2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Built Different: Tactile Perception to Overcome Cross-Embodiment Capability Differences in Collaborative Manipulation

William van den Bogert*[1], Madhavan Iyengar*[2], Nima Fazeli[3]

***Abstract*— Tactile sensing is a widely-studied means of implicit communication between robot and human. In this paper, we investigate how tactile sensing can help bridge differences between robotic embodiments in the context of collaborative manipulation. For a robot, learning and executing force-rich collaboration require compliance to human interaction. While compliance is often achieved with admittance control, many commercial robots lack the joint torque monitoring needed for such control. To address this challenge, we present an approach that uses tactile sensors and behavior cloning to transfer policies from robots with these capabilities to those without. We train a single policy that demonstrates positive transfer across embodiments, including robots without torque sensing. We demonstrate this positive transfer on four different tactile-enabled embodiments using the same policy trained on force-controlled robot data. Across multiple proposed metrics, the best performance came from a decomposed tactile shear-field representation combined with a pre-trained encoder, which improved success rates over alternative representations.**

I. Introduction

Not all robots are built equal. Consider tasks that require force-rich physical interaction such as collaborative object carrying or object hand-over. High-end robotic platforms achieve these skills via impedance control and joint torque monitoring that enable the robot to balance position tracking and force exertion. However, a large number of industry-standard robot models produced by FANUC, ABB, and Kawasaki (among others) offer only position control with no access to joint torques. There are also affordable robotic platforms where the lack of dedicated joint-torque sensing is key for affordability. The central question that we address in this work is: "How can we train a force-rich compliant policy on high-end robot data, and then transfer such a policy to robots equipped only with inexpensive tactile feedback?"

An alternative approach to addressing the difference in capability between embodiments is to mount auxiliary force-torque sensing at the robot "wrist" and use the resulting signal as a proxy for joint torque measurements. However, this approach has a number of important issues that are remedied by tactile sensing. First, the sensor readings are contaminated by inertial artifacts from robot acceleration, but tactile sensors are robust to this movement. Second, high-quality force/torque sensors are expensive (often thousands of dollars) and fragile, whereas the open-source tactile sensors used in this work cost only about $130 in materials [1] and provide inherent passive compliance for contact-rich interaction. Third, the sensors can occupy a significant portion of the robot's payload budget, whereas modern vision-based tactile sensors are constructed with lightweight silicone pads. Fourth, summarizing an interaction with a single force vector may alias pertinent details of the task, but tactile sensing provides far richer feedback, enabling slip detection [2], in-hand pose estimation, and extrinsic contact estimation [3]. There are other potential alternatives to both tactile sensing and auxiliary force-torque sensors, i.e. force-sensing resistors or primitive strain gauges. We argue that such alternatives require additional custom hardware, whereas modern research robots increasingly integrate tactile sensors, as reflected in recent literature [4], [1], [5], [6].

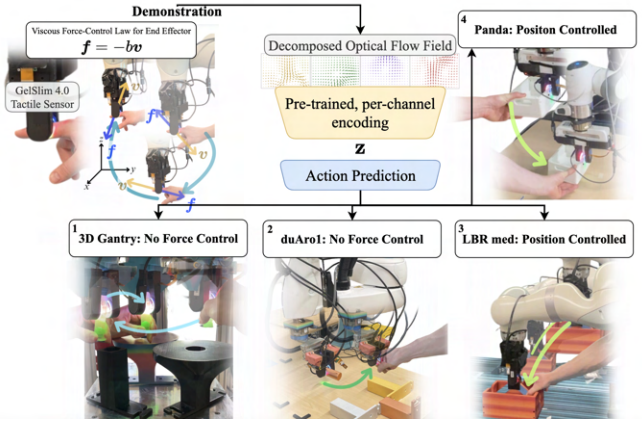

Fig. 1. Our framework for training a non-compliant robot to mimic compliant behavior during a collaborative task. The demonstration robot uses force feedback to enable a human to demonstrate the maneuvering of the grasped object throughout the workspace while tactile data is collected. A policy is trained from these demonstrations, then rolled out four unique, purely position-controlled embodiments.

A. Contributions

This paper considers tactile sensing as a means of bridging embodiment differences to enable collaborative manipulation across robotic systems. We present a method that learns compliant behavior on an impedance-capable, torque-sensing robot, then transfers a single tactile policy to tactile-equipped robots without access to impedance control and joint-torque feedback. Our key contributions are:

1) **Tactile sensing as a practical proxy for force feedback** We show that inexpensive tactile sensors can

*Equal contribution

[1] William van den Bogert is with the Mechanical Engineering Department at the University of Michigan, MI, USA willvdb@umich.edu

[2] Madhavan Iyengar is with the Computer Science Department at the University of Michigan, MI, USA miyen@umich.edu

[3] Nima Fazeli is with the Robotics Department at the University of Michigan, MI, USA nfz@umich.edu

provide sufficient feedback to enable compliant behavior typically achieved via impedance control, even on robots without torque sensing.

2) **Structured tactile representations for policy transfer.** We demonstrate that tactile representations derived from shear displacement fields, and particularly their Helmholtz-Hodge decompositions, improve data efficiency and generalization in learning our compliant policy when we test across four distinct embodiments.

B. Problem Statement

We consider two robotic systems: (1) a high-end *demonstration* robot with joint-torque sensing and impedance control, and (2) a *student* robot without torque sensing or impedance control. Our goal is to learn a policy from demonstrations on the high-end robot that enables the student robot, when equipped with tactile sensing, to exhibit similar compliant behavior during collaborative manipulation.

Workspace Assumption. We assume the Cartesian workspace of the student robot $\mathcal{W}_S$ is contained within that of the demonstration robot $\mathcal{W}_D$:

$$\mathcal{W}_S \subseteq \mathcal{W}_D \in \mathrm{SE}(3).$$

This ensures demonstrations can be collected within the student robot's reachable space.

Demonstration Robot Control. On the demonstration robot, we implement an impedance control law g, mapping end-effector velocity $\boldsymbol{v} \in \mathbb{R}^{\dim(\mathcal{W}_S)}$ and position $\boldsymbol{x} \in \mathcal{W}_S$ to wrench $\boldsymbol{f} \in \mathbb{R}^{\dim(\mathcal{W}_S)}$:

$$\boldsymbol{f} = g(\boldsymbol{v}, \boldsymbol{x})$$

For our compliance problem, we focus on a purely viscous control regime, assuming g depends only on velocity:

$$\boldsymbol{f} = g(\boldsymbol{v}) \tag{1}$$

Assuming g is invertible, we express the demonstration robot's compliant policy $\pi_D : \mathbb{R}^{\dim(\mathcal{W}_S)} \to \mathbb{R}^{\dim(\mathcal{W}_S)}$ as:

$$\boldsymbol{v} = \pi_D(\boldsymbol{f}) \tag{2}$$

where π_D maps sensed forces to end-effector velocities.

Relating Tactile Feedback to Forces. Let the vector field $\mathbf{u} \in V_{\mathbb{R}}$ represent the continuous deformation of a vision-based tactile sensor. We approximate this discretely via a shear-displacement tensor $\boldsymbol{\mathcal{U}} \in \mathbb{R}^{C \times H \times W}$, where C channels are defined over an $H \times W$ grid. Because the deformable material of the sensor exhibits roughly visco-elastic properties, we assume a time-invariant mapping $\mathcal{M} : V_{\mathbb{R}} \to \mathbb{R}^{\dim(\mathcal{W}_S)}$ of tactile feedback to end-effector forces:

$$\boldsymbol{f} = \mathcal{M}(\boldsymbol{u}) \approx \mathcal{M}'(\boldsymbol{\mathcal{U}})$$

where $\mathcal{M}' : \mathbb{R}^{C \times H \times W} \to \mathbb{R}^{\dim(\mathcal{W}_S)}$ is an approximation which is learnable from data [7], [8], [5].

Student Policy via Mapping Composition. Substituting the tactile-to-force mapping into Eq. 2 yields:

$$\boldsymbol{v} \approx \pi_D(\mathcal{M}'(\boldsymbol{\mathcal{U}}))$$

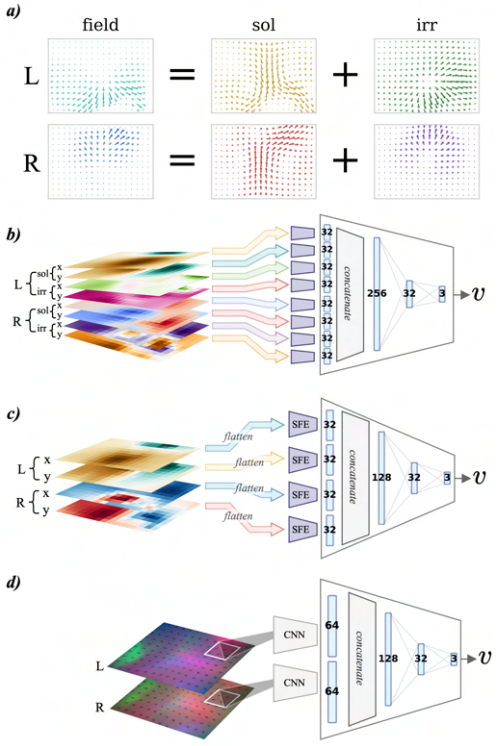

Fig. 2. (a) Example of Helmholtz-Hodge Decomposition (HHD). (b-d) The various tactile representations and their associated neural network architectures: (b) HHD, (c) shear-field, and (d) tactile image input to our model.

We define the student policy π_S as the composition of the force-to-motion policy π_D with the tactile-to-force map $\mathcal{M}'$:

$$\boldsymbol{v} \approx \pi_S(\boldsymbol{\mathcal{U}}) \tag{3}$$

where $\pi_S = \pi_D \circ \mathcal{M}'$.

By approximating π_S via supervised imitation learning of a dataset of a tactile-equipped *demonstration robot*'s behavior, a tactile-equipped *student robot* can perform π_S without impedance-control or related capabilities.

This formulation highlights why vision-based tactile sensing can serve as a practical proxy for force feedback, enabling compliant policy transfer across embodiments lacking torque sensing. We implement this formulation with tactile sensing fingertips, though it can be applied to other tactile anatomies (i.e. robot skin) as well.

Transferring our policy across embodiments presents a series of challenges. Policy success on unseen robotic systems is inhibited by the diverse set of properties (base stiffness, grasp strength, control architecture, environmental conditions, grasped objects) exhibited by embodiments. We aim to overcome these challenges by selecting a structured representation for the shear-displacement tensor $\boldsymbol{\mathcal{U}}$ based on the Helmholtz-Hodge decomposition.

II. Related Works

A. Tactile Perception

Tactile information is critical for humans and robots alike. Studies of human tactile perception demonstrate its role in

force estimation and fine motor control [9], [10]. In robotics, tactile sensing has long been used as an input for wrench estimation [11], including with GelSight [12], [13] and modified GelSight [6] sensors. To the best of our knowledge, the role of tactile sensing in compliant robotics remains underexplored, but there are works in this research domain. [14], [15] used tactile sensing to detect object compliance. [16], [17] demonstrated compliant grasping control with tactile-equipped end-effectors, as a means to improve grasp safety and quality. However, these works are limited to end-effector compliance or force perception, without extension to whole-robot compliant behavior or cross-embodiment collaborative task transfer, for which we learn a policy in this work.

B. Tactile Policy Learning

Whole-robot tactile policies have been learned in both reinforcement learning (RL) and behavior cloning (BC) frameworks. For a robot tasked with peg-in-hole insertion, [18] shows that an RL policy succeeds most frequently with multimodal vision and tactile (in the form of force-torque) feedback. These modalities enable deformable object manipulation policies as well [19], [20]. Similar RL work was later implemented by [4] with vision-based tactile sensors, indicating that a force-rich task can succeed with just fingertip tactile feedback.

Our work focuses on tactile policies generated from BC. Behavior cloning state-of-the-art Diffusion Policy [21] has been integrated with tactile feedback before [22], [23]. Diffusion Policy is well-suited to intelligently generate trajectories of end-effector poses from multimodal feedback. Our work requires a more streamlined architecture, in which a single-step velocity action is generated in response to tactile feedback. This choice reflects our goal of cloning the behavior of the demonstration robot's viscous control regime with only vision-based tactile sensing—no impedance/admittance control, force-sensing resistors, joint current/torque monitoring, or other force sensing solutions.

C. Structured Representations

We use the GelSlim 4.0 vision-based tactile sensor, which captures high-dimensional deformation signals, but produces state distributions sensitive to lighting and grasp variability. To improve data efficiency, previous work has distilled raw tactile images into structured representations, such as a discrete field of shear displacements [2]. This field can be decomposed using Helmholtz-Hodge decomposition (HHD) [24], which has precedent in tactile sensing. [25] uses HHD as a constraint to learn a representation of the tactile shear field. In contrast, our work uses HHD to decompose the shear field obtained from optical flow measurements. [8], [5], and [7] use a similar method to perceive a contact wrench. Our work bypasses the force estimation step and learns a representation of the HHD to enable our collaborative policy. Since this representation mitigates sensitivity to lighting and grasp variability, the same collaborative policy works across tactile-equipped embodiments.

D. Cross-Embodiment Transfer

Prior work has explored cross-embodiment policy transfer using latent space alignment [26] and raw sensory alignment [27]. Transfer of vision-based policies has also been improved by in-painting the demonstration robot into the student robot's scene [28].

In contrast, we enable zero-shot transfer of a collaborative tactile policy by using the learned HHD representation as a shared state space between the demonstration and student robots. Importantly, this approach does not require any additional data collection on the student robot before transfer.

III. Methods

Our framework is intended to mimic the compliant behavior of impedance-controlled robots using BC and a structured tactile representation.

We focus on manipulators with parallel-jaw grippers where end-effector forces $\boldsymbol{f}$ arise from human interaction with a grasped object. While our method uses fingertip tactile sensing, it could also extend naturally to other tactile modalities (e.g., robot skin), though we leave this to future work.

A. Tactile Representations

To learn a tactile-informed policy that (1) mimics compliant behavior and (2) generalizes well across embodiments, we require a structured representation of sensor feedback.

The GelSlim 4.0 sensor [1] we use in this work enables high-resolution tactile sensing in the form of raw RGB images $\mathbf{I}$. To derive the shear-displacement tensor from these images, we used the **open-cv2** Python library. A function $\mathbf{Flow}$ calculates the optical flow components from the undeformed image $\mathbf{I}_0$ to the deformed image $\mathbf{I}_t$. We then use the Python library **naturalHHD** [29] to estimate the solenoidal (sol) and irrotational (irr) components of the Helmholtz-Hodge decomposition (HHD) [24], where:

$$\text{sol} + \text{irr} = (\text{x}, \text{y}) = \mathbf{Flow}(\mathbf{I}_0, \mathbf{I}_t)$$

where x and y are the horizontal and vertical components of the vectors, respectively. This decomposition, visualized in Fig. 2a, separates divergence-free and curl-free motion patterns, which we hypothesize as decoupling normal and shear force information further than raw optical flow.

For a parallel-jaw gripper with two GelSlim sensors, we obtain two decompositions (L, R), each with two fields and two components, resulting in an 8-channel tensor $\boldsymbol{\mathcal{U}} \in \mathbb{R}^{8\times13\times18}$ (see Sec. I-B).

B. Scalar Field Encoder

Each channel of $\boldsymbol{\mathcal{U}}$ is processed by a scalar field encoder (SFE), shown in Fig. 2b. This fully connected network is pre-trained on reconstruction of channels from a tactile manipulation dataset. The encoder outputs an embedding $\mathbf{z}_i \in \mathbb{R}^{32} = \text{SFE}(\boldsymbol{\mathcal{U}}_i)$ for $i \in \{0 \cdots 7\}$. Fig. 2b shows the reconcatenation of these fields post-encoding, providing a latent representation of the entire HHD $\mathbf{z}_{\boldsymbol{\mathcal{U}}} \in \mathbb{R}^{256}$.

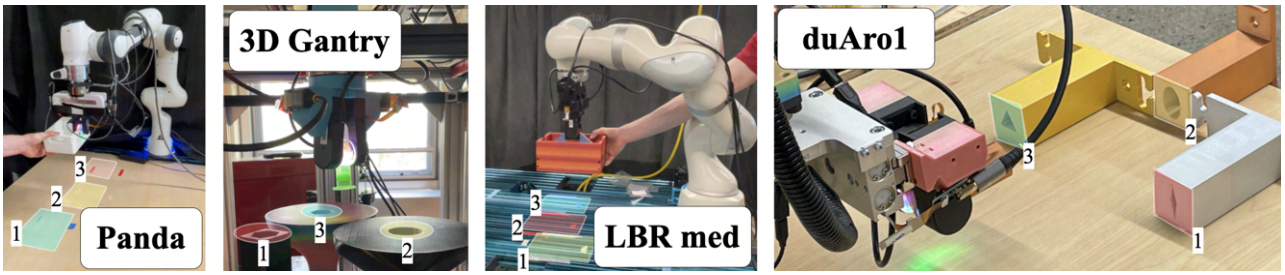

Fig. 3. Policies on these embodiments were evaluated on the task of maneuvering grasped objects to the three highlighted goal locations.

C. Data Collection of Compliant Behavior

To collect demonstrations, we exploit the impedance-control capabilities of the demonstration robot to generate compliant behaviors. Following Eq. 1, we implement a viscous control regime where the end-effector exerts a wrench $\boldsymbol{f} \in \mathbb{R}^3$ opposite to its velocity $\boldsymbol{v} \in \mathbb{R}^3$:

$$\boldsymbol{f} = -\mathbf{B}\boldsymbol{v} \tag{4}$$

where $\mathbf{B} > 0$ is a diagonal damping matrix. Intuitively, this models the robot as resisting motion proportionally to velocity, allowing it to "yield" naturally to human guidance. With this control active, we collect the synchronized dataset $\mathcal{D} = \{\boldsymbol{\mathcal{U}}, \boldsymbol{v}\}$ from the tactile-equipped demonstration robot.

D. Behavior Cloning Architecture

Given $\mathcal{D}$, we solve a supervised imitation learning problem to find $\pi_S : \mathbb{R}^{C\times H\times W} \to \mathbb{R}^3$ from Eq. 3.

We use a lightweight 3-layer multilayer perceptron (MLP) architecture, shown in Fig. 2, to map the embedding of tactile state $\mathbf{z}_{\boldsymbol{\mathcal{U}}}$ to velocity $\boldsymbol{v}$:

$$\boldsymbol{v} \approx \text{MLP}(\mathbf{z}_{\boldsymbol{\mathcal{U}}}) \tag{5}$$

We train this MLP on L2-norm regression to $\boldsymbol{v}$ given paired $\mathbf{z}_{\boldsymbol{\mathcal{U}}}$ in the dataset $\mathcal{D}$. The combination of the MLP and the concatenated SFE outputs is our estimate for π_S:

$$\boldsymbol{v} \approx \pi_S(\boldsymbol{\mathcal{U}}) \approx \text{MLP}(\text{SFE}(\boldsymbol{\mathcal{U}}_0) \oplus \cdots \oplus \text{SFE}(\boldsymbol{\mathcal{U}}_7)) \tag{6}$$

After training, π_S predicts the velocity that imitates π_D.

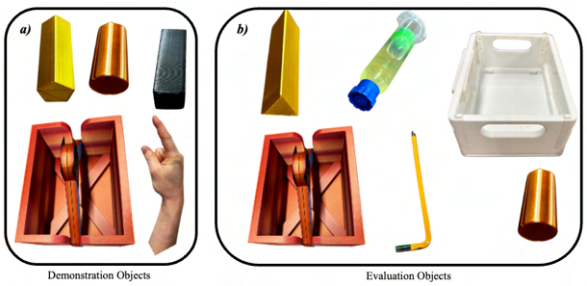

Fig. 4. Objects grasped by the robot during (a) demonstration collection and (b) evaluation.

IV. Experiments

In this section we describe our experimental setup, baseline methods, and evaluation metrics. With our experiments, we aim to show that (1) compliant behavior can in fact be enabled entirely through vision-based tactile feedback, serving as an effective proxy for joint torque sensing. (2) the compliant behavior we learn generalizes across diverse robot embodiments and grasped objects. (3) this generalization is best achieved through structured tactile representations, such as the HHD proposed in Sec. III-A.

A. Training

Data Collection. Training demonstrations were collected on a 7-DOF Kuka LBR Med manipulator constrained to a 3D translational workspace. The demonstration robot executed the viscous control regime from Eq. 4, while grasping one of the objects shown in Fig. 4a. A human operator manipulated the grasped object across the workspace, aiming to generate a wide range of velocities in both magnitude and direction. We collected 62 episodes, each using a different object, resulting in roughly 27,000 $(\boldsymbol{\mathcal{U}}, \boldsymbol{v})$ pairs.

Dataset Filtering. Many samples contained near-zero velocities, which can bias learning toward static actions. We filtered out data with $||\boldsymbol{v}|| < 1.5\times\text{median}(||\boldsymbol{v}||)$, resulting in roughly 3700 pairs. Our dataset had $\text{median}(||\boldsymbol{v}||) = 0.09$ m/s.

Data Augmentation. We augmented the filtered dataset by adding Gaussian noise ($\sigma = 0.1$) to $\boldsymbol{\mathcal{U}}$, expanding the dataset to roughly 15,000 usable samples.

Dataset Split. We used an 80–10–10 train–validation–test split such that around 13,000 data points were seen by the network during training.

B. Evaluation

1) Student Robot Embodiments: We evaluated trained policies on 4 student robots (Fig. 3), each differing in morphology and control capabilities:

1) **Panda:** a small manipulator set to position-control mode, cut off from native force monitoring.
2) **3D Gantry:** a 3D prismatic gantry adapted from a 3D printer, with no native force-control capabilities.
3) **LBR Med:** the same manipulator as the demonstration robot, though cut off from native force monitoring.
4) **duAro1:** a manipulator without native force monitoring.

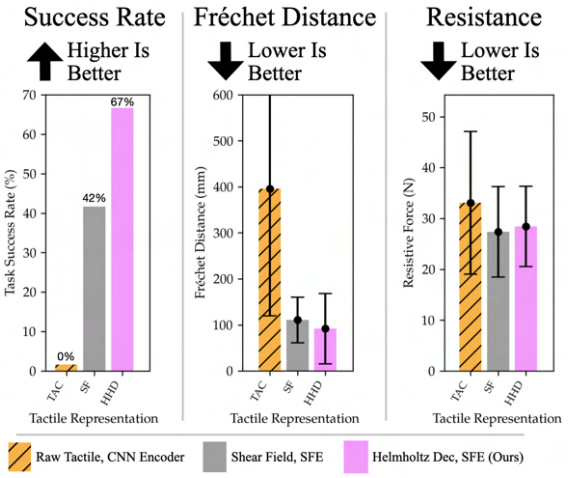

Fig. 5. Focusing on policies with pre-trained encoders only, this graph shows the effect of different tactile representations on compliant behavior over all test embodiments. Performance is improved when HHD is used over other representations. Error bars represent 1σ.

2) Goal-Reaching Tasks: In each trial, a student robot grasped an object from Fig. 4b and the human operator guided the object toward 1 of 3 goal locations (Fig. 3). All 3 goals were tested independently for each embodiment.

3) Metrics for Evaluation: We evaluated our method on six different metrics in order to holistically determine which policies perform best.

1) **Fréchet distance from straight-line path** was computed between the executed trajectory $\boldsymbol{x}_{1\ldots N}$ and the straight-line path from start pose $\boldsymbol{x}_s$ to goal $\boldsymbol{x}_g$, intending to capture global similarity in shape and ordering.
2) **Projected** (agnostic to x, y, or z) **root-mean-square error (RMSE)** was also computed between $\boldsymbol{x}_{1\ldots N}$ and the straight-line path to assess local deviations:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\|(\boldsymbol{x}_i - \boldsymbol{x}_s)\times(\boldsymbol{x}_g - \boldsymbol{x}_s)\|}{\|\boldsymbol{x}_g - \boldsymbol{x}_s\|}\right)^2}$$

Note that an immobile robot ($\boldsymbol{x}_i = \boldsymbol{x}_s$) gets RMSE $= 0$, so we ignore this metric if the goal is not reached.
3) **Path inefficiency** was defined as the ratio of the executed path length to the straight-line distance, where lower values indicate more efficient execution. This was also recorded for successful runs only. Otherwise, a student robot which does not move gets a perfect score.
4) **Resistive force** was measured as the average magnitude of the end-effector force applied over the entire trajectory, serving as a proxy for human effort and recorded via an external ATI Gamma force-torque sensor.
5) **Task duration:** was recorded for successful runs only, as another measure of efficiency.
6) **Success rate:** the fraction of trials where the student robot's final position falls near $\boldsymbol{x}_g$ (illustrated in Fig. 3), with a tolerance of 15% (visualized in Fig. 7) of the workspace size (via longest dimension of 3D rectangular prism) of the student robot.

As these metrics are non-standard, we aim to make a recommendation of a subset of metrics based on our results.

C. Baselines

We evaluated a suite of model variants spanning different tactile representations, encoders, and training strategies in order to thoroughly baseline our proposed method.

1) Tactile Representations: We investigated 3 methods of representing tactile information prior to its presentation to any learning algorithm.

- **Raw Tactile Image (TAC)** baselines test whether high-dimensional GelSlim images, without preprocessing, are sufficient for learning transferable collaborative behavior
- **Shear-Field (SF)** baselines apply the optical flow function as discussed in Sec. III-A, without the HHD.
- **Helmholtz-Hodge decomposition (HHD)** baselines apply the decomposition as discussed in Sec. III-A.

2) Encoders: We investigated 2 encoder architectures for each tactile representation: a ResNet-18 and another using a neural network with a smaller number of parameters.

- **ResNet-18 (ResNet)** is a commonly used residual learning encoder for vision [30]. We employ the 18-layer version.
- *Smaller Encoder #1:* **3-layer Convolutional Neural Network (CNN)** is a custom low-dimensional convolutional neural network we apply to the TAC representation only, since this representation is too high dimensional for the fully-connected scalar field encoder (SFE).
- *Smaller Encoder #2:* **Scalar Field Encoder (SFE)** is our proposed scalar field encoder discussed in Sec. III-B, which we apply to both SF and HHD representations.

3) Training Strategies: We investigated 2 training strategies for our behavior cloning architecture.

- **Pre-training (PT)** involves training the above encoders on reconstruction of their given tactile representation, given a dataset with a larger distribution of tactile information than the imitation learning dataset $\mathcal{D}$. The pre-training dataset is collected from a variety of tactile manipulation scenarios.

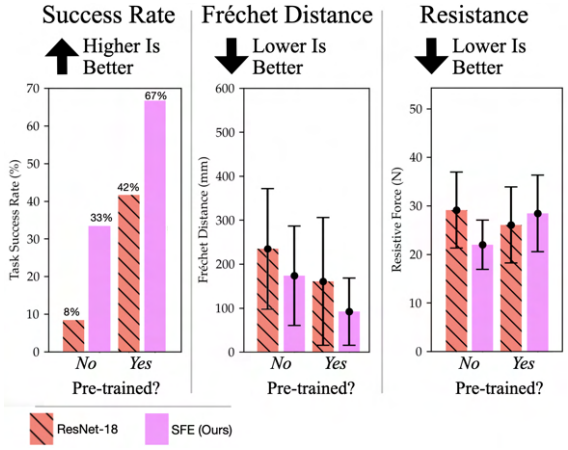

Fig. 6. Focusing on HHD policies only, this graph shows the effect of different tactile encoders on compliant behavior over all test embodiments. Performance is improved when our pre-trained SFE is used over other variants. Error bars represent 1σ.

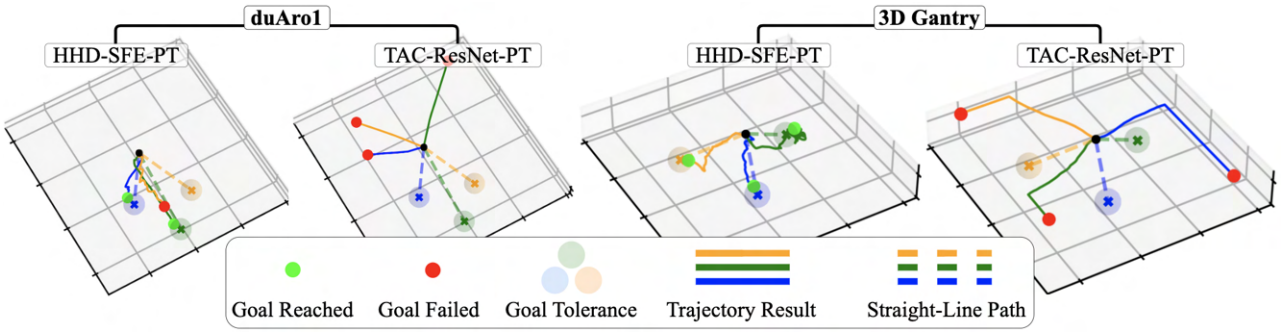

Fig. 7. Example trajectories from our results.

- **End-to-end training (E2E)** involves training the above encoders simultaneously with the MLP that predicts velocity actions, given only the imitation learning dataset $\mathcal{D}$.

Combining 3 representations, 2 encoders, and 2 training strategies yields 12 model variants, including our proposed approach (HHD + SFE + PT). Each policy was trained on the same dataset $\mathcal{D}$ and evaluated on all 4 robot embodiments. Each of the 3 goal-reaching tasks used a model trained with 1 of 3 independent random seeds. Thus, each of the 12 model variants received 12 independent evaluations.

V. Results

This section reports performance across all model variants and robot embodiments described in Sec. IV. To simplify references, we abbreviate model variants as:

{Tactile Representation}-{Encoder}-{Training Strategy}

where:

1) **Tactile Representation** is either **HHD** (Helmholtz-Hodge decomposition), **TAC** (raw GelSlim images) or **SF** (non-decomposed shear-field).
2) **Encoder** is either **SFE** (our scalar field encoder), **ResNet** (ResNet-18), or **CNN** (a 3-layer CNN).
3) **Training Strategy** is either **PT** (pre-trained encoder), or **E2E** (simultaneous training of encoder and MLP)

For example, HHD-SFE-PT is our proposed method. This section contains results for each model variant based on evaluations described in Sec. IV on the four different robot embodiments: Panda, 3D Gantry, LBR Med, and duAro1.

A. Key Findings

These results aggregate performance across all embodiments to isolate the improved effect of (1) the HHD representation over other choices and (2) pre-training the HHD SFE encoder over E2E and ResNet variants.

1) Effect of Tactile Representation: Fig. 5 compares models using different tactile representations, all paired with their smallest viable pre-trained encoder (SFE for HHD and SF, CNN for TAC). Here we present the metrics which are valid for both successful and unsuccessful trials, with success defined as in Sec. IV-B: success rate, Fréchet distance, and resistive force. HHD generates the highest success rate at 67% over SF (42%) and TAC (0%). On Fréchet distance, HHD performs slightly better than SF on average, while TAC consistently underperforms. The resistive-force metric does not reveal strong trends across methods and is therefore less informative for evaluating policy quality. These results suggest that structured representations—particularly HHD—are better suited for this problem than raw GelSlim images.

2) Effect of Encoder and Training Strategy: Fig. 5 isolates the HHD representation to compare encoders and training strategies. We again present the metrics which are valid for both successful and unsuccessful trials. Among the HHD-based models, our pre-trained SFE achieves the highest success rate (67%) compared to E2E SFE (42%) and ResNet-18 variants (pre-trained 33%, E2E 8%). Fréchet distance follows a similar trend, favoring the pre-trained SFE. Together, these findings indicate that lightweight, task-specific

TABLE I

Fréchet Distance From Straight-Line Path Results (mm) ↓ **Lower is better.** — Mean ± Std. Deviation Over 3 Goal Locations

Method	LBR Med	Panda	3D Gantry	duAro1
TAC-ResNet-E2E	989 ± 334	190 ± 66.7	286 ± 36.3	351 ± 127
TAC-CNN-E2E	521 ± 268	175 ± 27.8	212 ± 52.3	209 ± 76.5
TAC-ResNet-PT	922 ± 355	199 ± 54.6	268 ± 64.9	435 ± 153
TAC-CNN-PT	813 ± 197	166 ± 81.6	344 ± 59.8	260 ± 89.5
SF-ResNet-E2E	563 ± 264	102 ± 23.5	273 ± 67.6	106 ± 27.2
SF-SFE-E2E	379 ± 237	**35.0 ± 12.3**	99.3 ± 52.2	87.5 ± 49.0
SF-ResNet-PT	186 ± 97.7	80.5 ± 75.1	194 ± 125	211 ± 73.5
SF-SFE-PT	132 ± 69.9	103 ± 57.4	**91.1 ± 17.5**	116 ± 17.6
HHD-ResNet-E2E	360 ± 137	103 ± 63.7	257 ± 63.2	218 ± 118
HHD-SFE-E2E	267 ± 173	138 ± 14.5	133 ± 36.8	155 ± 88.4
HHD-ResNet-PT	295 ± 165	38.9 ± 9.26	247 ± 81.0	61.8 ± 22.7
HHD-SFE-PT	**76.8 ± 10.9**	38.4 ± 8.93	177 ± 106	**74.4 ± 33.1**

encoders benefit from pre-training on reconstruction tasks before behavior cloning, while larger generic encoders such as ResNet-18 appear less effective for this structured input.

B. Comprehensive Results

Figure 8 summarizes the 6 evaluation metrics from Sec. IV-B across the 12 model variants, averaged over all embodiments. Among these, **Fréchet distance** and **success rate** emerge as the most informative, providing clear indications of trajectory quality and task completion, respectively.

In contrast, **resistive force** shows little variation between models and correlates weakly with other measures of performance, making it less useful for differentiating methods. Finally, the remaining metrics—**RMSE**, **path inefficiency**, and **task duration**—offer complementary insights but are less reliable when success rates are low, since averages are then based on fewer trials.

C. Embodiment-Specific Results

Finally, we present embodiment-specific breakdowns of our two most informative metrics: Fréchet distance (Table I) and success rate (Table II). Fréchet distance results align with qualitative trajectory observations, some of which are visualized in Fig. 7. TAC is excluded from Table II due to no successes. Our proposed **HHD-SFE-PT** model ranks among the most successful on every robot.

These results demonstrate that the learned policy generalizes well across embodiments with different morphologies and control interfaces, even though no student-robot data were used during training.

VI. Discussion and Limitations

This work demonstrates that vision-based tactile sensing can enable compliant, collaborative manipulation on position-controlled robots without joint-torque sensing or impedance control. By training on demonstrations from an impedance-enabled robot and leveraging a structured tactile representation—the Helmholtz-Hodge decomposition (HHD) of shear displacement fields—we achieve zero-shot policy transfer across four diverse robot embodiments. Our results show that tactile can act as a practical proxy for force, enabling policies that mimic a viscous control regime (Eq. 1).

Among the tactile representations evaluated, HHD paired with our pre-trained scalar field encoder (SFE) achieved the highest overall success rates and lowest trajectory deviation. We hypothesize that this benefit arises because HHD dealiases task-relevant information such as normal and shear forces, while filtering out environmental and sensor-specific artifacts present in raw tactile images. Nevertheless, other SFE-based policies performed competitively, suggesting the encoder itself provides a broadly useful low-dimensional tactile representation. Multimodal BC frameworks, such as Diffusion Policy [21], could benefit from incorporating this encoder for contact-rich tasks.

We also evaluated six potential metrics and found that success rate and Fréchet distance were the most informative for assessing policy quality. Resistive force showed low discriminative power, as both failed and successful policies exhibited similar values, though reducing human-applied force remains an important goal.

Several limitations remain. First, our experiments focus on tasks solvable entirely through human guidance; richer collaborative scenarios where robots balance human intent with autonomous objectives are not addressed here. Second, we restrict compliance to translation only, leaving rotational compliance for future work which may apply reasoning from Sec. I-B toward an expanded Eq. 4. Third, demonstrations and evaluations were conducted by expert operators, which may limit generalization; incorporating broader user studies would provide stronger evidence of usability.

Finally, while our method enables compliance-like behavior on position-controlled robots, resistive forces remain higher than under native impedance control, reflecting the inherent limitations of indirect force feedback. This high resistance may be an artifact of pseudo-velocity control via commanding Δ-poses. Direct control of end-effector velocity should be implemented in future work.

Overall, our results establish vision-based tactile sensing and structured tactile representations as a promising proxy for force-torque feedback in enabling compliant, cross-embodiment collaborative manipulation, while highlighting avenues for improving policy robustness, tactile sensitivity, and broader task generalization.

References

[1] A. Sipos, W. van den Bogert, and N. Fazeli, "Gelslim 4.0: Focusing on touch and reproducibility," 2024.

[2] Y. Zhang, Z. Kan, Y. A. Tse, Y. Yang, and M. Y. Wang, "Fingervision tactile sensor design and slip detection using convolutional LSTM network," *CoRR*, vol. abs/1810.02653, 2018.

TABLE II

Task Success Rate Results (%) ↑ **Higher is better** — Mean Over 3 Goal Locations

Method	LBR Med	Panda	3D Gantry	duAro1
SF-ResNet-E2E	0.00	0.00	0.00	0.00
SF-SFE-E2E	33.33	66.67	**100.00**	0.00
SF-ResNet-PT	33.33	66.67	66.67	0.00
SF-SFE-PT	**66.67**	66.67	33.33	0.00
HHD-ResNet-E2E	0.00	0.00	33.33	0.00
HHD-SFE-E2E	33.33	**100.00**	0.00	0.00
HHD-ResNet-PT	33.33	0.00	**100.00**	**33.33**
HHD-SFE-PT	**66.67**	66.67	**100.00**	**33.33**

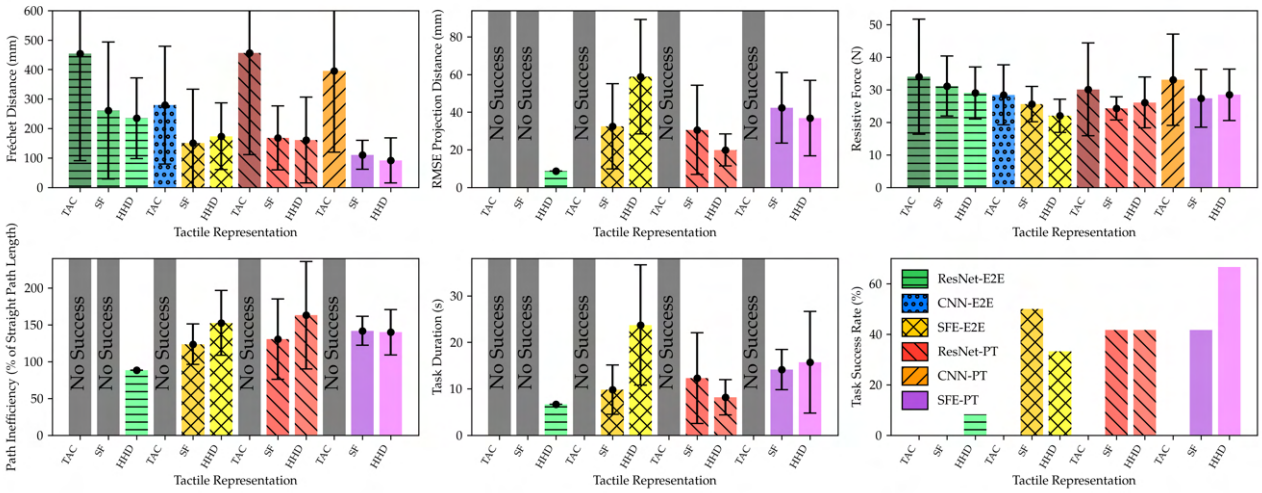

Fig. 8. All metrics evaluated for each method over all rollouts on all embodiments. A gray "No Success" bar indicates the lack of valid results given that there where no rollouts where the goal location was reached. Error bars represent 1σ.

[3] S. Kim and A. Rodriguez, "Active extrinsic contact sensing: Application to general peg-in-hole insertion," *CoRR*, vol. abs/2110.03555, 2021.

[4] S. Dong, D. K. Jha, D. Romeres, S. Kim, D. Nikovski, and A. Rodriguez, "Tactile-rl for insertion: Generalization to objects of unknown geometry," *CoRR*, vol. abs/2104.01167, 2021.

[5] E. Aucone, C. Sferrazza, M. Gregor, R. D'Andrea, and S. Mintchev, "Optical tactile sensing for aerial multicontact interaction: Design, integration, and evaluation," *IEEE Transactions on Robotics*, vol. 41, pp. 364–377, 2025.

[6] W. Li, M. Wang, J. Li, Y. Su, D. K. Jha, X. Qian, K. Althoefer, and H. Liu, "L^3 f-touch: A wireless gelsight with decoupled tactile and three-axis force sensing," *IEEE Robotics and Automation Letters*, vol. 8, no. 8, pp. 5148–5155, 2023.

[7] Y. Zhang, Z. Kan, Y. Yang, Y. A. Tse, and M. Y. Wang, "Effective estimation of contact force and torque for vision-based tactile sensors with helmholtz–hodge decomposition," *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 4094–4101, 2019.

[8] G. Zhang, Y. Du, H. Yu, and M. Y. Wang, "Deltact: A vision-based tactile sensor using a dense color pattern," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 10778–10785, 2022.

[9] L. A. Jones, *The Control and Perception of Finger Forces*, pp. 99–122. Cham: Springer International Publishing, 2014.

[10] A. M. Gordon and J. F. Soechting, "Use of tactile afferent information in sequential finger movements," *Experimental Brain Research*, vol. 107, pp. 281–292, 2004.

[11] Z. Zhou, R. Zuo, B. Ying, J. Zhu, Y. Wang, X. Wang, and X. Liu, "A sensory soft robotic gripper capable of learning-based object recognition and force-controlled grasping," *IEEE Transactions on Automation Science and Engineering*, vol. 21, no. 1, pp. 844–854, 2024.

[12] W. Yuan, S. Dong, and E. H. Adelson, "Gelsight: High-resolution robot tactile sensors for estimating geometry and force," *Sensors*, vol. 17, no. 12, p. 2762, 2017.

[13] Z. Lu, Z. Liu, X. Zhang, Y. Liang, Y. Dong, and T. Yang, "3d force identification and prediction using deep learning based on a gelsight-structured sensor," *Sensors and Actuators A: Physical*, vol. 367, p. 115036, 2024.

[14] A. Fath El Bab, M. Eltaib, M. Sallam, and O. Tabata, "Tactile sensor for compliance detection," *Sensors and Materials*, vol. 19, pp. 165–177, 06 2007.

[15] M. Burgess, "Learning object compliance via young's modulus from single grasps with camera-based tactile sensors," 2024.

[16] L. Jentoft, Q. Wan, and R. Howe, "Limits to compliance and the role of tactile sensing in grasping," *Proceedings - IEEE International Conference on Robotics and Automation*, pp. 6394–6399, 09 2014.

[17] H. Yang, J. Liu, W. Liu, W. Liu, Z. Deng, Y. Ling, C. Wang, M. Wu, L. Wang, and L. Wen, "Compliant grasping control for a tactile self-sensing soft gripper," *Soft Robotics*, vol. 11, pp. 230–243, Apr. 2024. Epub 2023 Sep 28.

[18] 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*, vol. 36, no. 3, pp. 582–596, 2020.

[19] Y. Wi, A. Zeng, P. Florence, and N. Fazeli, "Virdo++: Real-world, visuo-tactile dynamics and perception of deformable objects," 2022.

[20] M. V. der Merwe, Y. Wi, D. Berenson, and N. Fazeli, "Integrated object deformation and contact patch estimation from visuo-tactile feedback," 2023.

[21] 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.

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

[23] A. George, S. Gano, P. Katragadda, and A. B. Farimani, "Visuo-tactile pretraining for cable plugging," 2024.

[24] H. v. Helmholtz, "Über integrale der hydrodynamischen gleichungen, welche den wirbelbewegungen entsprechen.," 1858.

[25] Z. Zhang, H. Yang, and Z. Yin, "Gelflow: Self-supervised learning of optical flow for vision-based tactile sensor displacement measurement," in *Intelligent Robotics and Applications* (H. Yang, H. Liu, J. Zou, Z. Yin, L. Liu, G. Yang, X. Ouyang, and Z. Wang, eds.), (Singapore), pp. 26–37, Springer Nature Singapore, 2023.

[26] T. Wang, D. Bhatt, X. Wang, and N. Atanasov, "Cross-embodiment robot manipulation skill transfer using latent space alignment," 2024.

[27] H. Niu, J. Hu, G. Zhou, and X. Zhan, "A comprehensive survey of cross-domain policy transfer for embodied agents," in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI '24, 2024.

[28] L. Y. Chen, K. Dharmarajan, K. Hari, C. Xu, Q. Vuong, and K. Goldberg, "MIRAGE: Cross-Embodiment Zero-Shot Policy Transfer with Cross-Painting," in *Proceedings of Robotics: Science and Systems*, (Delft, Netherlands), July 2024.

[29] H. Bhatia, V. Pascucci, and P.-T. Bremer, "The natural helmholtz-hodge decomposition for open-boundary flow analysis," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 11, pp. 1566–1578, 2014.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015.