

Residual Rotation Correction using Tactile Equivariance

Yizhe Zhu¹, Zhang Ye¹, Boce Hu¹, Haibo Zhao¹, Yu Qi¹, Dian Wang^{2†}, and Robert Platt^{1†}

Abstract—Visuotactile policy learning augments vision-only policies with tactile input, facilitating contact-rich manipulation. However, the high cost of tactile data collection makes sample efficiency the key requirement for developing visuotactile policies. We present EquiTac, a framework that exploits the inherent $\text{SO}(2)$ symmetry of in-hand object rotation to improve sample efficiency and generalization for visuotactile policy learning. EquiTac first reconstructs surface normals from raw RGB inputs of vision-based tactile sensors, so rotations of the normal vector field correspond to in-hand object rotations. An $\text{SO}(2)$ -equivariant network then predicts a residual rotation action that augments a base visuomotor policy at test time, enabling real-time rotation correction without additional reorientation demonstrations. On a real robot, EquiTac accurately achieves robust zero-shot generalization to unseen in-hand orientations with very few training samples, where baselines fail even with more training data. To our knowledge, this is the first tactile learning method to explicitly encode tactile equivariance for policy learning, yielding a lightweight, symmetry-aware module that improves reliability in contact-rich tasks. <https://yizhezhu0925.github.io/equitac.github.io/>.

I. INTRODUCTION

As robots increasingly rely on touch to perform precise, contact-rich interactions, developing visuotactile policies has become a key challenge in robotic manipulation. While data-driven visuomotor policy learning has achieved remarkable progress, extending these advances to tactile sensing remains fundamentally data-constrained. Unlike visual data, tactile signals are harder to scale, since collecting sufficient tactile interactions to cover diverse contact conditions is extremely expensive, and publicly available tactile datasets are still limited in both scale and diversity [1], [2], [3]. Thus, learning tactile manipulation policies, or even tactile reasoning modules that augment vision-based policies, demands far greater sample efficiency than comparable vision-only pipelines.

To mitigate the above data constraints, equivariant learning, i.e., embedding task symmetries as inductive bias in neural networks, offers a direct route to improving sample efficiency and generalization. By baking in spatial regularities that would otherwise need to be learned from data, equivariant policy learning in robotic manipulation has consistently shown gains across imitation [4], [5] and RL [6] pipelines, enabling few-shot and on-robot learning [7], [8]. However, existing studies remain largely vision-centric: they design equivariance for image or scene geometry, while equivariance in tactile learning

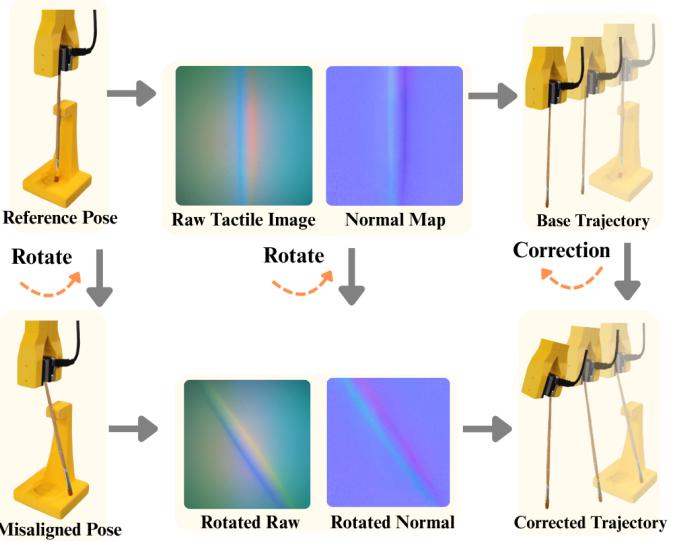


Fig. 1: **Equivariance in EquiTac.** When the tactile observation is rotated, the predicted action rotates consistently.

remains under-explored. In particular, prior equivariant robot-learning approaches do not explicitly model the object-in-gripper rotational symmetry that underlies tactile feedback, which is crucial for contact-rich manipulation.

To close this gap, this paper introduces *EquiTac*, the first equivariant tactile learning pipeline. Our core idea is to leverage in-hand object orientation symmetry: when the grasped object rotates within the gripper, the tactile signal should transform predictably. However, in vision-based tactile sensors, the raw RGB intensities are often distorted by internal illumination effects (e.g., RGB LEDs), making naive in-plane image rotation inconsistent with the true contact geometry. We address this by reconstructing surface normal maps from tactile images and decomposing them into equivariant (in-plane x, y vector) and invariant (out-of-plane normal- z) components. This representation restores rotation-consistent behavior and provides the correct carriers for $\text{SO}(2)$ -equivariant reasoning. Building on this representation, we propose a residual policy correction framework, where a rotation residual predicted by an equivariant network augments a trained manipulation policy at test time to correct the gripper’s orientation. Figure 1 illustrates the core idea of this equivariant correction: when the object rotates in hand, the equivariant network observes a rotated normal map and predicts a corresponding target gripper rotation, from which an angular residual is computed to correct the base policy. By injecting this symmetry-aware tactile correction, the policy generalizes zero-shot to unseen in-hand object orientations without retraining.

[†]Equal Advising. Corresponding to dianwang@stanford.edu

¹Yizhe Zhu, Zhang Ye, Boce Hu, Haibo Zhao, Yu Qi and Robert Platt are with Northeastern University, Massachusetts, MA, USA {zhu.yizhe, ye.zhang1, hu.boce, zhao.haib, qi.yu2, r.platt@northeastern.edu}

²Dian Wang is with Stanford University, California, CA, USA dianwang@stanford.edu

Concretely, we make the following contributions:

- 1) We propose EquiTac, which encodes in-hand rotational symmetry via normal-map parameterization and an SO(2)-equivariant network to support symmetry-consistent tactile representations. To the best of our knowledge, this is the first work to introduce equivariance into tactile learning.
- 2) We develop an equivariant tactile residual correction pipeline, a lightweight module that predicts a target gripper rotation from a single tactile frame and computes an angular residual to correct an existing policy zero-shot for unseen in-hand orientations.
- 3) We empirically validate EquiTac on in-hand angle estimation and residual policy correction, demonstrating higher prediction accuracy and strong zero-shot generalization to unseen object orientations.

II. RELATED WORK

Visuotactile Manipulation. Tactile feedback is critical for contact-rich manipulation, especially when vision alone cannot reliably determine an object’s in-hand orientation [9], [10], [11], [12], [13], [14]. Recent visual–tactile imitation-learning methods fuse modalities to reconstruct object pose [15], [16], [17], but they generally do not adjust policy outputs online using contact information. These approaches achieve promising results, but they generally lack the ability to actively adjust policy outputs in real time using contact information. Some methods incorporate shear-force sensing and deploy auxiliary models running at different inference rates to refine the original policy’s actions online [18]. However, they do not use the object’s rotation information to update the policy outputs in real time. Other studies designed an “orientation readjustment” phase during data collection to satisfy downstream requirements [19]. However, such task-specific adjustment strategies are not broadly generalizable and substantially increase data collection costs. In contrast, our method preserves the original vision-based policy structure while making real-time, contact-driven adjustments without adding an extra, manually crafted orientation-adjustment stage.

Residual learning in robotics. Residual learning has proven effective in robotics. Methods can be grouped into three categories: reinforcement learning; human corrections; and interactive imitation learning. RL-based work [20], [21], such as Policy Decorator [22], EXPO [23] and ResiP [21], learns residual corrections through interaction with off-policy algorithms or by combining behavioral cloning with RL. Methods involving human corrections, such as TRANSIC [24] and CR-Dagger [25], collect teleoperation data or force-aware adjustments from operators and use this to train residuals with supervision. Interactive imitation learning, as demonstrated by HG-Dagger [26], enables experts to intervene during execution and provide corrective demonstrations. However, these approaches often require thousands of online interactions, substantial human supervision during deployment, and complex sim-to-real transfer. Furthermore, the residual networks are often similar in size to the base policy. Our method takes a

different approach. We encode rotational equivariance from tactile data to achieve zero-shot generalization, eliminating the need for online RL or sim-to-real transfer. The correction module can be trained with supervision on a single example. This reduces sample complexity and system overhead, enabling direct deployment in a real environment.

Equivariance in Robotics. Equivariance has been shown to boost performance and improve sample efficiency [6], [8], [27], [28], [7], [29], [30], [31], [32], [33], [34], [35], allowing policies to learn effectively from far fewer demonstrations; it has been applied across open-loop and closed-loop settings as well as diverse dataset generation. Recently, the idea has been extended to closed-loop diffusion policies: Equivariant Diffusion Policy [36] augments diffusion policies with an SO(2)-equivariant architecture to leverage task symmetries, yielding better generalization and data efficiency. Beyond planar symmetries, 3D-Spherical Projection [37] achieves SO(3) equivariance from a single RGB camera by projecting features onto a sphere for real-time visuomotor control. However, these approaches typically do not explicitly model object-in-gripper rotation equivariance, which is crucial for contact-rich or fine manipulation tasks. To address this gap, we develop a new equivariant framework that enables real-time, visuo-tactile action correction using tactile and visual sensing.

III. BACKGROUND

A. Equivariance

Let G be a symmetry group, a function $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is *equivariant* with respect to G if applying a group transform before f is equivalent to transforming the output after f :

$$f(\rho_{\text{in}}(g)x) = \rho_{\text{out}}(g)f(x), \quad \forall g \in G.$$

In this work, we focus on planar rotations $G = \text{SO}(2)$. Here, ρ_{in} and ρ_{out} are called the group representations that define how the input and output transform under a rotation $g \in \text{SO}(2)$. For example, $\rho = 1$ corresponds to the trivial (invariant) representation acting on scalars under rotation, and $\rho = R_g = \begin{bmatrix} \cos g & -\sin g \\ \sin g & \cos g \end{bmatrix}$ represents the standard representations of $\text{SO}(2)$ acting on vectors.

B. Flow Matching Policy

Flow-based policy learning [38], [39] is a class of imitation learning methods that model action generation as a continuous-time transport process using learned velocity fields. These methods learn to transform noise samples into structured actions through ordinary differential equation (ODE) integration, conditioned on multi-modal observations. Formally, given an observation encoding c (which may include visual features, tactile signals, and proprioceptive states) and a flow time $t \in [0, 1]$, the policy learns a time-dependent velocity field $v_\theta(x, t, c)$ that transports a noise sample $x_0 \sim \mathcal{N}(0, \sigma^2 I)$ toward a target action x_1 . During training, the method constructs interpolated states via $x_t = (1-t)x_0 + t x_1$, where the ground-truth velocity is $\frac{dx_t}{dt} = x_1 - x_0$. The model is trained to

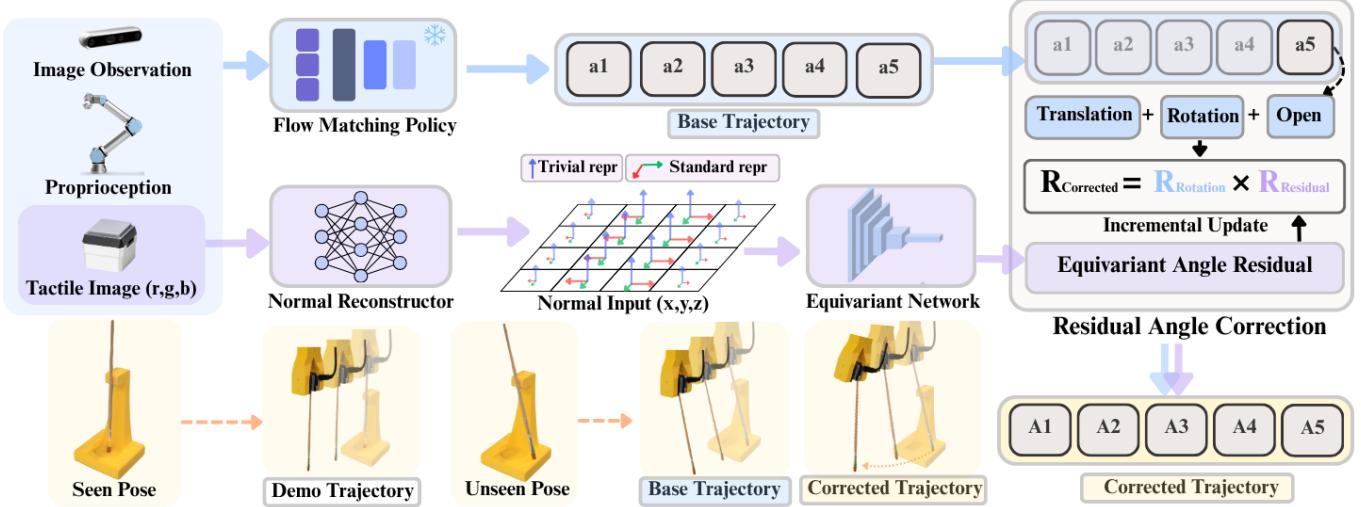


Fig. 2: **Overview of our tactile-guided manipulation framework with equivariant orientation correction.** The system begins with a Flow Matching Policy (top) that predicts basic action chunks from multimodal inputs including robot proprioception, tactile images, and three camera views. During action execution (middle), tactile images are processed through a Normal Reconstructor to obtain normal maps, which are fed into an $\text{SO}(2)$ -equivariant network. The equivariant network will predict the angular residual between the object's current and target orientations, enabling real-time correction of the action chunk to correct for misalignment. The bottom row shows (i) the data-collection setup with ideal object placement and (ii) the results of executing the base and corrected trajectories under placement deviations at rollout.

minimize the velocity-matching objective:

$$\mathcal{L} = \mathbb{E}_{x_0, x_1, t} \| v_\theta(x_t, t, c) - (x_1 - x_0) \|^2.$$

At test time, the policy generates actions by sampling $x_0 \sim \mathcal{N}(0, \sigma^2 I)$ and numerically integrating the learned ODE $\frac{dx}{dt} = v_\theta(x, t, c)$ with initial condition $x(0) = x_0$. The integration is performed from $t = 0$ to $t = 1$ using N Euler steps: $x_{i+1} = x_i + \frac{1}{N} v_\theta(x_i, t_i, c)$ where $t_i = \frac{i}{N}$, yielding the final action x_N .

IV. METHOD

A. Problem Statement

We study closed-loop visuotactile policy learning for contact-rich robotic manipulation that can generalize to unseen object orientations, formulated as imitation learning from expert demonstrations. Specifically, we consider an expert demonstration dataset $\mathcal{D} = \{(\mathcal{O}, \mathcal{A})\}_{t=1}^N$, where $\mathcal{O} = \mathcal{O}_v \times \mathcal{O}_p \times \mathcal{O}_{\text{tac}}$ includes visual observations from RGB cameras, proprioceptive readings from the robot, and tactile measurements from a touch sensor. $\{a_1, a_2, \dots, a_m\} \in \mathcal{A}$ denotes an action chunk consisting of a sequence of robot actions in the next m time steps.

The goal is to learn a policy that integrates a visuomotor base policy π_b with a tactile residual policy π_r . The base policy $\pi_b : \mathcal{O} \rightarrow \mathcal{A}$ first predicts an action chunk from multi-modal observations, then the tactile residual policy $\pi_r : \mathcal{O}^{\text{tac}} \times \mathcal{A} \rightarrow \mathcal{A}$ generates the corrected action chunk from the tactile feedback and the base actions. Together, the goal for the composed policy $\pi = \pi_r \circ \pi_b$ is to achieve robust manipulation under varying in-hand object orientations.

B. Overview of EquiTac

Figure 2 shows the overview of EquiTac. Given a multi-modal observation consisting of RGB camera images, robot proprioception, and a tactile image, we first use a flow-matching policy π_b to produce a base action chunk. In parallel, an equivariant tactile residual policy π_r will generate an in-hand rotation residual, which is applied to the base action chunk to correct the angular error caused by the unseen in-hand object pose.

Our equivariant tactile residual policy π_r has three main steps. First, we use a learnable mapping to convert the raw tactile image into a normal map. Second, taking the normal map as the input, an equivariant network predicts a target *in-hand yaw* rotation. Let $\{x, y, z\}$ be the fingertip local frame with z the contact normal, in-hand yaw is the rotation about z , i.e., the rotation about the fingertip surface normal (Figure 3). Finally, we calculate a rotation residual between the predicted target yaw and the current gripper orientation, and apply it incrementally to the base action sequence to refine the gripper's rotation.

C. Surface Normal Map Reconstruction

To enable rotational equivariance, the tactile representation must transform consistently with the object's true in-hand rotation. Vision-based tactile sensors illuminated by fixed RGB

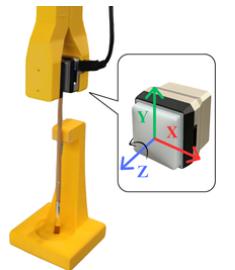


Fig. 3: Sensor on the gripper fingertip, z -axis denotes the finger normal.

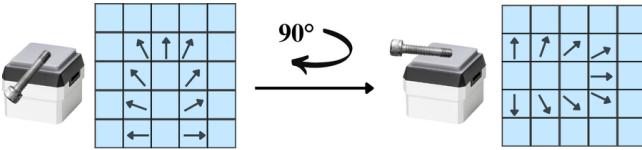


Fig. 4: **Equivariance of the normal map.** When the object rotates in hand, the normal map co-rotates as a vector field.

LEDs do not satisfy this property directly, as identical local geometry can yield different colors as the object rotates. Consequently, naively using RGB image rotation as the representation does not align with contact geometry.

We resolve this by reconstructing a surface normal map $N \in \mathbb{R}^{3 \times H \times W}$ using a lightweight MLP, where each pixel of the normal map stores a unit normal vector, $\vec{n} = (n_x, n_y, n_z)$. Under an in-hand object rotation $g \in \text{SO}(2)$ about the fingertip normal, all the pixels in the normal map will rotate accordingly. Moreover, for each individual vector \vec{n} , the in-plane components (n_x, n_y) rotate as a 2-vector, while n_z remains invariant,

$$g \cdot \vec{n} = \begin{bmatrix} R_g & 0 \\ 0 & 1 \end{bmatrix} (n_x, n_y, n_z)^T,$$

where R_g is the 2×2 rotation matrix for g . Figure 4 shows an example of such transformation. We denote $g \cdot N$ as this vector field rotation. This representation provides the correct carriers for $\text{SO}(2)$ -equivariant reasoning and ensures the representation co-rotates with the object.

We follow the standard GelSight calibration procedure [40], [2] to obtain pixel-wise ground truth geometry for training the normal reconstructor. Specifically, a metal calibration sphere is pressed against the sensor surface at multiple locations to generate contact samples with known geometry. For each contact, the local surface gradient is computed analytically from the known spherical geometry. Using these gradients as supervision, an MLP is trained to directly predict the surface gradient at each pixel from its color and spatial location (R, G, B, U, V). The predicted gradient map is then converted into a surface normal map via standard gradient-to-normal conversion.

D. Equivariant Angular Residual Tracking

After acquiring the normal map N , we aim to train an equivariant network $\phi : N \mapsto (\cos -\alpha_t, \sin -\alpha_t)$ that predicts the target in-hand yaw α_t in the form of a unit vector on a unit circle. The model satisfies

$$\phi(g \cdot N) = g \cdot \phi(N) = R_g(\cos -\alpha_t, \sin -\alpha_t)^T,$$

where R_g is the standard 2×2 rotation matrix. Thus, when the object rotates on the finger by $g \in \text{SO}(2)$, the input normal map N becomes $g \cdot N$, and the output target in-hand yaw vector will counter-rotate accordingly, compensating for the in-hand object rotation. This property facilitates zero-shot generalization, achieving precise angular estimation without training on tactile images with different object orientations.

In practice, although an ideal equivariant function can theoretically generalize to unseen orientations without additional training, the discretization of $\text{SO}(2)$ in implementation leads to incomplete coverage of the rotation space.

To address this, we augment the training data by randomly rotating the entire normal map with uniformly sampled angles, thereby preserving geometric consistency with the equivariant structure of the model.

E. Equivariant Angular Residual Correction

We use Flow Matching (III-B) as the base policy π_b to progressively denoise an initial noisy action and obtain a base action chunk $\{a_i = (T_i^b, p_i)\}_{i=1}^m$, where $T_i^b \in \mathbb{R}^{4 \times 4}$ is a transformation matrix representing the center of the gripper fingertips in the world frame, and p_i is the gripper command. The equivariant tactile residual policy π_r uses ϕ to estimate the in-hand yaw target α_t , then apply a proportional update to T_i^b by a rotation in the gripper frame. Specifically, given the current in-hand yaw α we define the residual transformation matrix as a rotation about the z-axis in the gripper frame (Figure 3),

$$T^r = \begin{bmatrix} \cos K_p(\alpha_t - \alpha) & -\sin K_p(\alpha_t - \alpha) & 0 & 0 \\ \sin K_p(\alpha_t - \alpha) & \cos K_p(\alpha_t - \alpha) & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

where $K_p \in (0, 1)$ is a proportional gain. The residual policy correction is achieved by right multiplication $T_i^b \cdot T^r$. This proportional update mechanism ensures that each value in the angular residual gradually adjusts the rotation angle in the current action chunk toward the target. In the actual implementation, our overall policy runs at two rates: the base Flow Matching policy emits action chunks at a lower rate, while a lightweight equivariant module provides high-frequency in-hand yaw corrections.

F. Implementation Details

Our network architecture follows the theoretical formulation of $\text{SO}(2)$ equivariance described in IV-D, which is a four-layer C_8 -equivariant convolutional neural network, implemented using the ESCNN [41] library. We adopt a double-angle representation on the output, where the network is trained to predict $(\cos -2\alpha_t, \sin -2\alpha_t)$. At inference time, the orientation is recovered via $\alpha_t = -\frac{1}{2} \arctan 2(\sin -2\alpha_t, \cos -2\alpha_t)$. This double-angle representation has two advantages. First, it maps both α_t and $\alpha_t + \pi$ to the same output vector, preventing the ambiguity caused by many objects that are symmetric over π rotation. Second, it ensures $\alpha_t \in [-\frac{\pi}{2}, \frac{\pi}{2}]$, restricting corrections to the forward-facing semicircle and preventing inward-pointing actions that could cause collisions.

V. MODEL EVALUATIONS

A. Ablation on Model Design and Input Representation

We conduct ablation studies to evaluate the impact of the equivariant network ϕ and input representation on in-hand object orientation prediction. Experiments are performed on

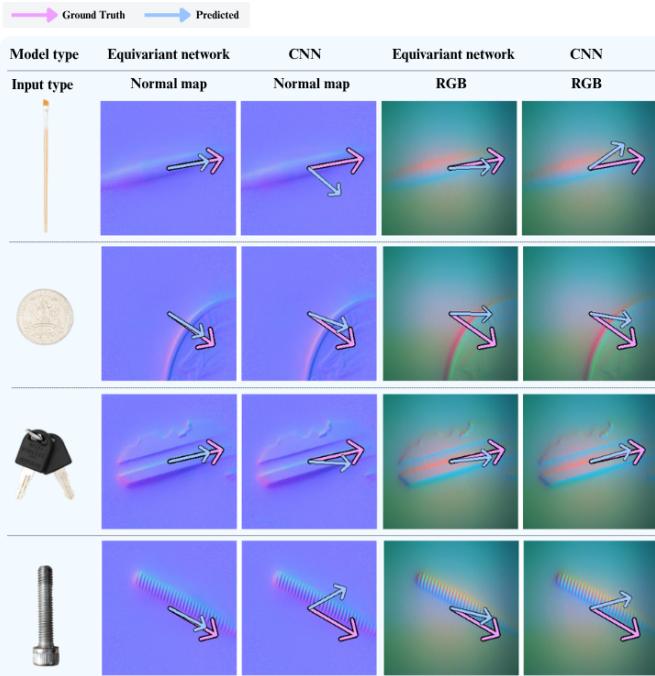


Fig. 5: Qualitative comparison of angular estimation across different model configurations.

TABLE I: **Ablation study on model design and input representation.** Angular estimation errors (in degrees) across four objects, with each row showing a variant of the model with one component removed.

Configuration	Brush	Coin	Keys	Screw	Mean
Full Model (Ours)	2.7	2.4	3.0	3.7	2.9
No Normal Map	29.2	26.3	30.4	23.8	27.4
No Equivariance	10.5	19.6	13.3	10.5	13.5
No Data Augmentation	9.3	12.5	7.6	8.3	9.4

four representative objects: *Brush*, *Coin*, *Key*, and *Screw*. For each object, ϕ is trained on a single object orientation with data augmentation and tested on 100 randomly sampled orientations, which are kept consistent across all variants for fair comparison. The full model is compared against three modified versions.

- **No Normal Map:** The model takes raw tactile RGB images as input instead of reconstructed surface normal maps.
- **No Equivariance:** All SO(2)-equivariant layers are replaced with standard conv layers, while keeping the same backbone architecture and data augmentation.
- **No Data Augmentation:** The model is trained without applying any rotational data augmentation techniques.

As shown in Table I, the full model with all components achieves a mean angular estimation error of only 2.9° , indicating its high precision. Qualitative results across different configurations are visualized in Figure 5, highlighting the improvement in angular alignment achieved by our full model.

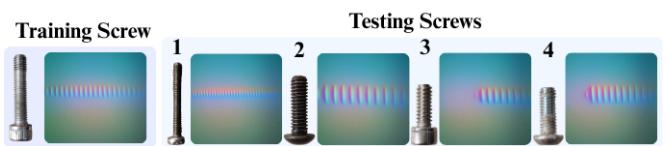


Fig. 6: **Screw geometries for generalization evaluation.** Left: training screw. Right: four novel test screws with diverse thread patterns and textures.

TABLE II: **Generalization to unseen screw geometries.** Angular errors (in degrees) when trained on a single screw type and tested on four novel thread patterns.

Method	Training	Test 1	Test 2	Test 3	Test 4
Ours	3.7	4.6	5.3	4.1	5.6

Replacing the normal map with raw RGB tactile inputs causes the most significant degradation, increasing the mean error to approximately 27° . This highlights that normal-map representations are essential for capturing geometry-consistent rotation cues and accurately reflecting object orientation. Removing SO(2)-equivariance in the network while retaining the normal map increases the mean error to about 14° , confirming that the equivariant structure further improves angular estimation accuracy. Finally, to assess the role of data augmentation, we train the equivariant network without rotation augmentation. This yields an error of 9.4° on average, which is slightly higher than the full model but still lower than the non-equivariant baseline trained with augmentation. This result suggests that while data augmentation provides a benefit, the primary performance gains arise from the model’s equivariant architecture and the use of normal-map inputs.

B. Robustness to Input Variations

We evaluate the generalization capability of our model by training the network ϕ on tactile data from a single screw instance and testing it on several unseen screws with similar but distinct geometries (Figure 6). Each test screw has a different thread pattern, and the model is evaluated on 100 randomly sampled object orientations per instance.

As shown in Table II, the model accurately estimates the angular residuals for these unseen screw types. Despite being trained on a single screw geometry, it maintains low prediction errors on novel instances, with only a modest increase compared to the training instance. This demonstrates a strong generalizability of our model, attributed to the equivariant structure and geometry-aware normal map representation.

VI. MANIPULATION EXPERIMENTS

A. Experiment Setup

In this section, we evaluate our method using a real robot system consisting of a UR5 robot, three Intel RealSense D455 RGBD cameras (RGB channels only), and a GelSight tactile sensor mounted at the center of the gripper, with its center

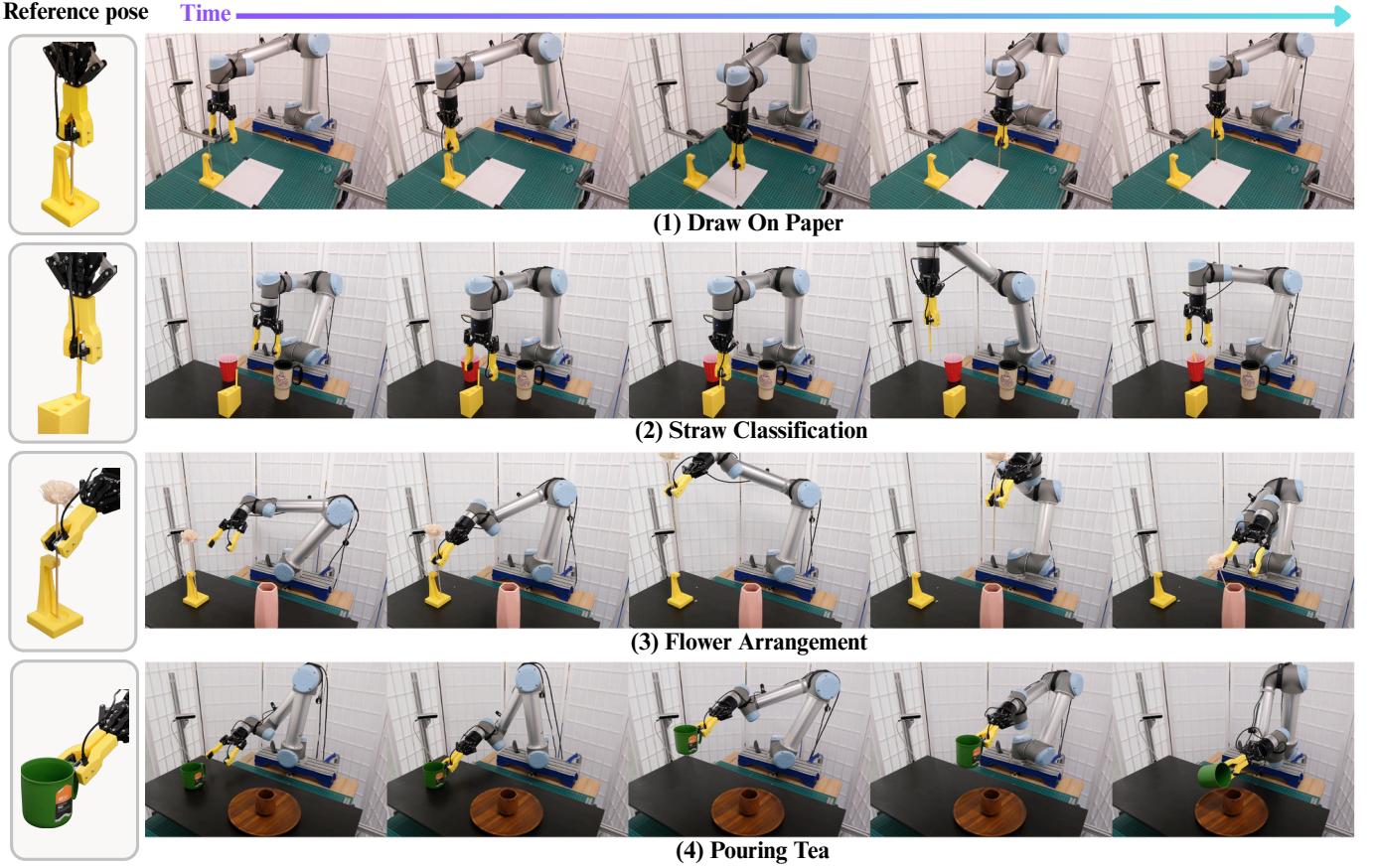


Fig. 7: **Overview of real-world object manipulation experiments.** The left column presents the reference in-hand pose for each object, while the right columns show the robot performing four daily-life tasks including drawing on paper, straw classification, flower arrangement, and pouring tea using tactile feedback to maintain reference in-hand poses.

defined as the tool center point (TCP). During data collection, demonstrations are collected using a SpaceMouse controller, with the manipulation object placed in a fixed canonical orientation. During inference, the object is initialized with a random $\text{SO}(2)$ rotation in the gripper's z -plane relative to the canonical orientation, or subjected to external perturbations, to evaluate the model's robustness to in-hand rotational variations.

We evaluate our correction module across four challenging real-world tasks, as shown in Figure 7.

- 1) **Draw on Paper.** The robot grasps a hanging paintbrush and draws a rectangle on paper.
- 2) **Flower Arrangement.** The robot grasps a flower stem and inserts it into a vase.
- 3) **Straw Classification.** The robot picks up straws with different surface textures and places them into their corresponding cups.
- 4) **Pouring Tea.** The robot grasps a teacup and performs a pouring motion.

In our experiments, we consider three evaluation conditions as shown in Figure 8. (1) No Var.: the object is evaluated under the same initial orientation distribution as the training data, without any variation. (2) Var. Init.: the object is ini-



Fig. 8: **Illustration of the three evaluation conditions in the Flower Arrangement task.** (a) No Var.: the flower stem starts upright, matching the training distribution; (b) Var. Init.: the stem is initialized with unseen tilted orientations; (c) Pert.: external human disturbance is applied during execution, inducing in-hand orientation shifts.

tialized with unseen orientations, such as the paintbrush or flower starting from a non-vertical pose. (3) Pert.: external perturbations from human are applied during inference, e.g., the object is poked by a human to alter its in-hand angle. Both Var. Init. and Pert. introduce variations in the in-hand object orientation that are not present during training. The goal is

TABLE III: **Real-world Experiment.** Success rate (%) of 4 physical experiments over 20 evaluation episodes

Method	Draw (20 demos)			Flower (10 demos)			Straw (40 demos)			Pouring (10 demos)		
	No Var.	Var. Init.	Pert.	No Var.	Var. Init.	Pert.	No Var.	Var. Init.	Pert.	No Var.	Var. Init.	Pert.
FM (RGB)	1.00	0.05	0.00	0.90	0.00	0.05	0.45	0.25	0.40	0.90	0.45	0.40
FM (RGB w/T)	0.95	0.15	0.00	0.90	0.05	0.00	0.90	0.35	0.50	1.00	0.30	0.35
EquiTac (Ours)	1.00	0.90	0.95	1.00	0.85	0.90	0.90	0.85	0.90	1.00	0.95	0.95

TABLE IV: Performance comparison under different tactile information configurations under both **No Var.** and **Pert.** conditions.

Method	Draw		Flower		Straw		Tea	
	w/o	w/	w/o	w/	w/o	w/	w/o	w/
EquiTac (RGB)	1.00	0.15	0.95	0.3	0.80	0.60	1.00	0.50
EquiTac (Normal)	1.00	0.95	1.00	0.80	0.90	0.90	1.00	0.95

to evaluate whether our correction module enables adaptive robustness under real-world out-of-distribution conditions.

B. Results

We compare our method against two baselines:

- 1) **FM (RGB):** A baseline Flow Matching policy that takes multi-view RGB images and proprioceptive input.
- 2) **FM (RGB w/T):** A Flow Matching policy that additionally incorporates tactile images alongside multi-view RGB and proprioceptive inputs.

As shown in [Table III](#), our method performs similarly as the baseline FM (RGB w/T) under no variations. However, when variations in initial orientation or human perturbations are introduced, our method significantly outperforms the baselines. For example, in the *Draw on Paper* and *Flower Arrangement* tasks, where precise contact stability and orientation control are critical, the baseline models (*FM (RGB)* and *FM (RGB w/T)*) show large drops in success rate under perturbations, decreasing from 1.00 to 0.15 and from 1.00 to 0.30, respectively. In contrast, our proposed *EquiTac* maintains a high success rate of 0.95 under both perturbation conditions. This improvement demonstrates that our tactile correction module effectively detects in-hand object pose deviations and generates corresponding compensatory adjustments in real time.

To better understand the effect of tactile representation, we conduct an ablation study comparing normal-map inputs with raw tactile RGB images in the equivariant correction network. As shown in [Table IV](#), replacing the normal map with RGB inputs causes a clear drop in performance across all tasks, especially under perturbation conditions. The reason is that tactile RGB images do not preserve rotation information, making it harder for the network to estimate and correct object orientation accurately. In contrast, normal maps provide direct cues about surface geometry and contact orientation, helping the model maintain stable in-hand poses and recover quickly from external disturbances. These results confirm that combining the equivariant structure with normal-map tactile

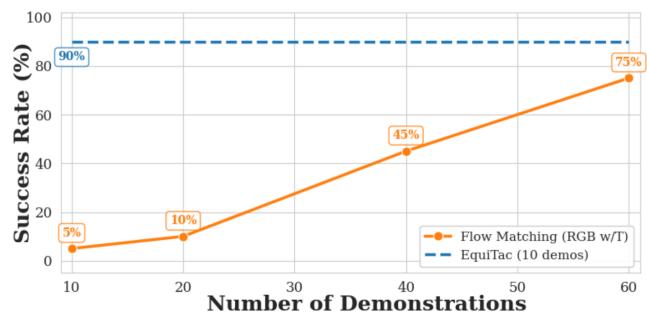


Fig. 9: **Sample efficiency comparison on the Flower Arrangement task.** The Flow Matching (RGB w/T) baseline is trained with 10–60 demonstrations, while EquiTac achieves higher success rates using only 10 demonstrations.

inputs is key to achieving consistent and reliable correction behavior.

C. Sample Efficiency Evaluation

To evaluate the sample efficiency of our approach, we augmented the training dataset, which explicitly demonstrates reorientation with variations in the initial object orientation and external perturbations during manipulation. As shown in [Figure 9](#), with only 10 demonstrations, *EquiTac* achieves a success rate of 90%, whereas the baseline reaches at most 75% even when trained with 60 demonstrations. This result demonstrates that *EquiTac* achieves higher sample efficiency by leveraging equivariant tactile representation, allowing it to learn robust correction behaviors from fewer examples.

VII. CONCLUSION AND LIMITATION

Conclusion. In this paper, we propose EquiTac, a tactile-equivariant residual correction framework for contact-rich manipulation. By reconstructing surface normal maps and leveraging SO(2)-equivariant representations, our method can precisely estimate the in-hand orientation residual estimation from a single tactile image. We further integrate this equivariant module with a flow-matching visuomotor policy, allowing real-time action correction without additional demonstrations.

Limitations. Although EquiTac demonstrated excellent performance in practical experiments, the current approach only exploits the rotational symmetry of objects relative to the gripper within the SO(2) plane. However, in many contact-rich operations, the gripped object may undergo SE(3) motion relative to the gripper, including rotations and translations. In addition, collecting multimodal real-world data introduces

high cost and limits scalability. In future work, we plan to extend EquiTac to capture SE(3) changes in tactile feedback and take advantage of tactile simulation to further improve generalization and robustness.

REFERENCES

- [1] F. Yang, C. Ma, J. Zhang, J. Zhu, W. Yuan, and A. Owens, “Touch and go: Learning from human-collected vision and touch,” *arXiv preprint arXiv:2211.12498*, 2022.
- [2] H.-J. Huang, M. Kaess, and W. Yuan, “Normalflow: Fast, robust, and accurate contact-based object 6dof pose tracking with vision-based tactile sensors,” *IEEE Robotics and Automation Letters*, 2024.
- [3] L. Fu, G. Datta, H. Huang, W. C.-H. Panitch, J. Drake, J. Ortiz, M. Mukadam, M. Lambata, R. Calandra, and K. Goldberg, “A touch, vision, and language dataset for multimodal alignment,” *arXiv preprint arXiv:2402.13232*, 2024.
- [4] J. Yang, Z.-a. Cao, C. Deng, R. Antonova, S. Song, and J. Bohg, “Equibot: Sim (3)-equivariant diffusion policy for generalizable and data efficient learning,” *arXiv preprint arXiv:2407.01479*, 2024.
- [5] Z. Zhang, Z. Xu, J. N. Lakamsani, and Y. She, “Canonical policy: Learning canonical 3d representation for equivariant policy,” *arXiv preprint arXiv:2505.18474*, 2025.
- [6] D. Wang, R. Walters, and R. Platt, “So(2)-equivariant reinforcement learning,” *arXiv preprint arXiv:2203.04439*, 2022.
- [7] H. Zhao, D. Wang, Y. Zhu, X. Zhu, O. L. Howell, L. Zhao, Y. Qian, R. Walters, and R. Platt, “Hierarchical equivariant policy via frame transfer,” in *Forty-second International Conference on Machine Learning*, 2025.
- [8] X. Zhu, D. Wang, O. Biza, G. Su, R. Walters, and R. Platt, “Sample efficient grasp learning using equivariant models,” *arXiv preprint arXiv:2202.09468*, 2022.
- [9] O. M. Andrychowicz, B. Baker, M. Chociej, R. Jozefowicz, B. McGrew, J. Pachocki, A. Petron, M. Plappert, G. Powell, A. Ray *et al.*, “Learning dexterous in-hand manipulation,” *The International Journal of Robotics Research*, vol. 39, no. 1, pp. 3–20, 2020.
- [10] G. Khandate, S. Shang, E. T. Chang, T. L. Saidi, Y. Liu, S. M. Dennis, J. Adams, and M. Ciocarlie, “Sampling-based exploration for reinforcement learning of dexterous manipulation,” *arXiv preprint arXiv:2303.03486*, 2023.
- [11] S. Suresh, G. Gallego, M. Kaess, M. Bauza, and A. Rodriguez, “Tactile slam: Real-time inference of shape and pose from planar pushing,” in *ICRA*, 2021.
- [12] P. Sodhi, M. Kaess, M. Mukadam, and S. Anderson, “Learning tactile models for factor graph-based estimation,” in *ICRA*, 2021.
- [13] M. Bauzá Villalonga, A. Rodriguez, B. Lim, E. Valls, and T. Sechopoulos, “Tactile object pose estimation from the first touch with geometric contact rendering,” in *Proceedings of CoRL (PMLR)*, 2021, pp. 1015–1029.
- [14] T. Kelestemur, D. Surovik, S. Hart *et al.*, “Tactile pose estimation and policy learning for unknown object manipulation,” in *AAMAS*, 2022.
- [15] B. Huang, Y. Wang, X. Yang, Y. Luo, and Y. Li, “3d-vitac: Learning fine-grained manipulation with visuo-tactile sensing,” *arXiv preprint arXiv:2410.24091*, 2024.
- [16] F. Nonnengießer, A. Kshirsagar, B. Belousov, and J. Peters, “In-hand object pose estimation via visual-tactile fusion,” *arXiv preprint arXiv:2506.10787*, 2025.
- [17] B. Huang, J. Xu, I. Akinola, W. Yang, B. Sundaralingam, R. O’Flaherty, D. Fox, X. Wang, A. Mousavian, Y.-W. Chao *et al.*, “Vt-refine: Learning bimanual assembly with visuo-tactile feedback via simulation finetuning,” *arXiv preprint arXiv:2510.14930*, 2025.
- [18] H. Xue, J. Ren, W. Chen, G. Zhang, Y. Fang, G. Gu, H. Xu, and C. Lu, “Reactive diffusion policy: Slow-fast visual-tactile policy learning for contact-rich manipulation,” *arXiv preprint arXiv:2503.02881*, 2025.
- [19] X. Zhu, B. Huang, and Y. Li, “Touch in the wild: Learning fine-grained manipulation with a portable visuo-tactile gripper,” *arXiv preprint arXiv:2507.15062*, 2025.
- [20] T. Johannink, S. Bahl, A. Nair, J. Luo, A. Kumar, M. Loskyll, J. A. Ojea, E. Solowjow, and S. Levine, “Residual reinforcement learning for robot control,” in *2019 international conference on robotics and automation (ICRA)*. IEEE, 2019, pp. 6023–6029.
- [21] L. Ankile, A. Simeonov, I. Shenfeld, M. Torne, and P. Agrawal, “From imitation to refinement-residual rl for precise assembly,” in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 01–08.
- [22] X. Yuan, T. Mu, S. Tao, Y. Fang, M. Zhang, and H. Su, “Policy decorator: Model-agnostic online refinement for large policy model,” *arXiv preprint arXiv:2412.13630*, 2024.
- [23] P. Dong, Q. Li, D. Sadigh, and C. Finn, “Expo: Stable reinforcement learning with expressive policies,” *arXiv preprint arXiv:2507.07986*, 2025.
- [24] Y. Jiang, C. Wang, R. Zhang, J. Wu, and L. Fei-Fei, “Transic: Sim-to-real policy transfer by learning from online correction,” *arXiv preprint arXiv:2405.10315*, 2024.
- [25] X. Xu, Y. Hou, Z. Liu, and S. Song, “Compliant residual dagger: Improving real-world contact-rich manipulation with human corrections,” *arXiv preprint arXiv:2506.16685*, 2025.
- [26] M. Kelly, C. Sidrane, K. Driggs-Campbell, and M. J. Kochenderfer, “Hg-dagger: Interactive imitation learning with human experts,” in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 8077–8083.
- [27] H. Huang, D. Wang, R. Walters, and R. Platt, “Equivariant transporter network,” *arXiv preprint arXiv:2202.09400*, 2022.
- [28] B. Hu, X. Zhu, D. Wang, Z. Dong, H. Huang, C. Wang, R. Walters, and R. Platt, “Orbitgrasp: se(3)-equivariant grasp learning,” *arXiv preprint arXiv:2407.03531*, 2024.
- [29] H. Huang, O. Howell, D. Wang, X. Zhu, R. Walters, and R. Platt, “Fourier transporter: Bi-equivariant robotic manipulation in 3d,” *arXiv preprint arXiv:2401.12046*, 2024.
- [30] Y. Qi, Y. Ju, T. Wei, C. Chu, L. L. Wong, and H. Xu, “Two by two: Learning multi-task pairwise objects assembly for generalizable robot manipulation,” in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 17383–17393.
- [31] B. Hu, H. Tian, D. Wang, H. Huang, X. Zhu, R. Walters, and R. Platt, “Push-grasp policy learning using equivariant models and grasp score optimization,” *arXiv preprint arXiv:2504.03053*, 2025.
- [32] X. Zhu, Y. Qi, Y. Zhu, R. Walters, and R. Platt, “Equat: An se (3)-equivariant multi-task transformer for open-loop robotic manipulation,” *arXiv preprint arXiv:2505.21351*, 2025.
- [33] D. Wang, M. Jia, X. Zhu, R. Walters, and R. Platt, “On-robot learning with equivariant models,” in *Proceedings of The 6th Conference on Robot Learning (CoRL)*. PMLR, 2023, pp. 1345–1354.
- [34] D. Wang, R. Walters, X. Zhu, and R. Platt, “Equivariant q learning in spatial action spaces,” in *Proceedings of the 5th Conference on Robot Learning (CoRL)*. PMLR, 2022, pp. 1713–1723.
- [35] X. Zhu, D. Wang, G. Su, O. Biza, R. Walters, and R. Platt, “On robot grasp learning using equivariant models,” *Autonomous Robots*, vol. 47, no. 8, pp. 1175–1193, 2023.
- [36] D. Wang, S. Hart, D. Surovik, T. Kelestemur, H. Huang, H. Zhao, M. Yeatman, J. Wang, R. Walters, and R. Platt, “Equivariant diffusion policy,” *arXiv preprint arXiv:2407.01812*, 2024.
- [37] B. Hu, D. Wang, D. Klee, H. Tian, X. Zhu, H. Huang, R. Platt, and R. Walters, “3d equivariant visuomotor policy learning via spherical projection,” *arXiv preprint arXiv:2505.16969*, 2025.
- [38] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter *et al.*, “pi0: A vision-language-action flow model for general robot control,” *arXiv preprint arXiv:2410.24164*, 2024.
- [39] F. Zhang and M. Gienger, “Affordance-based robot manipulation with flow matching,” *arXiv preprint arXiv:2409.01083*, 2024.
- [40] S. Wang, Y. She, B. Romero, and E. Adelson, “Gelsight wedge: Measuring high-resolution 3d contact geometry with a compact robot finger,” in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 6468–6475.
- [41] G. Cesa, L. Lang, and M. Weiler, “A program to build e (n)-equivariant steerable cnns,” in *International conference on learning representations*, 2022.