Feel the Force:
Contact-Driven Learning from Humans

Ademi Adeniji[*1,2]   Zhuoran Chen[*3]   Vincent Liu[1]   Venkatesh Pattabiraman[1]   Siddhant Haldar[1]

Raunaq Bhirangi[1]   Pieter Abbeel[2]   Lerrel Pinto[1]

[1]New York University   [2]UC Berkeley   [3]New York University Shanghai
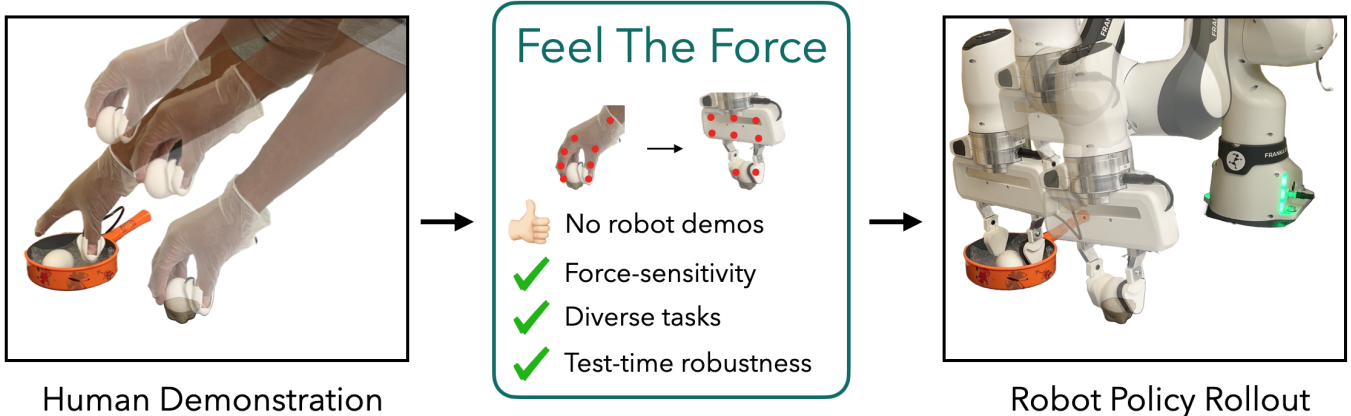
[*]Equal Contribution

Fig. 1: FTF allows zero-shot transfer of tactile human demonstrations to a Franka Robot.

Abstract—Controlling fine-grained forces during manipulation remains a core challenge in robotics. While robot policies learned from robot-collected data or simulation show promise, they struggle to generalize across the diverse range of real-world interactions. Learning directly from humans offers a scalable solution, enabling demonstrators to perform skills in their natural embodiment and in everyday environments. However, visual demonstrations alone lack the information needed to infer precise contact forces. We present FEELTHEFORCE (FTF): a robot learning system that models human tactile behavior to learn force-sensitive manipulation. Using a tactile glove to measure contact forces and a vision-based model to estimate hand pose, we train a closed-loop policy that continuously predicts the forces needed for manipulation. This policy is re-targeted to a Franka Panda robot with tactile gripper sensors using shared visual and action representations. At execution, a PD controller modulates gripper closure to track predicted forces—enabling precise, force-aware control. Our approach grounds robust low-level force control in scalable human supervision, achieving a 77% success rate across 5 force-sensitive manipulation tasks. Code and videos are available at https://feel-the-force-ftf.github.io.

Index Terms—Learning from Touch, Learning from Humans, Imitation Learning

I. INTRODUCTION

Humans excel at manipulating the physical world not only through vision but also through a rich and nuanced sense of touch. Everyday actions—like delicately placing an egg in a bowl or unstacking a cup—depend on the ability to modulate fine-grained contact forces in real time. This form of low-level contact reasoning remains a core unsolved challenge in robotics [30]. Although recent advances in vision-based and proprioceptive imitation learning have enabled robots to perform increasingly complex tasks, they often fail when subtle force adjustments are needed, particularly under partial observability or contact uncertainty [24]. This gap arises from the mismatch between the high-bandwidth, tactile-rich control humans employ and the sparse, delayed signals available to most robotic systems.

Prior work in tactile imitation learning [22] has explored ways to integrate force feedback into robot control. However, such work relies heavily on teleoperation which is hard to scale to diverse real-world environments [15]. Furthermore, receiving tactile feedback through data-collection interfaces requires expensive haptic feedback setups [18] or the demonstrator adjusting continuous gripper closure based on visual cues such as object deformation [34]. In contrast, humans interact with the physical world constantly, providing a vast, underutilized source of contact-rich demonstrations in everyday settings. Works such as [28] propose mobile data collection systems involving glove-based motion capture. However, without force-sensing, such methods rely on task-specific priors, such as manually offsetting desired actions to approximate increased target forces.

Even if we can obtain tactile information, a number of

learning challenges stand in the way of leveraging this data for robust force-sensitive manipulation. Many methods propose to simply feed observed tactile information into the policy [31, 18, 24, 22]. However, this passive use of tactile information suffers from two main issues. Firstly, the learned model is highly reliant on the observed force distribution. Due to the embodiment gap between the human and the robot, the executable force distributions differ. If during deployment, the robot leaves the force distribution determined by the human training data, the policy will fail to generalize. Secondly, since more information is being fed into the policy in order to predict more precise actions, much more data is often required reducing learning efficiency.

In this work, we ask: Can we endow robots with robust, force-aware control by learning efficiently from human tactile experiences? We present FEELTHEFORCE (FTF): a novel framework that bridges human tactile demonstrations and robotic force-sensitive manipulation. FTF models human tactile-proprioceptive signals during manipulation and trains a closed-loop imitation learning policy to predict hand trajectories and desired contact forces. We collect data using a tactile glove and train a transformer-based policy that outputs hand trajectories, which are retargeted to robot end-effector poses, along with the contact force applied by the human at each timestep. At deployment, a low-level PD controller modulates the gripper closure to track the predicted force, enabling precise force reproduction across tasks without any robot data during training. This inference-time, PD-based controller continuously adjusts the robot's behavior to stabilize around desired tactile feedback—enabling robust execution even under morphological mismatch or sensor noise. Unlike prior approaches that require action-aligned teleoperation, our method decouples the learning and execution phases, leveraging human expertise to guide robot force control.

This formulation offers two key advantages: (1) it eliminates the need for large-scale robot data and expensive haptic teleoperation and (2) it enables generalization from the human embodiment to the robot embodiment to solve force-sensitive tasks robustly. We transfer the learned policy to a Franka Panda robot with fingertip tactile sensors and evaluate on 5 force-sensitive manipulation tasks.

In summary, we demonstrate that:

- FTF robustly solves all 5 force-sensitive tasks evaluated with a 77% success rate where baselines fail showing that active force prediction and reproduction is more effective than passive use of multi-modal force inputs.
- FTF achieves higher success rates than baselines trained on robot teleoperation data showing that the natural data collection enabled by the tactile glove can be effective for tactile data collection.
- FTF is able to achieve a success rate of 67% on a task with adversarial disturbances during deployment, displaying robustness to test-time shifts in the tactile data distribution.

II. RELATED WORK

Tactile Sensing and Haptic Learning. Tactile sensing plays a crucial role in enabling dexterous manipulation, particularly in tasks requiring fine-grained force control. Prior work has explored learning policies directly from tactile inputs, using sensors such as GelSight [32], BioTac [8], and other vision-based tactile arrays [16]. These methods have enabled tasks like slip detection [4], grasp stability prediction [27], and in-hand manipulation [21]. However, most tactile learning approaches rely on robot-collected data, which can be costly to obtain and may not generalize well across tasks or hardware. In contrast, our method leverages human tactile signals alone for training and sidesteps the need for robot interaction data.

Imitation Learning from Human Demonstrations. Learning from human demonstrations has a long history in robotics, encompassing techniques like behavior cloning [23], inverse reinforcement learning [20], and more recent deep imitation frameworks [33, 7]. Most approaches rely on visual or kinematic supervision, requiring aligned action spaces or kinesthetic teaching. Some methods attempt to use wearable devices like motion capture suits or teleoperation rigs [28, 34], but these are often expensive or intrusive. While a few studies explore imitation from human force data [13, 29], they typically assume action-space correspondence or require ground-truth contact forces. Our work is distinguished by using raw tactile and proprioceptive human signals, without requiring direct imitation of motor actions or precise temporal alignment between human and robot data.

Human to Robot Embodiment Transfer Transferring skills across embodiments—especially from human to robot—is a core challenge in imitation learning. Recent work [9] introduces fingertip kinematic retargeting to map human video demonstrations to coarse robot actions, which are subsequently refined through reinforcement learning. Other approaches [1, 2] focus on extracting manipulation affordances from human videos to define high-level interaction goals for robot control. [31, 6, 18, 25] leverage hand-held data collection tools with grippers designed to match the target robot's embodiment, improving correspondence between human demonstrations and robot execution. Our method builds upon [11] by using visual hand keypoints to bridge the morphological gap between humans and robots, enabling data collection in the demonstrator's natural embodiment while still recovering precise, executable actions for the robot.

III. FTF

Learning force-sensitive manipulation from humans is challenging due to the complexity of inferring fine-grained contacts and forces solely from visual observations. To address this, we introduce FTF, a framework for collecting tactile data from human demonstrations using a low-cost force-sensing glove and learning policies that predict both actions and desired forces from combined visual and tactile inputs. FTF enables the human-to-robot embodiment transfer through a key point based unified observation and action space [10], while
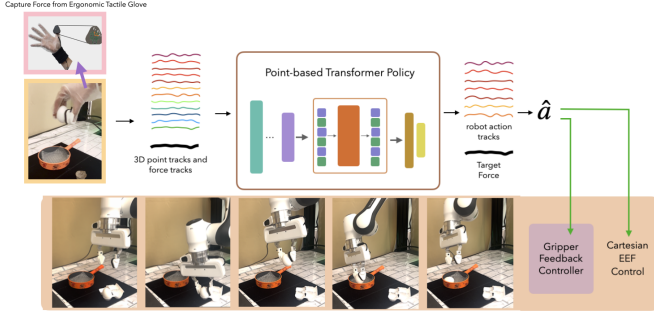
Fig. 2: FTF allows zero-shot transfer of tactile human demonstrations to a Franka Robot.

(a) AnySkin augmented glove, worn by a human data-collector. The straightforward electronics of the sensor interface both reduces excessive wiring and also allows for a bluetooth setting (right).

(b) Middle: Franka Panda gripper with AnySkin on one fingertip, emulating the human wearable (left). We attach a plain silicone cap on the other fingertip.

Fig. 3: Hardware setup for human demonstration and robot replication using AnySkin [22].

allowing human-like force modulation at inference through PD control on predicted forces.
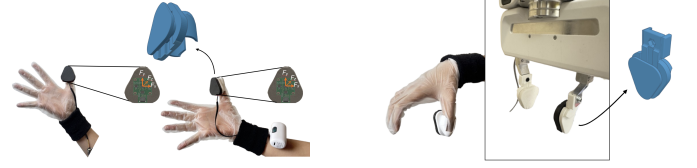
a) Assumptions: (1) The pose of the human hand in the first frame is the same as that of the robot gripper at reset. This can be relaxed by initializing the robot to arbitrary human hand poses, which we do not investigate in this work. (2) We operate in a calibrated scene where the intrinsic and extrinsic matrices for each camera is known. In practice, this is a one-time process that only takes a few minutes when the robot system is set up.

A. Data Acquisition for Human-to-Robot Force Transfer

To collect human demonstrations, FTF enables task execution through natural human movements. During data acquisition, as the human performs the task, two calibrated RealSense cameras record visual observations of the hand and environment. At deployment, the same camera setup monitors a Franka Panda arm in the same environment. A custom tactile glove is designed to collect force data from human demonstrations, which is transferred to the robot using gripper-mounted tactile sensors.

a) Human Hardware Design: During human data collection, force interactions are captured using a custom ergonomic tactile glove, inspired by AnySkin [3], which uses 3D-printed magnetometer-based sensors on the underside of the thumb (palm side) to avoid obstructing manipulation. The glove's transparent design preserves the hand's natural visual appearance, while an embedded PCB collects force data and streams it via USB to a nearby desktop computer. A schematic of the glove design is shown in Figure 3a. While each magnetometer captures full 3D force vectors, we aggregate the force reading by taking the norm of the center magnetometer's force vector. Since the force readings are streamed at 200 fps while the camera readings are 30 fps, we aggregate force readings across time to produce an effective fps of nearly 30. We provide the sensor norm to applied force (Newton) mapping in Figure 5.

b) Robot Hardware Design: During robot deployment, we 3D print custom gripper tips for the robot end-effector with a mount for fixing the AnySkin tactile sensors. The tactile sensor is only mounted on one of the gripper jaws to emulate the setup on the tactile glove. This ensures a one-to-one correspondence between the sensors used for human data

collection and deployment. The robot gripper force sensing design is shown in Figure 3b.

B. Embodiment Agnostic Scene Representation

The human hand motion data from tactile gloves is converted into a point-based representation to enable robot policy learning from human demonstrations.

1) Human-to-Robot Embodiment Transfer: For each time step t of a human video, we use Mediapipe [19] to extract image key points p_h^t on the human hand. Using point triangulation, the corresponding hand key points from two fixed, calibrated camera views are projected to 3D hand key points $\mathcal{P}_h^t$. We use point triangulation for 3D projection due to its higher accuracy as compared to sensor depth from the camera [10]. The robot position $\mathcal{R}_{pos}^t$ is computed as the midpoint between the tips of the index finger and thumb in $\mathcal{P}_h^t$. The robot orientation $\mathcal{R}_{ori}^t$ is computed as

$$
\begin{aligned}
\Delta\mathcal{R}_{ori}^t &= \mathcal{T}(\mathcal{P}_h^0, \mathcal{P}_h^t) \\
\mathcal{R}_{ori}^t &= \Delta\mathcal{R}_{ori}^t \cdot \mathcal{R}_{ori}^0
\end{aligned}
\tag{1}
$$

where $\mathcal{T}$ computes the rigid transform between hand key points on the first frame of the video, $\mathcal{P}_h^0$, and $\mathcal{P}_h^t$. The robot end effector pose is then represented at $T_r^t \leftarrow \{\mathcal{R}_{pos}^t, \mathcal{R}_{ori}^t\}$. Finally, the robot pose T_r^t is converted to N robot key points through a set of N rigid transformations T about the computed robot pose such that

$$
(\mathcal{P}_r^t)^i = T_r^t \cdot T^i, \quad \forall i \in \{1, ..., N\}
\tag{2}
$$

The robot's gripper state $\mathcal{R}_g$ is considered closed when the distance between the tip of the index finger and thumb is less than 7cm, otherwise open. The continuous force value measured for each step, $\mathcal{R}_f^t$, is also included in the robot state. This process has been illustrated in Figure 2.

2) Scene Key Point Representation: The environment is represented as key points through sparse human annotations, following prior work [17, 10]. Given a single demonstration frame, a human user annotates semantically meaningful key points on task-relevant objects in the scene. Using DIFT [26],

an off-the-shelf semantic correspondence model, the annotations are propagated to the first frames of all other demonstrations, minimizing human effort. For each demonstration, Co-Tracker [14], an off-the-shelf point tracker, then tracks the initialized key point through each trajectory, efficiently handling occlusions and maintaining temporal consistency. To obtain 3D object key points, we triangulate the tracked key points from the two camera views, grounding them in the robot's base frame. During inference, DIFT is used to localize keypoints in the first frame, after which Co-Tracker tracks them during execution. This approach leverages large pre-trained vision models to generalize across novel object instances and scenes without additional training, requiring only a single frame of user input per task.

C. Policy Learning

For policy learning, we use a transformer policy architecture [11, 10] that takes as input the robot points $\mathcal{P}_r$ and object points $\mathcal{P}_o$ along with the binarized gripper state $\mathcal{R}_g$ and continuous force value $\mathcal{R}_f$. Since the gripper state and force value are 1D and the points are 3D, we repeat the value 3 times when appending to the point tracks to ensure dimensional consistency. A history of observations for each key point is flattened into a single vector and encoded using a multilayer perceptron (MLP) encoder. Each encoded point track and the history of gripper and force values are fed as a separate token into the transformer policy, which predicts the future tracks for each robot point $\hat{\mathcal{P}}_r$, the robot gripper state $\hat{\mathcal{G}}_r$, and future gripper force predictions $\hat{\mathcal{F}}_r$. using a deterministic action head. Following prior works in policy learning [34, 5], we use action chunking with exponential temporal averaging to ensure temporal smoothness of the predicted point tracks. The policy is trained using a mean squared error loss. The transformer is non-causal in this scenario, and the training loss is only applied to the robot point tracks.

D. Inference

Algorithm 1: FORCEFEEDBACKGRIPPERCONTROL($\hat{F}_t$)

Algorithm 1

1: Initialize $\tau \leftarrow 0$
2: **repeat**
3:     $\Delta g_t^\tau = k \cdot (\hat{F}_t - F_t^\tau)$
4:     $g_t^{\tau+1} = g_t^\tau + \Delta g_t^\tau$
5:     Execute gripper action $g_t^{\tau+1}$
6:     Read $F_t^{\tau+1}$ from AnySkin
7:     $\tau \leftarrow \tau + 1$
8: **until** $||\hat{F}_t - F_t^\tau|| \le \epsilon$

---

a) Robot pose from predicted key points: The predicted robot points $\hat{\mathcal{P}}_r$ are mapped back to the robot pose using constraints from rigid-body geometry. We first consider the key point corresponding to the robot's wrist $\hat{\mathcal{P}}_r^{wrist}$ as the robot position $\hat{\mathcal{R}}_{pos}$. The robot orientation $\hat{\mathcal{R}}_{ori}$ is computed using Eq. 1 considering $\mathcal{R}_{ori}^0$ is fixed and known. Finally, the robot pose $\hat{\mathcal{R}}_{pose}$ is defined as $(\hat{\mathcal{R}}_{pos}, \hat{\mathcal{R}}_{ori})$.

Algorithm 2 FTF Policy Inference

1: Obtain object keypoints on first frame using DiFT on annotated dataset frame.
2: **for** t in rollout **do**
3:     Compute action chunk $(\hat{\bar{a}}_t, ..., \hat{\bar{a}}_{t+H}) = \pi(a|s_t)$ and obtain $\hat{a}_t$ with temporal aggregation.
4:     Parse action: $(\hat{F}_t, \hat{g}_t, \hat{a}_t^{eef}) \leftarrow \hat{a}_t$
5:     **if** $\hat{g}_t > closethreshold$ **then**
6:         Call FORCEFEEDBACKGRIPPERCONTROL($\hat{F}_t$)
7:     **else if** $\hat{g}_t < openthreshold$ **then**
8:         Open gripper
9:     **end if**
10:    Execute $\hat{a}_t^{eef}$ on robot
11:    Read next state s_{t+1} using Co-Tracker
12: **end for**

---

b) Inference-time PD force controller: To deploy the tactile policy on the robot arm, we need a means for the robot gripper exerting the force predicted by the policy at each step. For this, we design an outer-loop PD controller that adjusts the target gripper closure setpoints to stabilize the measured forces. If at some timestep t, the policy predicts a force $\hat{F}_t$ to be applied, the controller is:

$$\Delta g_t^\tau = k \cdot (\hat{F}_t - F_t^\tau) \qquad (3)$$

where τ is the inner loop timestep of the PD controller and F_t^τ is the force read by the robot at timestep t of the policy and timestep τ of the controller. At each step the gripper closure is updated as $g_t^{\tau+1} = g_t^\tau + \Delta g_t^\tau$.

The PD controller runs until the convergence condition $||\hat{F}_t - F_t^\tau|| < \epsilon$. After the controller converges to the desired $\hat{F}_t$, the policy predicts the next action for step $t+1$. We find $k = 0.001$ and $\epsilon = 5$ to work well across all tasks. Finally, the action $\hat{\mathcal{A}}_r = (\hat{\mathcal{R}}_{pose}, \hat{\mathcal{G}}_r, g_t)$ is executed on the robot using end-effector position control at a 6Hz frequency.

IV. EXPERIMENTS

Our experiments are designed to answer the following questions: (1) How well does FTF work for learning force-sensitive tasks compared to baselines that learn to use human force data passiveley? (2) How does FTF compare to tactile policy learning methods that learn from robot teleoperation data? (3) Is FTF robust to test-time disturbances?

Experimental Setup We evaluate FTF on a Franka Panda robot, operating in a real-world tabletop manipulation environment. Two Intel RealSense D435 cameras are mounted to provide third-person RGB images to our policy. For baselines we also collect 30 demonstrations on the Franka robot per task using a VR-based teleoperation framework [12]. Demonstrations are recorded at 20Hz and subsampled to approximately 6Hz. For methods outputting robot actions, we use absolute actions with orientation represented with a 6D rotation representation [35].

| Place soft bread on plate | Unstack single plastic cup from stack | Place egg in pot | Place bag of chips on plate | Twist and lift bottle cap |

Fig. 4: Manipulation tasks evaluated with FTF

Task Descriptions Our manipulation tasks involve variations designed to evaluate the scope of force-sensitive manipulation capabilities achievable with FTF. The tasks are depicted in Figure 4. We provide a description of each task below. For each task, we collect 30 human demonstrations and 30 teleoperated demonstrations.

a) `Place soft bread on plate`: The robot arm picks up a highly deformable piece of bread from the table and places it on the plate without crushing it. The positions of the bread are varied for each evaluation.

b) `Unstack single plastic cup from stack`: The robot arm unstacks a single plastic cup from a stack of 3 upside down cups on the table.

c) `Place egg in pot`: The robot arm gently picks up and places an egg in a pot without crushing it. The position of the egg is varied for each evaluation.

d) `Place bag of chips on plate`: The robot arm picks and places a transparent bag of chips into a plate without crushing any of the chips inside.

e) `Twist and lift bottle cap`: The robot arm twists and lifts the cap of a bottle to remove it.

Baselines We compare FTF with 5 baselines - *Tactile Point Policy* [10], *Continuous-Gripper Tactile Point Policy*, *FTF + Tactile P3-PO* [17], *Tactile P3-PO*, and *Continuous-Gripper Tactile P3-PO*. We describe each method below.

a) *Tactile Point Policy* [10] performs behavior cloning from point tracks extracted from human data as well as force readings from the tactile glove and predicts future tracks which are converted into robot actions. This baseline provides a comparison to methods such as [22] that use force input to improve the precision of learned policies but in the context of human data.

(b) *Continuous-Gripper Tactile Point Policy* is similar to *Tactile Point Policy* but predicts continuous gripper closure. The gripper closure value is measured as the distance between the index and thumb tracked points from the human data renormalized to the range of the robot gripper.

(c) *FTF + Tactile P3-PO* extends *Tactile P3-PO* by predicting both robot actions and future contact forces. The model is trained on teleoperated robot data, using force signals collected during teleoperation as input, and outputs predicted forces alongside actions. This baseline evaluates whether incorporating force prediction improves control performance in robot teleoperation setting and compares the utility of robot-collected versus human-collected force data.

(d) *Tactile P3-PO* [17] predicts teleoperated robot actions from robot tracks obtained by unprojecting robot and object points of interst into 3D space. The method also inputs force readings from the robot gripper collected during teleoperation into the Transformer policy. This method provides a similar comparison to *Tactile Point Policy* but on teleoperated robot data and ground truth robot actions.

(e) *Continuous-Gripper Tactile P3-PO* is similar to *Tactile P3-PO* but predicts continuous gripper closure. The continuous gripper values are obtained directly from the robot gripper during teleoperator using an adaptation to the VR-based teleoperation framework [12] that allows the teleoperator to output continuous gripper closures based on visual feedback during data collection.

FTF is highly effective at learning low-level force control strategies for force-sensitive tasks. Table I compares the success rates of FTF with baseline methods. FTF stands out as the only approach capable of reliably solving tasks that require precise force application within a narrow range. For example, in the `unstack single plastic cup from stack` task, FTF is the only method able to lift the cup without inadvertently lifting others underneath. In contrast, the binary gripper baseline can lift the cup, but it fails to isolate the cup and ends up lifting all three cups, causing task failure. Similarly, for delicate tasks like `place soft bread on plate`, `place egg in pot`, and `place bag of chips on plate`, the binary gripper fails by crushing the object. In the `twist and lift bottle cap` task, although the binary gripper achieves 11/15 successes by firmly grasping the rigid cap, it often applies excessive force, leading to failures such as lifting the bottle or damaging the gripper pad. In contrast, FTF achieves 13/15 successes with more controlled and precise force application.

The robustness of *Tactile Point Policy's* underlying policy allows the binary gripper baseline to occasionally complete the task, albeit without adhering to the force constraints. Continuous grippers, on the other hand, struggle significantly across tasks. This is because relying solely on finger separation and mapping it to the robot is imprecise, leading to excessive oscillations that undermine stable grasps. Fine-grained positional control of the gripper would likely require more data than we collected in this study, due to the complexity of this mapping.

FTF generally outperforms teleoperation baselines that use force data passively. In Table II, we also provide results comparing FTF with baseline gripper strategies trained on robot teleoperation demonstrations. FTF generally performs better than the binary and continuous baselines across most tasks, showing that predicting contact force improves performance in the teleoperation setting. However, force signals collected during teleoperation are less reliable than those from human demonstrations. For instance, in `place egg in pot` and `twist and lift bottle cap`, the FTF fails to outperform the binary gripper, indicating inconsistencies or noise in the teleoperated force stream.

We also observe that the continuous gripper baseline struggles due to the sample inefficiency of learning precise gripper closure. While it occasionally performs better—for example,

TABLE I: Performance comparison of different gripper action spaces in Human Demo

Task	FTF	Binary Gripper	Continuous Gripper
Place soft bread on plate	13/15	0/15	0/15
Unstack single plastic cup from stack	9/15	0/15	0/15 (2/15 picked 3 cups)
Place egg in pot	13/15	0/15	0/15
Place bag of chips on plate	10/15	0/15	0/15
Twist and lift bottle cap	13/15	11/15 (1/15 break gripper pads)	0/15

TABLE II: Performance comparison of different gripper action spaces in Robot Teleop Demo

Task	FTF	Binary Gripper	Continuous Gripper
Place soft bread on plate	5/15	0/15	3/15
Unstack single plastic cup from stack	4/15	0/15 (6/15 picked 3 cups)	0/15 (2/15 picked 2 cups)
Place egg in pot	0/15	0/15	0/15
Place bag of chips on plate	3/15	0/15	0/15
Twist and lift bottle cap	9/15	12/15	8/15

picking up two cups instead of all three in the stacking task—it tends to work only for rigid objects like the bottle cap, where the required closure remains consistent. In contrast, it fails on deformable objects like chip bags, where the required gripper behavior varies more between demonstrations and test-time executions.

TABLE III: Analysis of FTF with test time adversarial disturbance

Task	FTF
Place bag of chips on plate	10/15

FTF is robust to test-time disturbances. In the *placing a bag of chips on a plate* task, we introduce disturbances by physically interacting with the bag by holding it down to the table, pressing on the top during the lift, or slightly reorienting it. As shown in Table III, despite the changes in the observed force profiles, FTF is able to adapt and still produce the desired forces with a 67% success rate.

V. CONCLUSION AND LIMITATIONS

We present FTF, a novel framework for learning force-sensitive manipulation from human tactile demonstrations. By leveraging a tactile glove and vision-based hand pose estimation, FTF captures rich contact force signals from natural human interactions without relying on teleoperation or robot-collected data. Our system trains a closed-loop policy to predict hand trajectories and desired contact forces, which are then retargeted to a robot using a PD controller that enables precise and robust force control. Through experiments across diverse manipulation tasks, we demonstrate that FTF significantly outperforms prior baselines and and remains robust under perturbations. These results highlight the power of modeling human tactile behavior—paving the way for more effective robot learning from human experience. Existing limitations of FTF include: 1) Shear forces are aggregated with normal forces leading to loss of directional information. For more dexterous tasks, force components can be collected, learned, and stabilized independently. 2) The data collection infrastructure is limited to a fixed, calibrated camera setting.

Leveraging egocentric cameras and using stereo triangulation for 3D point extraction could allow for in-the-wild data collection.

REFERENCES

[1] Shikhar Bahl, Abhinav Gupta, and Deepak Pathak. Human-to-robot imitation in the wild. *arXiv preprint arXiv:2207.09450*, 2022.

[2] Shikhar Bahl, Abhinav Gupta, and Deepak Pathak. Human-to-robot imitation in the wild, 2022. URL https://arxiv.org/abs/2207.09450.

[3] Raunaq Bhirangi, Venkatesh Pattabiraman, Enes Erciyes, Yifeng Cao, Tess Hellebrekers, and Lerrel Pinto. Anyskin: Plug-and-play skin sensing for robotic touch, 2024. URL https://arxiv.org/abs/2409.08276.

[4] Roberto Calandra, Andrew Owens, Manu Upadhyaya, Wenzhen Yuan, Justin Lin, Edward H. Adelson, and Sergey Levine. The feeling of success: Does touch sensing help predict grasp outcomes?, 2025. URL https://arxiv.org/abs/1710.05512.

[5] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *arXiv preprint arXiv:2303.04137*, 2023.

[6] Cheng Chi, Zhenjia Xu, Chuer Pan, Eric Cousineau, Benjamin Burchfiel, Siyuan Feng, Russ Tedrake, and Shuran Song. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. *arXiv preprint arXiv:2402.10329*, 2024.

[7] Sudeep Dasari and Abhinav Gupta. Transformers for one-shot visual imitation. In *Conference on Robot Learning*, pages 2071–2084. PMLR, 2021.

[8] Jeremy A. Fishel and Gerald E. Loeb. Sensing tactile microvibrations with the biotac — comparison with human sensitivity. In *2012 4th IEEE RAS EMBS International Conference on Biomedical Robotics and Biomechatronics (BioRob)*, pages 1122–1127, 2012. doi: 10.1109/BioRob.2012.6290741.

[9] Irmak Guzey, Yinlong Dai, Georgy Savva, Raunaq Bhirangi, and Lerrel Pinto. Bridging the human to robot dexterity gap through object-oriented rewards, 2024. URL https://arxiv.org/abs/2410.23289.

[10] Siddhant Haldar and Lerrel Pinto. Point policy: Unifying observations and actions with key points for robot manipulation, 2025. URL https://arxiv.org/abs/2502.20391.

[11] Siddhant Haldar, Zhuoran Peng, and Lerrel Pinto. Baku: An efficient transformer for multi-task policy learning, 2024. URL https://arxiv.org/abs/2406.07539.

[12] Aadhithya Iyer, Zhuoran Peng, Yinlong Dai, Irmak Guzey, Siddhant Haldar, Soumith Chintala, and Lerrel Pinto. Open teach: A versatile teleoperation system for robotic manipulation, 2024. URL https://arxiv.org/abs/2403.07870.

[13] Advait Jain and Charles C Kemp. Improving robot manipulation with data-driven object-centric models of everyday forces. *Autonomous Robots*, 35:143–159, 2013.

[14] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together, 2024. URL https://arxiv.org/abs/2307.07635.

[15] Alexander Khazatsky. Droid: A large-scale in-the-wild robot manipulation dataset, 2024. URL https://arxiv.org/abs/2403.12945.

[16] Mike Lambeta, Po-Wei Chou, Stephen Tian, Brian Yang, Benjamin Maloon, Victoria Rose Most, Dave Stroud, Raymond Santos, Ahmad Byagowi, Gregg Kammerer, Dinesh Jayaraman, and Roberto Calandra. Digit: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation. *IEEE Robotics and Automation Letters*, 5(3):3838–3845, 2020. doi: 10.1109/LRA.2020.2977257.

[17] Mara Levy, Siddhant Haldar, Lerrel Pinto, and Abhinav Shirivastava. P3-po: Prescriptive point priors for visuo-spatial generalization of robot policies, 2024. URL https://arxiv.org/abs/2412.06784.

[18] Fangchen Liu, Chuanyu Li, Yihua Qin, Ankit Shaw, Jing Xu, Pieter Abbeel, and Rui Chen. Vitamin: Learning contact-rich tasks through robot-free visuo-tactile manipulation interface, 2025. URL https://arxiv.org/abs/2504.06156.

[19] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019.

[20] Andrew Y. Ng and Stuart J. Russell. Algorithms for inverse reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, page 663–670, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc. ISBN 1558607072.

[21] Chaoyi Pan, Marion Lepert, Shenli Yuan, Rika Antonova, and Jeannette Bohg. In-hand manipulation of unknown objects with tactile sensing for insertion, 2023. URL https://arxiv.org/abs/2210.13403.

[22] Venkatesh Pattabiraman, Yifeng Cao, Siddhant Haldar, Lerrel Pinto, and Raunaq Bhirangi. Learning precise, contact-rich manipulation through uncalibrated tactile skins, 2024. URL https://arxiv.org/abs/2410.17246.

[23] Dean A. Pomerleau. *ALVINN: an autonomous land vehicle in a neural network*, page 305–313. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1989. ISBN 1558600159.

[24] Carmelo Sferrazza, Younggyo Seo, Hao Liu, Youngwoon Lee, and Pieter Abbeel. The power of the senses: Generalizable manipulation from vision and touch through masked multimodal learning, 2023. URL https://arxiv.org/abs/2311.00924.

[25] Nur Muhammad Mahi Shafiullah, Anant Rai, Haritheja Etukuru, Yiqian Liu, Ishan Misra, Soumith Chintala, and Lerrel Pinto. On bringing robots home. *arXiv preprint arXiv:2311.16098*, 2023.

[26] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion, 2023. URL https://arxiv.org/abs/

2306.03881.

[27] Filipe Veiga, Herke van Hoof, Jan Peters, and Tucker Hermans. Stabilizing novel objects by learning to predict tactile slip. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5065–5072, 2015. doi: 10.1109/IROS.2015.7354090.

[28] Chen Wang, Haochen Shi, Weizhuo Wang, Ruohan Zhang, Li Fei-Fei, and C. Karen Liu. Dexcap: Scalable and portable mocap data collection system for dexterous manipulation, 2024. URL https://arxiv.org/abs/2403.07788.

[29] Tuomas E Wiste, Skyler A Dalley, H Atakan Varol, and Michael Goldfarb. Design of a multigrasp transradial prosthesis. 2011.

[30] William Xie and Nikolaus Correll. Towards forceful robotic foundation models: a literature survey, 2025. URL https://arxiv.org/abs/2504.11827.

[31] Kelin Yu, Yunhai Han, Qixian Wang, Vaibhav Saxena, Danfei Xu, and Ye Zhao. Mimictouch: Leveraging multimodal human tactile demonstrations for contact-rich manipulation, 2025. URL https://arxiv.org/abs/2310.16917.

[32] Wenzhen Yuan, Siyuan Dong, and Edward H. Adelson. Gelsight: High-resolution robot tactile sensors for estimating geometry and force. *Sensors*, 17(12), 2017. ISSN 1424-8220. URL https://www.mdpi.com/1424-8220/17/12/2762.

[33] Tianhao Zhang, Zoe McCarthy, Owen Jow, Dennis Lee, Ken Goldberg, and Pieter Abbeel. Deep imitation learning for complex manipulation tasks from virtual reality teleoperation. *CoRR*, abs/1710.04615, 2017. URL http://arxiv.org/abs/1710.04615.

[34] Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023.

[35] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. *CoRR*, abs/1812.07035, 2018. URL http://arxiv.org/abs/1812.07035.
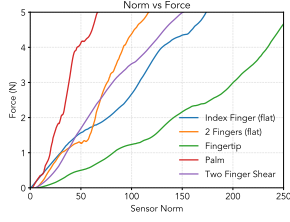
(a) Transformation between the sensor norm and applied force, for various modes of contact, ranging from low-pressure (Palm) to high pressure (Index Fingertip)

(b) Different modes of contact for data collection; Two fingers laid flat (top left), index fingertip (bottom left), palm (middle), one finger laid flat (top right) and two fingers pressing at an angle to emulate a combination of both normal and shear forces

Fig. 5: (Left) Mapping between sensor norm and applied force across various contact modes. (Right) Data collection setup illustrating different types of contact used in the mapping process.

Sensor Norm to Force (Newton) conversion In order to demonstrate a transform between the sensor norm and applied force, we collect data by pressing on an AnySkin sensor mounted on a weighing scale and record synchronous data from both the sensor and the scale streamed through USB at 10Hz. We press the sensor in 5 different manners gradually increasing the force from 0 to 5N, in order to capture different pressures and diverse modes of contact. The sensor norm to applied force comparison is illustrated in Figure 5.
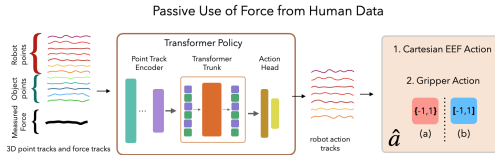
(a) Point track baselines from human data with passive use of force. (a) uses a binary gripper action space by thresholding human hand closure and (b) retargets continuous human hand closure to continuous gripper.
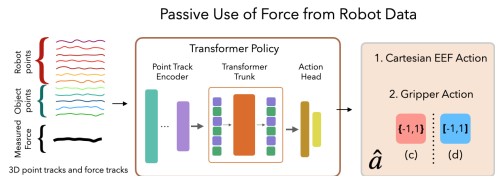
(b) Action imitation baselines from robot data with passive use of force. (c) uses a binary gripper action space and (d) uses continuous gripper action space.

Fig. 6: Comparison of gripper action space across human (left) and robot (right) baselines under passive force conditions.

TABLE IV: Performance comparison of masked vs umasked force tracks inputted to FTF

Task	Masked Force	Unmasked Force
Lift and hold bread	10/10	10/10

FTF does not require force input to perform effectively. We implemented a variant of FTF that masks the force data fed to the Transformer, constraining the model to predict desired forces solely based on the environment and robot state. We evaluate FTF on a simple task `Lift and hold bread`, involving picking up a soft piece of bread without crushing it and suspending the grasp in the air in Table IV. This modification did not result in any degradation of force prediction or task performance, suggesting that FTF can achieve effective force control even without explicit force input.
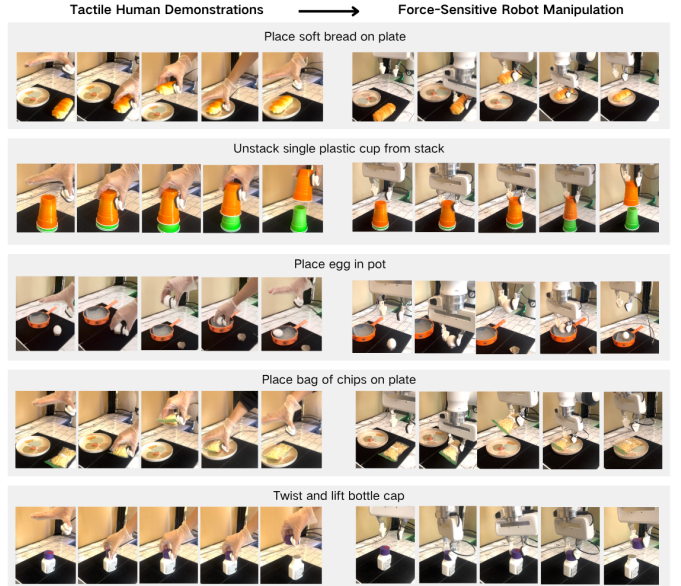

Fig. 7: Visual comparison of tactile human demonstrations (left) and force-sensitive robot manipulation rollouts (right) learned from the human demonstrations.