that precise manipulation tasks possess consistent local geometric and interaction dynamics across different scenarios, despite variations in broader environmental contexts. For



(a) Early Fusion

(b) Late Fusion

(c) Normal-Only BYOL

Figure 2: Fusion strategies explored for incorporating surface normal input. (a) Early fusion concatenates RGB and normal maps as a 6-channel input. (b) Late fusion uses separate encoders for each modality. (c) BYOL-style pretraining on normal maps for geometric feature extraction.

example, inserting a plug into a socket maintains identical geometric and interaction properties regardless of whether the socket is mounted on a wooden table or a metallic panel.

Franka Kitchen (Simulation). We use the Franka Kitchen benchmark [14], which features 7 manipulation tasks in a simulated kitchen environment and a dataset of 566 human-collected demonstrations. The observations are RGB frames rendered at 224×224 resolution, and surface normals are estimated using a pretrained monocular normal predictor (DSINE [6]). We evaluated 100 rollouts and reported the mean number of completed tasks (maximum 4).

xArm Manipulation Suite (Real World). Our physical setup uses a UFACTORY xArm 7 robot equipped with a two-fingered xArm Gripper. The observations for policy learning include 128×128 RGB images captured by a fisheye camera mounted on the robot's wrist; the estimated surface normal maps have the same 128×128 resolution, estimated via the same DSINE model used in simulation. Figure 3 shows an example of the fisheye view and the corresponding estimated surface normal. Observations and actions are explicitly grounded in the robot's end-effector frame, ensuring robust generalization to novel spatial configurations during policy deployment. Expert demonstrations are collected through a VR teleoperation framework [15] (HTC Vive trackers + Unity), recorded at 30 Hz and downsampled to 6 Hz for training and deployment. We collect 50 episodes per task, each lasting 5–10 s. Figure 4 illustrates the real-world xArm environment.

The following tasks are performed using the xArm setup:

- **Plug in Socket:** The robot arm holds a plug within the gripper and is tasked with inserting the plug into a socket. The robot's initial position is randomly sampled in a 6×6 cm area around the socket, 10 cm above the socket.
- **USB Insertion:** The robot arm holds a USB stick within the gripper and is tasked with going down and inserting the USB stick into a USB socket. The robot's initial position is randomly sampled in a 6×6 cm area around the socket, 10 cm above the socket.

(a) Original fisheye view

(b) Estimated surface normal view

Figure 3: Example of input data: original fisheye view from the robot's wrist camera and the corresponding estimated surface normal map.

Figure 4: XArm Environment Setting for real-world tasks.

- **Card Swiping:** The robot arm holds a credit card within the gripper and is tasked with swiping the card through a card machine. The robot's initial position is sampled in a $4 \times 4 \times 2$ cm bounding box in front of the card machine.

4.2 Does adding surface normal information improve generalization to spatial and visual scene variations?

To test the benefit of surface normal features in improving policy robustness, we evaluate in two settings: (1) with background visual changes introduced via green-screen domain augmentation, and (2) in the standard real-world scene. Crucially, we leverage surface normal estimation alongside standard RGB images from a wrist-mounted camera as input observations. Surface normals provide robust geometric cues that are invariant to changes in background textures, colors, and lighting conditions, thus enhancing the policy's robustness to environmental variations.

4.2.1 Comparison under Background Variation (Real World)

To test robustness to appearance shifts, we apply semantic augmentations: demonstrations are recorded against a green screen background, facilitating procedural background replacement via scene generation techniques [16] using RoboEngine [17]. This method intentionally targets visual

regions irrelevant to task execution, enabling the policy to focus on task-critical geometric and visual features, thereby reducing susceptibility to visual distractions. The visual content behind the foreground (i.e., the arm and object) is randomized, while surface normals (derived from the original, non-augmented scene) remain unchanged.

Table 1: Performance under visual domain shift (background change). Success rates over 10 trials.

Method	Plug in Socket	USB Insertion	Card Swiping
Visual BC [18]	0/10	0/10	0/10
Visuo-Normal BC (ours)	3/10	1/10	2/10
Visual BC + Aug.	**5/10**	**2/10**	1/10
Visuo-Normal BC + Aug. (ours)	**5/10**	**2/10**	**4/10**

Implementation. A randomly-initialized and then frozen ResNet-18 encoder [19] processes these enriched visual inputs (RGB and/or surface normal maps concatenated channel-wise for early fusion), and the resulting encoded features are passed into a transformer-based policy [18, 20] for predicting actions. The training employs action chunking [21] and minimizes a mean squared error loss between the predicted and ground-truth action chunks.

Conclusion. The results from Table 1 show that surface normal integration enhances policy robustness under background variations. Without augmentation, Visual BC fails entirely across all tasks, whereas Visuo-Normal BC demonstrates moderate robustness, achieving successes in each task. When combined with semantic augmentation, Visuo-Normal BC + Aug. matches or surpasses the performance achieved by visual-only methods with augmentation, highlighting the complementary benefits of geometric priors and visual augmentation techniques.

4.2.2 Comparison in Standard Scene Configuration (Real World)

We evaluate the same methods in the original unmodified environment to verify that surface normals still provide a benefit even without domain augmentation. Each policy is evaluated over 10 trials per task, using the same randomization ranges.

Table 2: Performance under fixed background (no augmentation). Success rates over 10 trials.

Method	Plug in Socket	USB Insertion	Card Swiping
Visual BC [18]	4/10	4/10	1/10
Visuo-Normal BC (ours)	**5/10**	**4/10**	**3/10**

Conclusion. The results from the surface normal ablation study, presented in Table 2, clearly demonstrate the benefit of incorporating surface normals into behavior cloning policies even without appearance variation during training. In all tasks, the Visuo-Normal BC achieves a higher or same success rate compared to the Visual BC. This result underscores the advantage of integrating geometric information via surface normals, particularly for tasks requiring precise alignment and contact interactions, highlighting its value as a geometric prior.

Overall Discussion on Surface Normal Generalization. These real-world results collectively validate our hypothesis that surface normals provide meaningful geometric invariance, substantially improving the policy's robustness to visual and spatial perturbations. The successful deployment highlights the practical value of integrating explicit geometric representations, such as surface normals, into visuomotor learning frameworks, particularly when deployed in real-world environments characterized by diverse and dynamic visual conditions. Additionally, surface normal estimation offers computational advantages compared to other 3D geometric priors, with the estimation process for 128×128 images averaging only 0.05 seconds per image. This efficiency is compatible with the high-frequency deployment of real-world robots and further highlights the practicality of surface normals as geometric priors in real-world robotic applications.

4.3 How do early fusion and late fusion strategies compare in downstream policy performance?

To assess the impact of different fusion strategies for incorporating surface normals into vision encoders, we also include controlled evaluations in the Franka Kitchen simulated environment.

Implementation. We pretrained ResNet-18 vision encoders under different modality configurations using different self-supervised learning method:

- **RGB + Normals (separate encoders):** RGB and surface normal inputs are encoded separately using two ResNet-18 encoders. Their outputs are fused at a higher layer. We test two variants: (1) Dynamo for RGB and BYOL for normals, and (2) Dynamo for both modalities.
- **RGB + Normals (concatenated):** RGB and surface normal maps are concatenated channel-wise to form a 6-channel input and processed by a single ResNet-18 encoder (early fusion).
- **RGB only:** Standard 3-channel RGB input encoded using a single ResNet-18 trained with Dynamo.
- **Surface Normal only:** Surface normals alone are encoded using a single ResNet-18 trained with Dynamo.

All pretrained encoders are frozen and used as visual backbones for a VQ-BeT [2] policy head. The imitation policy is then trained using behavior cloning on the frozen representations.

Table 3: Downstream policy performance on Franka Kitchen. Metric: average number of successful task completions (out of 4).

Modality Setup	SSL Method (RGB / Normal)	Result (./4)
RGB + Normals (late fusion)	Dynamo / BYOL	**3.39**
RGB + Normals (late fusion)	Dynamo / Dynamo	3.06
RGB + Normals (early fusion)	Dynamo / —	3.22
Surface Normal only	— / Dynamo	1.32
RGB only	Dynamo / —	3.28

Conclusion As detailed in Table 3, surface normals alone are insufficient for effective policy learning, yielding the poorest results when used in isolation. However, they contribute useful geometric insights when paired with RGB inputs, which is particularly evident in the late fusion approach. We achieved the best performance by training RGB inputs with the Dynamo self-supervised method and surface normals with the BYOL self-supervised method. However, since Dynamo is originally designed and tuned for RGB images, it is not surprising that it does not work well out of the box for the surface normal modality. Furthermore, we found that early fusion performed slightly worse than the baseline that used only RGB inputs. Again, this may be improved by tuning Dynamo further. Collectively, these results show the effectiveness of adding surface normals but also demonstrate that the SSL method used to extract features from them is important. This implies that aligning each input modality with a compatible self-supervised learning objective is critical for improving the performance of the downstream policy learning.

5 Discussion and Conclusion

We have demonstrated that augmenting behavior cloning with surface normal information significantly enhances the robustness of visuomotor policies to drastic spatial changes. Surface normals, by encoding local orientation in a scale- and lighting-invariant form, provide consistent geometric cues that the policy can leverage to generalize beyond its training distribution. Between our two integration strategies, auxiliary prediction yielded the best trade-off between performance and architectural simplicity.

Future Work We plan to extend this approach to real-world robot experiments and explore combining normals with temporal memory modules for dynamic scene understanding. Additionally, investigating learned normal estimators trained jointly with the policy may further improve performance. Lastly, we can also use Surface Normal Prediction as an auxiliary

References

[1] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion, 2024.

[2] Seungjae Lee, Yibin Wang, Haritheja Etukuru, H. Jin Kim, Nur Muhammad Mahi Shafiullah, and Lerrel Pinto. Behavior generation with latent actions, 2024.

[3] Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware, 2023.

[4] Dean Pomerleau. Alvinn: An autonomous land vehicle in a neural network. In D.S. Touretzky, editor, *Proceedings of (NeurIPS) Neural Information Processing Systems*, pages 305 – 313. Morgan Kaufmann, December 1989.

[5] Derek Hoiem, Alexei A. Efros, and Martial Hebert. Recovering surface layout from an image. *International Journal of Computer Vision*, 75:151–172, 2007.

[6] Gwangbin Bae and Andrew J. Davison. Rethinking inductive biases for surface normal estimation, 2024.

[7] David F. Fouhey, Abhinav Gupta, and Martial Hebert. Data-driven 3d primitives for single image understanding. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2013.

[8] Yuchen Yang, Xuanyi Liu, Xing Gao, Zhihang Zhong, and Xiao Sun. X as supervision: Contending with depth ambiguity in unsupervised monocular 3d pose estimation, 2024.

[9] Xiaoyang Wu, Daniel DeTone, Duncan Frost, Tianwei Shen, Chris Xie, Nan Yang, Jakob Engel, Richard Newcombe, Hengshuang Zhao, and Julian Straub. Sonata: Self-supervised learning of reliable point representations. In *CVPR*, 2025.

[10] Rui Wang, David Geraghty, Kevin Matzen, Richard Szeliski, and Jan-Michael Frahm. Vplnet: Deep single view normal estimation with vanishing points and lines. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 686–695, 2020.

[11] Ainaz Eftekhar, Alexander Sax, Roman Bachmann, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans, 2021.

[12] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning, 2023.

[13] Zichen Jeff Cui, Hengkai Pan, Aadhithya Iyer, Siddhant Haldar, and Lerrel Pinto. Dynamo: In-domain dynamics pretraining for visuo-motor control, 2024.

[14] Abhishek Gupta, Vikash Kumar, Corey Lynch, Sergey Levine, and Karol Hausman. Relay policy learning: Solving long-horizon tasks via imitation and reinforcement learning, 2019.

[15] Aadhithya Iyer, Zhuoran Peng, Yinlong Dai, Irmak Guzey, Siddhant Haldar, Soumith Chintala, and Lerrel Pinto. Open teach: A versatile teleoperation system for robotic manipulation. *arXiv preprint arXiv:2403.07870*, 2024.

[16] Eugene Teoh, Sumit Patidar, Xiao Ma, and Stephen James. Green screen augmentation enables scene generalisation in robotic manipulation. *arXiv preprint arXiv:2407.07868*, 2024.

[17] Chengbo Yuan, Suraj Joshi, Shaoting Zhu, Hang Su, Hang Zhao, and Yang Gao. Roboengine: Plug-and-play robot data augmentation with semantic robot segmentation and background generation. *arXiv preprint arXiv:2503.18738*, 2025.

[18] Siddhant Haldar, Zhuoran Peng, and Lerrel Pinto. Baku: An efficient transformer for multi-task policy learning. *arXiv preprint arXiv:2406.07539*, 2024.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[20] Venkatesh Pattabiraman, Yifeng Cao, Siddhant Haldar, Lerrel Pinto, and Raunaq Bhirangi. Learning precise, contact-rich manipulation through uncalibrated tactile skins. *arXiv preprint arXiv:2410.17246*, 2024.

[21] Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware. In *Proceedings of Robotics: Science and Systems*, Daegu, Republic of Korea, July 2023.