PP-Tac: Paper Picking Using Tactile Feedback in Dexterous Robotic Hands

Pei Lin[1,2⋆] Yuzhe Huang[1,3⋆] Wanlin Li[1⋆] Jianpeng Ma[1] Chenxi Xiao[2†] Ziyuan Jiao[1†]
[1]Beijing Institute for General Artificial Intelligence [2]ShanghaiTech University [3]Beihang University
⋆equal contributors †corresponding authors
https://peilin-666.github.io/projects/PP-Tac

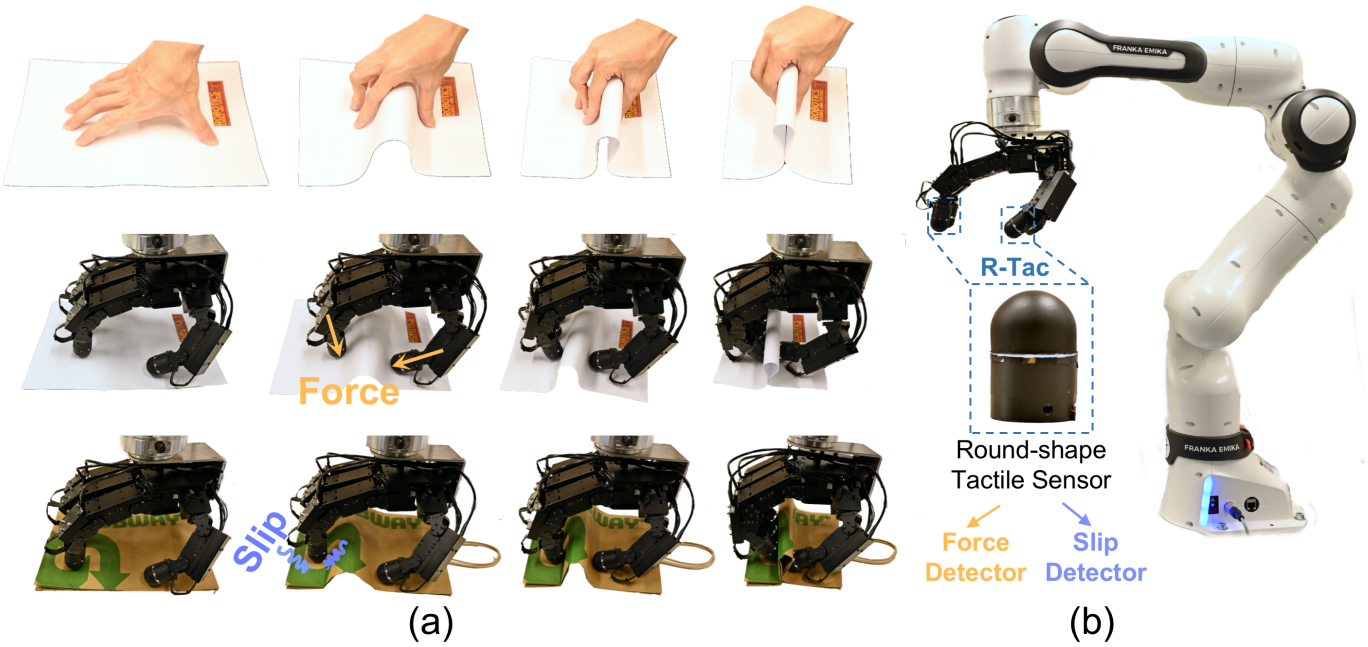

Fig. 1: **Overview of PP-Tac.** The system leverages tactile feedback from the proposed round-shaped sensor (R-Tac), integrated into a dexterous robotic hand, to grasp thin, deformable, paper-like objects. (a) Hand motions are generated by a diffusion-based policy inspired by human strategies, such as sliding and pinching. (b) The hardware setup includes a robotic arm, a dexterous hand, and four fingertip-mounted tactile sensors that simultaneously detect force and slip events.

***Abstract*—Robots are increasingly envisioned as human companions, assisting with everyday tasks that often involve manipulating deformable objects. Although recent advances in robotic hardware and embodied AI have expanded their capabilities, current systems still struggle with handling thin, flat, and deformable objects such as paper and fabric. This limitation arises from the lack of suitable perception techniques for robust state estimation under diverse object appearances, and the absence of planning techniques for generating appropriate grasp motions. To bridge these gaps, this paper introduces PP-Tac, a robotic system for picking up paper-like objects. PP-Tac features a multi-fingered robotic hand with high-resolution round-shaped tactile sensors R-Tac for omnidirectional tactile sensing. This hardware configuration enables real-time slip detection and online force control that mitigates slips. Furthermore, grasp motion generation is achieved through a trajectory synthesis pipeline, which first constructs a dataset of fingers' pinching motions. Then, a diffusion-based policy is trained to control the hand-arm robotic system. Experiments demonstrate that PP-Tac can effectively grasp paper-like objects of varying material, thickness, and stiffness, achieving an overall success rate of 87.5%. To the best of our knowledge, this work is the first attempt to grasp paper-like objects using a tactile dexterous hand.**

I. Introduction

Robots are increasingly popular as assistive agents in everyday life, particularly within household environments [36]. These robots are designed to perform various domestic tasks, often involving the grasp of thin, deformable objects such as paper and fabric [48]. For instance, clothes-folding tasks [25] require high dexterity and adaptability to accommodate variations in fabric size, texture, and stiffness, while document organization tasks [1] demand picking capabilities for diverse paper types and form factors. Beyond domestic settings, handling deformable objects is essential in industrial and logistical applications, such as fabricating fabrics [5] and packing objects using plastic bags and cardboard [12].

Despite their significance, picking up paper-like objects remains challenging in robotics [48]. In particular, the main challenges are three-fold: 1) Vision systems, commonly used for manipulation, struggle to perceive contact information during interactions with deformable objects due to limited sensing modalities and occlusion, resulting in an inaccurate

environment model for motion planning [26]; 2) These objects are often flat in shape, lacking salient features for contact points and thus hindering the synthesis of stable grasps [9]. 3) The appearance of such objects exhibits high variability, as their shape undergoes continuous and unpredictable deformation during manipulation. These dynamic shape variations significantly impair the generalizability of vision-based methods.

In contrast, humans excel at picking up paper-like objects by leveraging coordinated multi-fingered motion and tactile sensing. As shown in Fig. 1 (a), the process typically begins with establishing contact with the fingers, followed by sliding motions to deform the material and to generate a contact point for the pinched grasp. Such motion is made possible by hand's high Degree of Freedoms (DoFs), which enables establishing multiple contact points adaptively during sliding motion. During this process, tactile sensing is also crucial as it allows humans to perceive the object's deformation and decide the appropriate forces. These real-time adjustments ensure the successful execution of the picking-up action.

Inspired by human strategies, this paper introduces a robotics system coined *PP-Tac*: Paper-like object Picking using Tactile feedback. The PP-Tac system comprises two key hardwares: **A dexterous robotic hand, and the associated hemispherical and high-resolution Vision-Based Tactile Sensors (VBTS) R-Tac.** The fingertip-mounted tactile sensors provide real-time contact feedback during grasping operations. Featuring a spherical sensing area and a high-frame-rate monochrome camera, this design enables faster response times and simpler calibration processes compared to conventional RGB-based tactile sensors. An illustration of the system is shown in Fig. 1(b). In addition to the tactile sensor, this paper also presents **A diffusion-based motion generation policy (PP-Tac policy)** that imitates human picking-up skills. The proposed method first employs efficient trajectory optimization to generate expert data replicating human sliding and pinching motions. Second, generalizability to diverse flat objects were acheived by training a diffusion policy using these trajectories, leveraging proprioceptive data and tactile feedback for adaptive control of the dexterous robotic hand.

Comprehensive real-world experiments were conducted to evaluate the PP-Tac system. PP-Tac achieved an overall success rate of 87.5% in grasping everyday thin and deformable paper-like objects, including plastic bags, paper bags, and silk towels on flat surfaces. Fig. 1(a) illustrates examples of our arm-hand system successfully picking up paper-like objects. The PP-Tac also demonstrates significant adaptability in picking up paper-like objects on various uneven surfaces. Additionally, an ablation study further validated the contributions of each system component, highlighting the critical role of VBTS feedback and motion generation policies in achieving robust performance.

To the best of our knowledge, this work represents the first demonstration of deformable object picking using a dexterous hand equipped with VBTS. Overall, our contributions include:

1) We present R-Tac, a novel spherical tactile sensor designed with ease of fabrication, calibration, and scalable deployment. To demonstrate its utility, we integrate R-Tac into each fingertip of a fully actuated dexterous robotic hand, enabling real-time contact feedback during manipulation tasks.
2) We propose a novel trajectory-optimization-based data generation framework. The proposed framework does not rely on tactile or physical simulation, which is computationally expensive, and is capable of achieving robust sim-to-real transfer.
3) We present the *PP-Tac policy*, a diffusion-based control strategy that utilizes only tactile and robot proprioceptive feedback for manipulating paper-like objects. This approach demonstrates robust generalization across diverse materials and surface properties.
4) We provide the implementation and systematic experiments of the proposed algorithms on a real robot system. Both hardware and code for *PP-Tac system* are released to support further research and community development.

II. Related Work

A. Deformable Objects Manipulation

Deformable Object Manipulation (DOM) aims to handle soft objects that alter shape during interaction, which has been a long-standing challenging task in robotic research. Challenges mainly arised from uncertainties in perception and complex soft-body dynamics [15, 3, 29]. Early approaches solved such problem using visual perception for state estimation [48, 35], enabling tasks like rope-handling [31, 35], cloth-folding [40, 25] and picking up paper with marker [**?**]. However, vision-based methods often fall short in solving real-world DOM problems due to varying object appearance, object's physical property that are usually unknown in advance, visual occlusions [23, 6], and variable lighting conditions [46, 20]. These challenges hinder the development of scalable vision-based DOM solutions for diverse environments.

Tactile sensing, particularly Vision-Based Tactile Sensors (VBTS), has demonstrated significant potential for solving DOM tasks [48]. Leveraging their high-resolution tactile feedback, VBTS has demonstrated high performance in object shape reconstruction [32, 10, 28, 45], localization [18, 24, 7], and slip detection [42, 13]. Prior work has explored VBTS for deformable object manipulation [38], but existing implementations rely on gripper-mounted sensors, which lack the dexterity of multi-fingered hands due to limited DoF. Our experiments reveal that gripper-based approaches struggle with thin deformable objects, and lack sufficient adapability for objects placed on non-flat surfaces, which highlights the need for integrating dexterous robotic hands with VBTS for robust manipulation [19].

B. Dexterous Robotic Hand with Tactile Sensing

Current dexterous hands are often equipped with tactile sensors. Commonly used tactile sensors typically incorporate mechanisms such as capacitive [30], piezoresistive [17], or

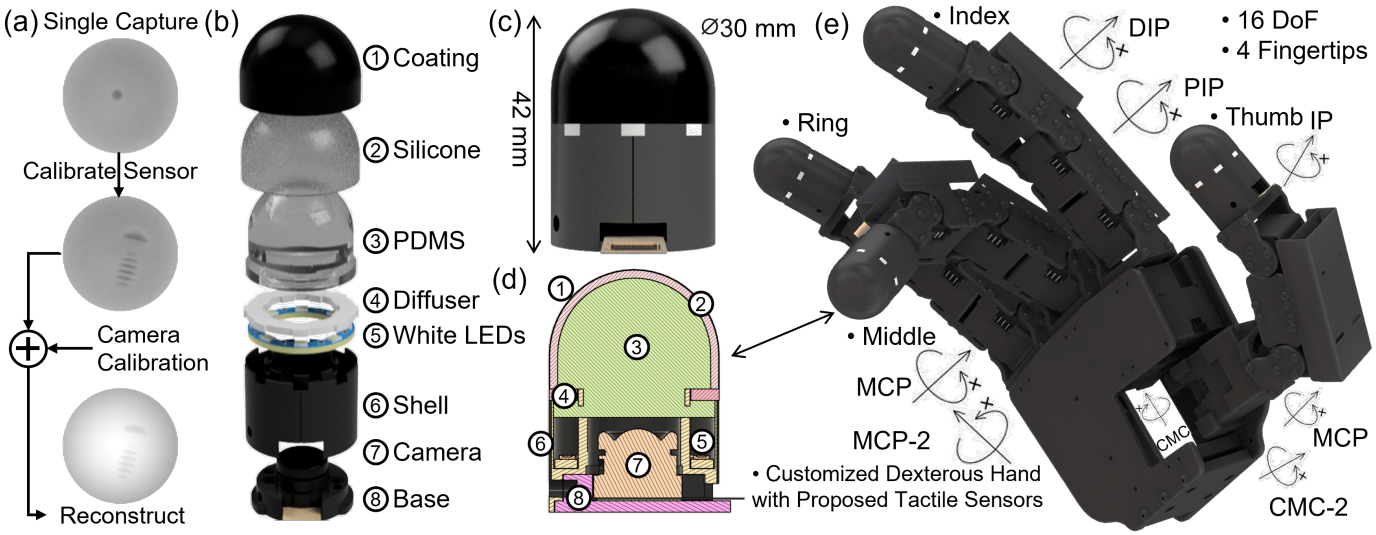

Fig. 2: **The hardware design of the R-Tac and its integration into the four-fingered dexterous robotic hand system.** (a) illustrates the pipeline of depth reconstruction. (b) illustrates the exploded view of the sensor, detailing each component. (c) shows the dimensions of the sensor. (d) shows the schematic design. (e) illustrates the robotic hand equipped with four sensors as its distal links.

magnetic-based [14] technologies. These designs can be fabricated in various shapes and sizes, allowing them to conform to the form factor of different robotic fingers. However, the sensing principles behind these technologies limit their spatial resolution and robustness under varying environmental conditions. To enhance sensing quality, recent work has devoted efforts to develop various VBTS sensors, especially with curved elastomer surfaces [10, 44, 21, 2, 4]. However, most of these VBTS sensors are not yet commercially available, and remain challenging to deploy at scale in hand. This is mainly due to the challenge arised from sensor's calibration, as the illumination from RGB chromatic light sources results in uneven light intensity distributions on curved elastomer surface, necessitating extensive data collection for calibration. During such data collection, these sensors [10, 44, 21, 2] often require specialized test beds (*e.g.*, fabricated using CNC machines) to collect large datasets, increasing the calibration complexity. Moreover, transmitting realtime chromatic video streams imposes higher bandwidth requirements, which can limit the overall frame rate in large-scale deployments. To address the above issues, we propose R-Tac that is structurally simple to fabricate, compact, and easy to calibrate.

Current robotic hands equipped with VBTS have been used in grasping and in-hand orientation tasks. For instance, Do *et al*. uses DenseTact [10] attached to an Allegro Hand, to grasp and manipulate small screws [11]. Qi *et al*. integrates fingertip VBTS [33] and DIGIT [41] on an Allegro Hand to rotate objects in hand. To the best of our knowledge, existing research has not yet explored VBTS-equipped dexterous hands for manipulating thin, deformable objects such as paper sheets.

III. Hardware Design

To provide sufficient dexterity to address the challenges of paper-picking tasks, we designed and fabricated a set of round-shape VBTS—R-Tac, which are integrated into Allegro Hand [34] through customization.

A. Fingertip-shaped Tactile Sensing

The design of R-Tac is guided by five key principles to ensure effective manipulation:

- **Round shape:** The hemispherical design enables omnidirectional tactile perception.
- **High resolution:** High resolution enables accurate depth reconstruction and slip detection during picking-up.
- **Convenient to fabricate & low-cost:** The components of the tactile sensor are either off-the-shelf or easy to fabricate, with a cost of around $60.
- **Efficient calibration:** The monochrome sensing principle simplifies lighting control and reduces manual effort for calibration, making it particularly suitable for large-scale deployment on multi-fingered robotic hands.
- **Efficient data transmission:** The monochrome camera produces lightweight data per frame, facilitating high-speed data transmission between systems.

Based on these 5 principles, the sensor design and its integration into the dexterous hand is illustrated in Fig. 2. Next, we detail each component and the calibration process.

1) Contact and Illumination Module: The core of the sensor is a contact module (elastomer) with a uniformly illuminated, deformable sensitive surface that maintains structural rigidity during contact. Inspired by the monochrome sensing principle [27], where intensity changes indicate deformation, we developed a hemispherical structure comprising a white LED ring, a stiff transparent internal skeleton, a soft semi-transparent perception layer, and a thin opaque protective layer that achieves the desired optical characteristics.

The LED ring (LUXEON 2835 4000K SMD LED) and a diffuser (double-sided frosted diffuser sheet) are first installed within the sensor shell. The skeleton is then manufactured

from PDMS (Dow Corning Sylgard 184 with Shore hardness 50 A) using a two-piece molding technique. The mixture (base: catalyst = 10: 1) is degassed and poured into the mold, and cured for 24 hours at room temperature. The perception layer is then manufactured similarly, using semitransparent silicone (Smooth-On Ecoflex with Shore hardness 00-10), and the layer is peeled off after 4 hours. Note that the measured depth range relies on the thickness of this layer, which is set to 2 mm. Finally, a silicone coating (Smooth-On Psycho Paint) is airbrushed onto the perception layer to form the opaque protective layer. The entire manufacturing process takes within 3 days, facilitating large-scale deployment.

2) Camera Module: A micro black-and-white CMOS camera (OV9281) with a wide 160° lens is used to capture the light intensity data. The camera operates up to 120Hz with a resolution of 640×480 and a latency of approximately 100ms.

3) Calibration: The uniform optical properties of the elastomer and illumination module (with a capture standard deviation as low as 6) enable the 3D geometry of the round shape sensor to be computed from single-channel pixel intensity in simply two steps using only 30 captures, without the need for a CNC machine. First, given the known intrinsic parameters K, camera calibration is performed using 29 captures in a 3D-printed indentation-based setup to estimate the extrinsic parameters of rotation matrix A and translation vector b, as well as the sensor surface reference projection D. Next, the depth mapping function M is calibrated by capturing a single image of a ball of known size pressed onto the sensor [27]. The complete mapping function from the pixel coordinates (u, v) to the sensor coordinates (x, y, z) can be expressed as:

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = A^{-1} \left((D(u,v) - M(I_\Delta(u,v)))K^{-1} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} - b \right), \quad (1)$$

which transforms grayscale intensity images to a depth map expressed in the sensor coordinates. A detailed explanation of camera calibration is provided in Appendix A. Reconstruction results and qualitative analysis are presented in Section VI-B.

4) Contact Force Estimation & Slip Detection: Our sensors are capable of detecting both contact forces and slip events. The contact force, modeled by elasticity theory, is proportional to the deformation depth as a linear function. The slip detection module is as follows:

- **Detection Model:** As illustrated in Fig. 3, when the slip occurs, distinct wrinkles become visible in the sensor's imaging. We apply a lightweight neural network architecture consisting of CNN (convolutional neural network) and MLP (multilayer perceptron) to detect the slip. The network processes a temporal sequence of the preceding five frames with a non-contact frame as input. CNN extracts the feature per image and then the feature maps are concatenated together to estimate the slip probability P_{slip} through a MLP.
- **Training:** To train the network, we collected approximately 20 minutes of tactile data from four sensors. The dataset comprises 40% slip samples and 60% non-slip samples, with each frame manually annotated. We choose binary cross-entropy as our loss function.
- **Inference:** In inference, it is necessary to set a threshold for P_{slip} for positive detection. Through manual adjustments, empirical results demonstrate that a threshold of 0.75 yields an optimal trade-off between sensitivity and accuracy, achieving a slip detection accuracy of 86%.

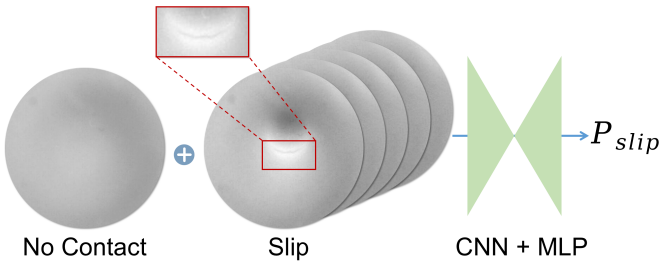

Fig. 3: **Slip detection.** The left tactile image shows no contact, while the middle tactile image highlights wrinkle features during slip. The network computes the probability of slipping P_{slip} using the no-contact tactile image and five most recent tactile images.

B. Robotic Hand System

We integrated the proposed R-Tac sensors into a fully actuated dexterous robotic hand. These tactile sensors are mounted at the distal end of each fingertip, facilitating contact characterization in the following paper-picking tasks. We designed and fabricated the robotic hand featuring 16 controllable DoFs, including the DIP, PIP, and MCP, MCP-2 joints for the index, middle, and ring fingers, as well as the CMC, CMC-2, MCP, and IP joints for the thumb. The robotic hand is driven by Dynamixel XC330-M288-T motors, which are all multiplexed through a U2D2 Hub. For each tactile sensor, it communicates with the PC via a USB interface. The entire assembly is mounted on a Franka Research 3, a 7-DoF robotic arm, which communicates with the PC via a high-speed Ethernet connection.

IV. Paper-like Object Picking Problem Statement

Next, we aim to address the challenge of grasping thin, deformable paper-like objects from flat surfaces. This appears as a commonly seen scenario in everyday tasks, such as organizing scattered document pages or retrieving napkins from dining plates. Although creases or irregularities in the material can sometimes provide grasping points, a particularly challenging scenario arises when the object is extremely flat and lacks discernible edges or salient grasping features. This research introduces a novel approach to tackle this paper-picking problem that was previously unexplored.

Motivated by the human strategy for grasping flat objects, our work is based on a biomimetic grasping pose optimized for paper picking, as illustrated in Fig. 4. By applying sufficient inward force, the robotic fingers can induce buckling of the material against the supporting surface. This buckling effect dynamically generates a pinchable region, enabling subsequent grasp execution.

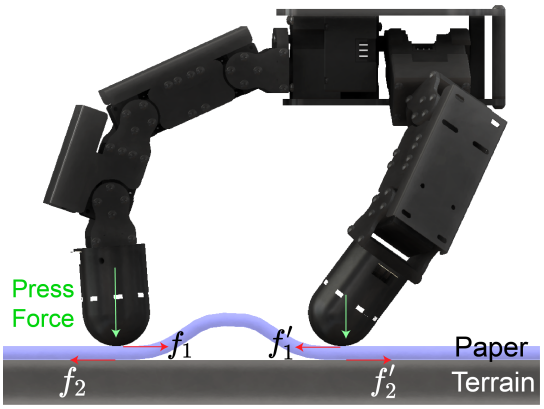

Fig. 4: **Force analysis during grasping flat objects.** The grasping process is made possible by the following forces: (1) the contact normal force exerted by the sensor on the object. (2) the static friction force (f_1, f_1') between fingers and the object, (3) a dynamic friction force (f_2, f_2') between the object and the terrain. When the static friction (f_1, f_1') exceeds the critical buckling resistance of the paper, the sheet deforms, creating a stable pinch region that facilitates successful grasping.

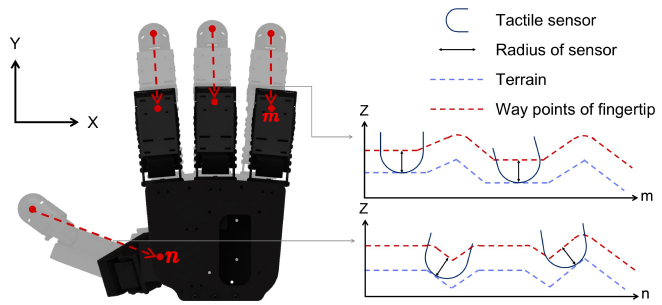

Fig. 5: **Fingertip trajectories from data synthesis.** Trajectories ensure fingertip sliding along the terrain surface. Adjusting the distance between waypoints and terrain affects sensor deformation. The right figure projects trajectories of two fingers onto the m-z and n-z planes, where m and n are straight-line projections of fingertip trajectories on the palm-aligned x-y plane, and the z-axis extends outward from the hand.

During buckling, the distance between contact points beneath the fingers decreases. When this reduction rate matches the fingertips' closure speed (*i.e.*, no relative motion between fingertips and material), two frictional forces govern the system: static friction (f_1, f_1') between the fingers and material, and dynamic friction (f_2, f_2') between the material and the supporting surface. Their magnitudes depend on the applied normal force and the respective coefficients of friction.

In particular, the above analysis assumes that the static friction between robotic fingers and the material exceeds both the maximum static friction at the material-terrain interface and the critical buckling resistance of the material. This framework can also be extended to scenarios with uneven supporting surfaces. Without loss of generality, we assume that height variations in the terrain are less than 3 cm.

One challenge is determining the control inputs for all joints and the wrist pose. Intuitively, this resembles human grasping behavior: when picking up a sheet of paper from a flat surface, the wrist must first elevate and then lower to establish stable contact. However, a finger-wrist coupling issue arises: the motion of one finger requires a specific wrist state, which in turn affects the movement of other fingers. In practice, our approach solved this problem by adopting a learning-based policy rather than a model-based optimization paradigm. This is due to its superior efficiency in deployment, as we found model-based optimization is too computationally expensive to adapt for online execution.

V. Policy Learning for Paper-Picking

Manipulating paper-like objects with visual perception remains challenging due to difficulties in detecting thickness and textural variability. To address this, we propose a vision-independent tactile-based approach. The core idea leverages tactile feedback to maintain contact conditions (as defined in Section IV), facilitating the creation of a buckling region for successful grasping. We implement this through the *PP-Tac policy*, developed in two stages: 1) Trajectory Optimization: Generate a dataset of grasping motions using trajectory optimization. 2) Diffusion Policy Training: Train a policy on this dataset to infer motions from tactile feedback and proprioceptive states, ensuring generalization to real-world robotic systems.

A. Grasp Motion Dataset Synthesis

We synthesize grasping motions through trajectory optimization in simulation, avoiding the need for complex teleoperation devices. While reinforcement learning (RL) offers an alternative, it requires soft-body simulation to model deformable object dynamics and VBTS elastomer behavior, often necessitating additional real-to-sim procedures for fidelity. In contrast, our approach uses rigid-body dynamics and transfers directly to real robots, as validated experimentally. The grasping process begins by establishing contact between the fingertips and the object's surface (see Appendix B for implementation details). Once contact is achieved, the fingers gradually close to pinch the object. Each finger follows an independent trajectory on the object's surface, and concurrently exerts target normal forces (Figs. 4 and 5).

To generate diverse fingertip trajectories, we first generated randomized terrain profiles and pre-recorded a grasping motion sequence. The grasping motion sequences were obtained through teleoperation, capturing natural and usual grasping motion. As illustrated in Figs. 4 and 5, we then extracted the (x,y) coordinates of the fingertip trajectories from the pre-recorded motion to serve as target positions. The corresponding z-coordinates were obtained by projecting these (x,y) points onto the terrain surface and sampling the z-value at each location.

To account for varying material properties, sliding trajectories are adjusted based on fingertip contact forces F. This is achieved by synthesizing motions that deform the tactile sensor's elastomer layer to different extents. The deformation

reading $\boldsymbol{d}_{tac}$ is proportional to the applied pressure, governed by the elastomers Young's modulus. Thus, contact force F is modulated by controlling $\boldsymbol{d}_{tac}$ via position control. Notably, the exact relationship between $\boldsymbol{d}_{tac}$ and F is not explicitly modeled, as precise force values are unnecessary for the algorithm. Our approach leverages rigid-body dynamics to control contact forces efficiently, avoiding complex deformable dynamics calculations. By adjusting the distance between the finger joint and the terrain, we can obtain trajectories with varying degrees of deformation. For example, when the distance between the finger joint and the terrain is equal to the sensor's radius (Fig. 5), the finger just makes contact with the terrain and $\boldsymbol{d}_{tac}$ just equals to 0. Finally, we get the target trajectories of 4 fingertips, with their positions denoted as $\boldsymbol{ee}_{target}$.

Given $\boldsymbol{ee}_{target}$, all of the finger joint angles and arm poses are solved through the following optimization problem:

$$\hat{\boldsymbol{\gamma}} = \arg\min_{\boldsymbol{\gamma}} \left(L_{\boldsymbol{ee}} + L_{\Delta} + L_{\boldsymbol{R},\boldsymbol{p}_{wrist}} \right), \tag{2}$$

$$L_{\boldsymbol{ee}} = w_{\boldsymbol{ee}} \ \mathbf{MSE}(\mathbf{fk}(\boldsymbol{\gamma}), \boldsymbol{ee}_{target}), \tag{3}$$

$$L_{\Delta} = w_{\Delta} \ \mathbf{MSE}\left(\bar{\boldsymbol{\gamma}}, \boldsymbol{\gamma}\right), \tag{4}$$

$$L_{\boldsymbol{R},\boldsymbol{p}_{wrist}} = w_{\boldsymbol{R},\boldsymbol{p}_{wrist}} \ \mathbf{MSE}\left((\bar{\boldsymbol{R}}, \bar{\boldsymbol{p}}_{wrist}), (\boldsymbol{R}, \boldsymbol{p}_{wrist})\right), \tag{5}$$

where $\boldsymbol{\gamma}$ is the optimization variables consisting of N_{data} frames' hand joint angles $\boldsymbol{q}$, wrist (end effector of arm) rotation $\boldsymbol{R}$ and wrist translation along the z-axis in world coordinates $\boldsymbol{p}_{wrist}$. N_{data} is the sequence length. The forward kinematics $\mathbf{fk}$ computes the four fingertips' trajectories by giving $\boldsymbol{\gamma}$. $\mathbf{MSE}$ denotes mean squared error. $L_{\boldsymbol{ee}}$ can minimize the error between the fingertip positions and their targets, while L_{Δ} regularizes the motion to remain close to the initial pose. Additionally, $L_{\boldsymbol{R},\boldsymbol{p}_{wrist}}$ minimizes wrist movement, ensuring the arm stays within its workspace. In practice, we choose SGD as our optimizer. After we filtering out the sequence with collision, we generated a dataset of 500,000 grasp samples, each comprising a sequence of $N_{data} = 100$ frames.

B. PP-Tac Policy

Once the dataset is prepared, we employ a diffusion policy to jointly control the hand and arm, enabling adaptation to varying terrain shapes and contact force conditions. We adopt a Denoising Diffusion Probabilistic Model (DDPM) framework [**?** 16, 8, 39], which predicts future actions (N_{pred} steps of x^{pred}) conditioned on historical states (N_{prefix} steps of x^{prefix}). In each frame, the state variables include:

$$(\boldsymbol{p}, \dot{\boldsymbol{p}}, \boldsymbol{q}, \dot{\boldsymbol{q}}, R, \Omega, p_{wrist}, \dot{p}_{wrist}, \boldsymbol{d}_{tac})$$

where $\boldsymbol{p} \in \mathbb{R}^{17\times 3}$ is hand joints' position in world coordinate, $\dot{\boldsymbol{p}} \in \mathbb{R}^{17\times 3}$ is the linear velocity of the hand joints relative to each parent frame, $\boldsymbol{q} \in \mathbb{R}^{16}$ is the rotation angle of controllable hand joints, $\dot{\boldsymbol{q}} \in \mathbb{R}^{16}$ is the angular velocity of controllable hand joints, $R \in \mathbb{R}^{6}$ is 6D rotation (represented as two-row vectors of rotational martix, which is from [**?**]) of wrist(end effector of arm), $\Omega \in \mathbb{R}^{6}$ represents the angular velocity of wrist rotation, $p_{wrist} \in \mathbb{R}$ is the wrist's height along arm's z-axis, $\dot{p}_{wrist} \in \mathbb{R}$ is the linear velocity of p_{wrist}, $\boldsymbol{d}_{tac} \in \mathbb{R}^4$ represents the deformation depth readings from four fingertip tactile sensors. Table II summarizes the notations used in this paper. The total state dimension is $\mathcal{D} = 152$. Such a over-parameterized input allows the network to extract more robust and expressive latent features for the diffusion policy.

The pipeline is illustrated in Fig. 6. Fig. 6 (right) illustrates a single denoising diffusion step. We apply an encoder-only transformer to predict future robot motion x_0^{pred} given prefix motion x^{prefix}, diffused future motion x_t^{pred}, diffusion step t, current frame index i, and target deformation depth $\bar{\boldsymbol{d}}_{tac}$. The input sequence is encoded into a latent vector of dimension $\mathbb{R}^{(1+N_{prefix}+N_{pred})\times\mathcal{D}}$, comprising: 1) A latent vector of $\mathcal{D}$-dimensional features representing t, i, and $\bar{\boldsymbol{d}}_{tac}$ extracted using a three 3-layer MLP network. 2) $N_{prefix} \times \mathcal{D}$ dimensions corresponding to the prefix states of N_{prefix} time steps. 3) $N_{pred} \times \mathcal{D}$ dimensions for the predicted states of N_{pred} time steps. Instead of predicting ϵ_t (formulated by [16]), we follow [43] to predict the state sequence itself $\hat{x}_0^{pred}$. Predicting $\hat{x}_0^{pred}$ is found to produce better results for the state sequence which contains motion data, and enables us to apply a target loss for each denoising step as following:

$$L = \|\hat{x}_0^{pred} - x_0^{pred}\|_2^2 + \lambda_{consist} L_{consist}, \tag{6}$$

$$L_{consist} = \|\mathbf{fk}(\boldsymbol{q}_0^{pred}) - \boldsymbol{p}_0^{pred}\|_2^2 \tag{7}$$

where $L_{consist}$ enforces consistency between joint angles and positions, and $\lambda_{consist}$ is a weight hyper-parameter.

During inference, we set $t = 1000$ and the diffused $x_{1000}^{pred} \sim \mathcal{N}(0, I)$ and iteratively denoise it to produce x_0^{pred}. To ensure real-time performance, we reduce denoising steps to 10 and set $N_{pred} = N_{prefix} = 5$, achieving motion generation in 11 ms on an RTX4090 GPU. The predicted $\boldsymbol{q}$ controls the hand, while R and p_{wrist} control the arm.

During grasping, preventing slip between the object and the fingertips is essential to maximize material deformation. To achieve this, a fingertip contact force controller is introduced, which adjusts the fingertip's deformation depth $\boldsymbol{d}_{tac}$. If slip is detected by the tactile sensors, we increase the desired deformation depth by a small increment $\Delta\boldsymbol{d}_{tac}$.

To deploy diffusion policy to real robots, we also need to tackle the domain gap between the real world and simulation. This is achieved by introducing four distinct ways to incorporate disturbances into x^{prefix} during training.

- Add random Gaussian noise to $\boldsymbol{\gamma}$ to simulate various control errors that may occur in real-world situations.
- Add Gaussian noise to the first frame and gradually amplify it in subsequent frames, simulating the fingers moving across a rising or descending terrain.
- Randomly choose from 2 to N_{prefix} temporal consistent frames to be static, simulating fingers getting stuck due to excessive pressure on complex terrain. And $\boldsymbol{d}_{tac}$ is set to its maximum threshold. The reason for adding the index of the frame into the input is also to avoid issues caused

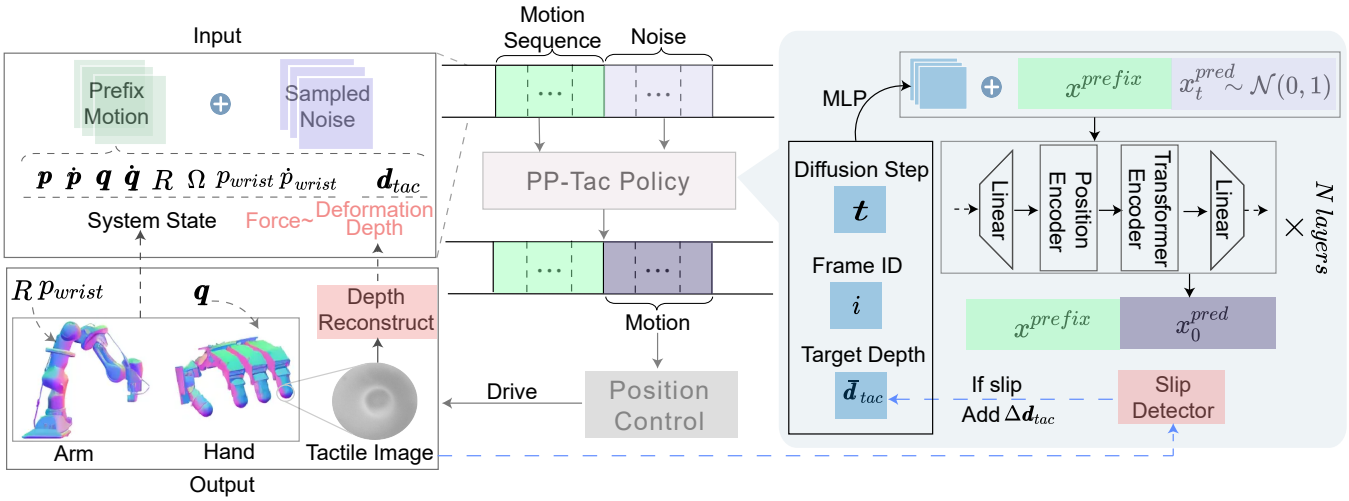

Fig. 6: **Inference pipeline of the proposed PP-Tac policy.** Conditioned on robot proprioception and the target force that needs to be exerted, PP-Tac can infer the action of the next steps. If slip is detected between the finger and the flat object underneath, an incremental amount of force will be exerted by the finger.

by the fingers getting stuck.

VI. Experiments

In this section, we present comprehensive experiments to evaluate our proposed PP-Tac pipeline. First, we detail the implementation of our algorithm (Section VI-A). Next, we show the quantitative and qualitative results of the depth reconstruction of our VBTS (Section VI-B). Then, we perform systematic comparisons of our system on different flat materials and supporting terrains (Section VI-C). We also compare our system with various manipulators to highlight its advantages and limitations (Section VI-D). Last, ablation studies are conducted to examine the influence of parameters and the necessary training steps (Section VI-E).

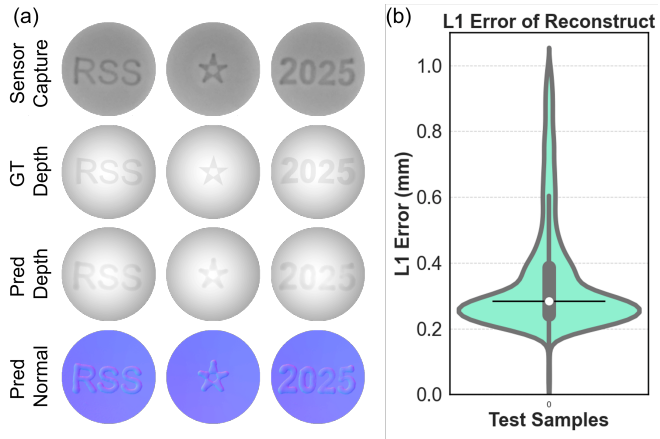

Fig. 7: **Reconstruction results.** (a) Gallery of reconstructed depth and normal maps from tactile images. (b) Depth reconstruction error of indentation test.

A. Implementation Details

For reproducibility, we provide the implementation details of the PP-Tac algorithm. Our diffusion policy is implemented as a four-layer Transformer encoder with a latent dimension of 512 and four attention heads. We split each synthesized data sequence into subsequences of length 10 for the diffusion process, and train the model for approximately 600,000 iterations on a single RTX 4090. During training, the diffusion step t is uniformly sampled from 0 to 1000. During inference, an acceleration technique is applied as follows. First, t is initialized to 1000 and directly denoised to x_0^{pred}. Subsequently, noise is added to the $t = 1000 - 100N_i$ level and denoised again to x_0^{pred}, where N_i is the inference step number. Thus, the entire inference process consists of 10 steps.

For terrain generation, we model the terrain beneath each finger as a cubic spline with a trajectory length of 100. Control points are placed at intervals of 25 along the trajectory, resulting in a total of 5 control points. To simulate ramps, the height of each control point is randomized by sampling uniformly within the range of $[0, 3]$ cm.

B. Depth Reconstruction of VBTS

To evaluate the performance of the tactile sensor in depth reconstruction, the sensor surface is pressed with three indenters, each with the text content "RSS", "★" and "2025". The qualitative results of the sensor output are shown in Fig. 7, which demonstrates the raw captured image from the sensor, the ground truth depth maps, predicted depth maps, and the corresponding calculated normal maps, respectively. These results demonstrate that the sensor can fully reconstruct fine surface details.

We quantify the reconstruction error using a violin plot, leveraging ground truth indentation information obtained from 3D-printed hemispherical shape indicators containing various testing indenters. We collected 215 testing configurations, each with paired sensor outputs and ground truth reprojection images. The sensor achieves a mean absolute error (L1 error) reconstruction loss of 0.35 mm, and a median loss of 0.28 mm, with 60% of reconstruction losses below 0.3 mm. In

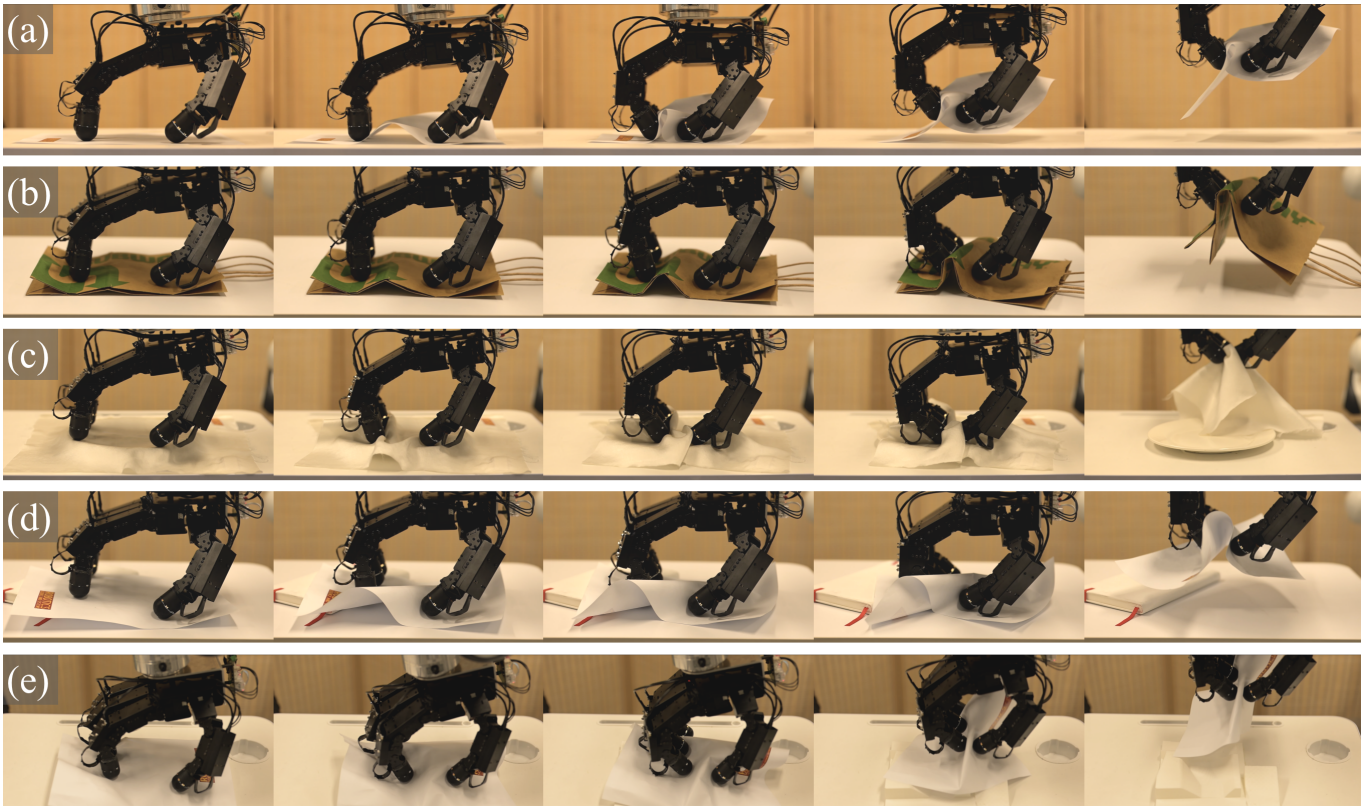

Fig. 8: **Gallery of Grasping Different Objects in Real-World Evaluations.** This figure demonstrates successful grasps of five flat objects on four different types of terrains, highlighting the effectiveness of our hardware and the PP-Tac algorithm. (a) A paper sheet on a flat desktop. (b) A stiff kraft paper bag on a flat desktop. (c) A soft napkin on a plate. (d) A paper sheet on a randomly arranged book. (e) Paper sheet on a random terrain. These evaluations showcase the robustness and adaptability of our approach.

terms of computational speed, the depth mapping process takes less than 10 ms, ensuring real-time performance for robotic applications.

C. Evaluation of PP-Tac Policy on Materials and Terrains

We conducted experiments to evaluate the system's ability to handle flat objects under varying conditions. The qualitative and quantitative results are shown in Fig. 8 and Fig. 9 respectively. Fig. 8 shows the typical successful grasp cases, highlighting that our hardware and PP-Tac algorithm can successfully handle flat objects placed above both the flat and uneven object surface. During the grasping process, the fingertip first contacts the material, followed by a gradual finger closure that buckles the material and creates pinchable regions. Finally, the object is pinched and lifted.

Fig. 9 provides quantitative analysis of the success rate with respect to the object material and the complexity of the terrain beneath. To facilitate this analysis, we conducted experiments using four flat objects in daily life: paper, plastic bag, cloth, and kraft paper bag, each of which presents unique challenges. The paper is extremely flat with no detectable hold points. Plastic bags, commonly encountered in daily life, are difficult to locate using conventional visual pipelines because of their transparency. The cloth is thick and highly deformable, while the kraft paper bags are stiff and have a multilayered structure. To assess the system's robustness, we also varied the terrain beneath the objects. The four types of terrain used include: a flat plane, a slope (10 degrees), a plane with a 2 cm thick book randomly placed on it, and an uneven terrain with random curvatures. The terrain shapes are shown in Fig. 9.

For statistical significance, we performed 20 grasping attempts for each combination of terrain and object. From results in Fig. 9, cloth and plastic bags are relatively easy to grasp due to their low stiffness, which allows them to buckle more easily under force. In contrast, paper and kraft paper bags are stiffer and resist buckling, leading to lower success rates.

The terrain beneath the object also significantly impacts grasp success. On flat terrains, such as a plane or a tilted slope, success rates for paper, plastic bags, and cloth were relatively high. This suggests that flat surfaces usually generate consistent frictional forces essential for a successful grasp. However, this advantage diminishes for stiffer flat objects, such as kraft paper bags. These stiff flat objects usually lack of initial buckling when placed on a flat surface, making it more challenging to form reliable grasp points afterward.

For uneven surfaces, the success rates varied according to the shape of the terrain. When a book was placed underneath the flat object, all objects maintained high success rates. These results can be attributed to the edge of the book and the partial void space created beneath the material, which made it easier

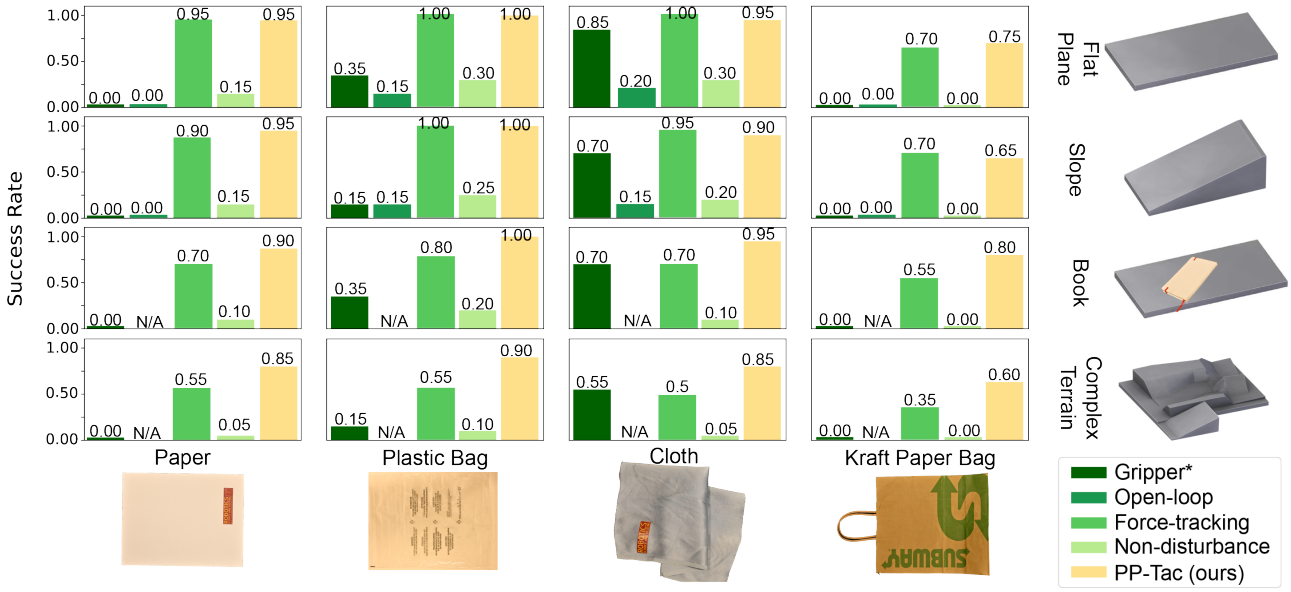

Fig. 9: **Experimental Results.** Evaluations were conducted to quantify the success rate of grasping four different flat objects (paper, plastic bag, cloth, and paper bag) across four terrain setups (plane, slope, book placement, and randomly generated complex terrain). Baseline conditions included: (1) Gripper*: grasp using a bi-finger gripper controlled by teleoperation; (2) Open-loop: baseline combines the PP-Tac-derived hand trajectory with compliant finger control via tactile feedback; (3)"Model based force tracking": combines the PP-Tac-derived hand trajectory with compliant finger control via tactile feedback; (4) Non-disturbance: grasp using our dexterous hand with tactile sensors, where the diffusion policy was trained without domain randomization disturbances; and (5) PP-Tac(ours): grasp using our full PP-Tac pipeline. Each condition was repeated 20 times. Note that open-loop grasp control is not feasible on uncertain terrains, and these cases are labeled as 'N/A'.

for the materials to buckle and separate with the terrain. In contrast, when the terrain was highly irregular, the success rate dropped for all objects. This is likely due to the challenges added to our force controllers, which increased the likelihood of the fingers slipping away from the material.

TABLE I: **Experimental results for varying paper quantities:** The system's performance was evaluated on paper materials with different buckling strengths, achieved by bonding 1, 3, 5, and 7 layers of paper with adhesive. For each configuration, 20 trials of grasps were conducted. The average number of slip events detected (No. Slip) and the final success rate (Succ. Rate) were recorded.

Paper Layers	No. Slip	Succ. Rate (%)
1	0.2	90
3	2.9	75
5	13.3	30
7	18.2	5

D. Comparison with Other System Configurations

To assess whether PP-Tac's system setup leveraging dexterous hand and tactile sensors can offer advantages, systematic comparisons with other robot configurations were conducted. Here, we constructed three baselines. To ensure fairness, each trial allowed only one grasp attempt.

- Bi-finger grippers controlled via human teleoperation with a camera mounted on the wrist to provide an egocentric view which can mimic the vision-based method [8]. This baseline can demonstrate the effectiveness of our hardware design.
- Open-loop control without tactile feedback: we pre-generated trajectories using the ground truth shape of the terrain and then replayed these trajectories rather than using the PP-Tac policy. Note that this trajectory-replay setting is unattainable in scenarios with high variations, such as the book setting and the complex terrain scenario in which the terrain shape is unknown.
- "Model based force tracking": due to the challenges outlined in Section IV, we employ the wrist trajectory generated by PP-Tac while actively controlling only the fingertips through real-time tactile feedback.

The evaluation results in Fig. 9 show that the PP-Tac pipeline outperforms all baselines. We observed that the teleoperation baseline using a gripper achieved some successful cases in grasping cloth and plastic bags, albeit with lower performance than PP-Tac. This is due to the ease of detecting the initial grasp point on these soft materials through human perception, and combined with human intelligence enabling grasp adjustments through visual feedback. However, for stiffer materials like paper and kraft paper, the bi-finger gripper failed completely. Therefore, we conclude that the PP-Tac pipeline is the most suitable configuration for handling flat objects. The open-loop baseline achieved a lower success rate compared to PP-Tac. The suboptimal performance primarily stems from control error. As mentioned in [37], Allegro Hand

[34] exhibits joint angle errors exceeding 0.1 radians, which will be further accumulated across the kinematic chain. These errors critically degrade performance in precision-sensitive tasks such as paper picking, highlighting the necessity of tactile feedback for robust control. While the "Model-based force tracking" achieves satisfactory performance in structured terrains by leveraging wrist trajectories generated by PP-Tac, its effectiveness becomes limited when confronted with irregular or complex terrains. This underscores the need for enhanced adaptability in unstructured environments.

E. Ablation Studies

1) Influence of Material Stiffness: We found that the material's stiffness (represented by its thickness), significantly influences the task's success rate. To demonstrate this effect, we created flat objects by stacking paper pages bonded with adhesive. The experimental results are shown in Table I. As the number of paper pages increased, the grasp success rate decreased significantly. Additionally, the increase in material stiffness also led to a higher number of detected slips.

2) Influence of Data Disturbance: We emphasize the importance of the data disturbance technique for domain randomization (introduced in Section V-B). To quantify its impact, we conducted ablation studies comparing grasp performance before and after adding four types of disturbances to the prefix motion x^{prefix}. Experimental results demonstrate that this technique significantly enhances performance. As shown in the "Non-disturbance" baseline in Section VI-C, removing data disturbance led to a notable performance drop across all experiments, often resulting in complete failure when grasping stiff objects, such as kraft paper bags. This underscores the improved generalization and higher grasp success rates enabled by domain randomization. However, a drawback of this technique is the increased training time, requiring approximately 400,000 additional iterations to achieve the same loss as training without data disturbance.

VII. Limitations

We have observed the following limitations in our system. One limitation is determining the initial force (sensor's target deformation depth) required for successful grasping. While our algorithm can adaptively adjust the force magnitude online, an appropriate initial value must still be manually set, which remains an empirical parameter-tuning process. If the initial value is too small, the grasp is more likely to fail due to the additional time and finger sliding distance needed for adaptation to a reasonable value. Conversely, if the initial value is too large, excessive friction may exceed the load capacity of the hand motors. In addition to the initial value, the adaptive algorithm for adjusting force also has room for improvement, particularly with highly stiff materials such as kraft paper bags on non-flat surfaces.

VIII. Conclusions

This paper presents PP-Tac, a coordinated hand-arm system designed to manipulate thin, flat objects such as paper and fabric. The system is equipped with a multi-fingered, vision-based tactile sensor that is easy to fabricate and deploy on the hand's fingertips. The sensor can detect contact on its curved surfaces, enabling the system to measure force and friction during contact. This capability helps minimize slip and increases the likelihood of material deformation when handling flat materials. Based on this hand design, the grasping motion is planned using a data-driven approach. We developed an efficient synthesis algorithm to generate sliding trajectories across various terrain shapes and sensor deformation conditions, resulting in a dataset of 500,000 trajectory samples. Using this dataset and a domain randomization technique, we trained a diffusion policy that enables adaptation to diverse terrains in real-world settings. Experimental results show that our system can successfully grasp flat objects of varying thicknesses and stiffness, achieving a success rate of 87.5%. Additionally, the proposed policy demonstrates robustness to external disturbances and adapts well to different support terrain surfaces.

IX. Acknowledgment

We thank Changyi Lin (Carnegie Mellon University) for insightful discussions. This work was supported in part by the National Natural Science Foundation of China (Grant No.52305007), by the State Key Laboratory of Mechanical System and Vibration (Grant No. MSV202519), and by Shanghai Frontiers Science Center of Human-centered Artificial Intelligence (ShangHAI), MoE Key Laboratory of Intelligent Perception and Human-Machine Collaboration (KLIP-HuMaCo).

References

[1] Enrique Amigó, Julio Gonzalo, and Felisa Verdejo. A general evaluation measure for document organization tasks. In *Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 643–652, 2013.

[2] Iris Andrussow, Huanbo Sun, Katherine J Kuchenbecker, and Georg Martius. Minsight: A fingertip-sized vision-based tactile sensor for robotic manipulation. *Advanced Intelligent Systems*, 5(8):2300042, 2023.

[3] Veronica E Arriola-Rios and Jeremy L Wyatt. A multimodal model of object deformation under robotic pushing. *IEEE Transactions on Cognitive and Developmental Systems*, 9(2):153–169, 2017.

[4] Osher Azulay, Nimrod Curtis, Rotem Sokolovsky, Guy Levitski, Daniel Slomovik, Guy Lilling, and Avishai Sintov. Allsight: A low-cost and high-resolution round tactile sensor with zero-shot learning capability. *IEEE Robotics and Automation Letters (RA-L)*, 9(1):483–490, 2023.

[5] Aude Billard and Danica Kragic. Trends and challenges in robot manipulation. *Science*, 364(6446):eaat8414, 2019.

[6] Tara Boroushaki, Junshan Leng, Ian Clester, Alberto Rodriguez, and Fadel Adib. Robotic grasping of fully-

occluded objects using rf perception. In *International Conference on Robotics and Automation (ICRA)*, pages 923–929. IEEE, 2021.

[7] Arkadeep Narayan Chaudhury, Timothy Man, Wenzhen Yuan, and Christopher G Atkeson. Using collocated vision and tactile sensors for visual servoing and localization. *IEEE Robotics and Automation Letters (RA-L)*, 7(2):3427–3434, 2022.

[8] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *International Journal of Robotics Research (IJRR)*, page 02783649241273668, 2023.

[9] Xinke Deng, Yu Xiang, Arsalan Mousavian, Clemens Eppner, Timothy Bretl, and Dieter Fox. Self-supervised 6d object pose estimation for robot manipulation. In *International Conference on Robotics and Automation (ICRA)*, pages 3665–3671. IEEE, 2020.

[10] Won Kyung Do and Monroe Kennedy. Densetact: Optical tactile sensor for dense shape reconstruction. In *International Conference on Robotics and Automation (ICRA)*, pages 6188–6194. IEEE, 2022.

[11] Won Kyung Do, Bianca Aumann, Camille Chungyoun, and Monroe Kennedy. Inter-finger small object manipulation with densetact optical tactile sensor. *IEEE Robotics and Automation Letters (RA-L)*, 2023.

[12] Mehmet Remzi Dogar and Siddhartha S Srinivasa. A framework for push-grasping in clutter. In *Proceedings of Robotics: Science and Systems (RSS)*, volume 2, 2011.

[13] Siyuan Dong, Daolin Ma, Elliott Donlon, and Alberto Rodriguez. Maintaining grasps within slipping bounds by monitoring incipient slip. In *International Conference on Robotics and Automation (ICRA)*, pages 3818–3824. IEEE, 2019.

[14] Satoshi Funabashi, Tomoki Isobe, Shun Ogasa, Tetsuya Ogata, Alexander Schmitz, Tito Pradhono Tomo, and Shigeki Sugano. Stable in-grasp manipulation with a low-cost robot hand by using 3-axis tactile sensors with a cnn. In *International Conference on Intelligent Robots and Systems (IROS)*, pages 9166–9173. IEEE, 2020.

[15] Rafael Herguedas, Gonzalo López-Nicolás, Rosario Aragüés, and Carlos Sagüés. Survey on multi-robot manipulation of deformable objects. In *Proceedings of IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*, pages 977–984. IEEE, 2019.

[16] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

[17] Mohsen Kaboli, Rich Walker, Gordon Cheng, et al. In-hand object recognition via texture properties with robotic hands, artificial skin, and novel tactile descriptors. In *Proceedings of IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, pages 1155–1160. IEEE, 2015.

[18] Justin Kerr, Huang Huang, Albert Wilcox, Ryan Hoque, Jeffrey Ichnowski, Roberto Calandra, and Ken Goldberg. Self-supervised visuo-tactile pretraining to locate and follow garment features. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023.

[19] Gagan Khandate, Siqi Shang, Eric T. Chang, Tristan Luca Saidi, Yang Liu, Seth Matthew Dennis, Johnson Adams, and Matei Ciocarlie. Sampling-based Exploration for Reinforcement Learning of Dexterous Manipulation. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023.

[20] Michael Krawez, Tim Caselitz, Jugesh Sundram, Mark Van Loock, and Wolfram Burgard. Real-time outdoor illumination estimation for camera tracking in indoor environments. *IEEE Robotics and Automation Letters (RA-L)*, 6(3):6084–6091, 2021.

[21] Mike Lambeta, Tingfan Wu, Ali Sengul, Victoria Rose Most, Nolan Black, Kevin Sawyer, Romeo Mercado, Haozhi Qi, Alexander Sohn, Byron Taylor, et al. Digitizing touch with an artificial multimodal fingertip. *arXiv preprint arXiv:2411.02479*, 2024.

[22] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. Ep n p: An accurate o (n) solution to the p n p problem. *International Journal of Computer Vision (IJCV)*, 81:155–166, 2009.

[23] Mengdi Li, Cornelius Weber, Matthias Kerzel, Jae Hee Lee, Zheni Zeng, Zhiyuan Liu, and Stefan Wermter. Robotic occlusion reasoning for efficient object existence prediction. In *International Conference on Intelligent Robots and Systems (IROS)*, pages 2686–2692. IEEE, 2021.

[24] Rui Li, Robert Platt, Wenzhen Yuan, Andreas Ten Pas, Nathan Roscup, Mandayam A Srinivasan, and Edward Adelson. Localization and manipulation of small parts using gelsight tactile sensing. In *International Conference on Intelligent Robots and Systems (IROS)*, pages 3988–3993. IEEE, 2014.

[25] Yinxiao Li, Danfei Xu, Yonghao Yue, Yan Wang, Shih-Fu Chang, Eitan Grinspun, and Peter K Allen. Regrasping and unfolding of garments using predictive thin shell modeling. In *International Conference on Robotics and Automation (ICRA)*, pages 1382–1388. IEEE, 2015.

[26] Yunzhu Li, Jiajun Wu, Russ Tedrake, Joshua B Tenenbaum, and Antonio Torralba. Learning particle dynamics for manipulating rigid bodies, deformable objects, and fluids. *arXiv preprint arXiv:1810.01566*, 2018.

[27] Changyi Lin, Ziqi Lin, Shaoxiong Wang, and Huazhe Xu. Dtact: A vision-based tactile sensor that measures high-resolution 3d geometry directly from darkness. In *International Conference on Robotics and Automation (ICRA)*, pages 10359–10366. IEEE, 2023.

[28] Changyi Lin, Han Zhang, Jikai Xu, Lei Wu, and Huazhe Xu. 9dtact: A compact vision-based tactile sensor for accurate 3d shape reconstruction and generalizable 6d force estimation. *IEEE Robotics and Automation Letters*

(RA-L), 2023.

[29] Pei Lin, Sihang Xu, Hongdi Yang, Yiran Liu, Xin Chen, Jingya Wang, Jingyi Yu, and Lan Xu. Handdiffuse: Generative controllers for two-hand interactions via diffusion models, 2023. URL https://arxiv.org/abs/2312.04867.

[30] Xiaofei Liu, Wuqiang Yang, Fan Meng, and Tengchen Sun. Material recognition using robotic hand with capacitive tactile sensor array and machine learning. *Transactions on Instrumentation and Measurement (TIM)*, 2024.

[31] Ashvin Nair, Dian Chen, Pulkit Agrawal, Phillip Isola, Pieter Abbeel, Jitendra Malik, and Sergey Levine. Combining self-supervised learning and imitation for vision-based rope manipulation. In *International Conference on Robotics and Automation (ICRA)*, pages 2146–2153. IEEE, 2017.

[32] Kei Ota, Devesh K Jha, Hsiao-Yu Tung, and Joshua Tenenbaum. Tactile-Filter: Interactive Tactile Perception for Part Mating. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023.

[33] Haozhi Qi, Brent Yi, Sudharshan Suresh, Mike Lambeta, Yi Ma, Roberto Calandra, and Jitendra Malik. General in-hand object rotation with vision and touch. In *Conference on Robot Learning (CoRL)*, pages 2549–2564. PMLR, 2023.

[34] Wonik Robotics. Allegro Hand, 2024. URL https://www.allegrohand.com/ah-v4-main.

[35] Jose Sanchez, Juan-Antonio Corrales, Belhassen-Chedli Bouzgarrou, and Youcef Mezouar. Robotic manipulation and sensing of deformable objects in domestic and industrial applications: a survey. *International Journal of Robotics Research (IJRR)*, 37(7):688–716, 2018.

[36] Brian Scassellati, Henny Admoni, and Maja Matarić. Robots for use in autism research. *Annual Review of Biomedical Engineering*, 14(1):275–294, 2012.

[37] Kenneth Shaw, Ananye Agarwal, and Deepak Pathak. Leap hand: Low-cost, efficient, and anthropomorphic hand for robot learning. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023.

[38] Yu She, Shaoxiong Wang, Siyuan Dong, Neha Sunil, Alberto Rodriguez, and Edward Adelson. Cable manipulation with a tactile-reactive gripper. *International Journal of Robotics Research (IJRR)*, 40(12-14):1385–1401, 2021.

[39] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021.

[40] Li Sun, Gerardo Aragon-Camarasa, Simon Rogers, and J Paul Siebert. Accurate garment surface analysis using an active stereo robot head with application to dual-arm flattening. In *International Conference on Robotics and Automation (ICRA)*, pages 185–192. IEEE, 2015.

[41] Sudharshan Suresh, Haozhi Qi, Tingfan Wu, Taosha Fan, Luis Pineda, Mike Lambeta, Jitendra Malik, Mrinal Kalakrishnan, Roberto Calandra, Michael Kaess, et al. Neuralfeels with neural fields: Visuotactile perception for in-hand manipulation. *Science Robotics*, 9(96):eadl0628, 2024.

[42] Ian H Taylor, Siyuan Dong, and Alberto Rodriguez. Gelslim 3.0: High-resolution measurement of shape, force and slip in a compact tactile-sensing finger. In *International Conference on Robotics and Automation (ICRA)*, pages 10781–10787, 2022.

[43] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *International Conference on Learning Representations (ICLR)*, 2023.

[44] Megha H Tippur and Edward H Adelson. Rainbowsight: A family of generalizable, curved, camera-based tactile sensors for shape reconstruction. In *International Conference on Robotics and Automation (ICRA)*, pages 1114–1120. IEEE, 2024.

[45] Benjamin Ward-Cherrier, Nicholas Pestell, Luke Cramphorn, Benjamin Winstone, Maria Elena Giannaccini, Jonathan Rossiter, and Nathan F Lepora. The tactip family: Soft optical tactile sensors with 3d-printed biomimetic morphologies. *Soft robotics*, 5(2):216–227, 2018.

[46] Xiaolong Wu and Cédric Pradalier. Illumination robust monocular direct visual odometry for outdoor environment mapping. In *International Conference on Robotics and Automation (ICRA)*, pages 2392–2398. IEEE, 2019.

[47] Zhengyou Zhang. A flexible new technique for camera calibration. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 22(11):1330–1334, 2002.

[48] Jihong Zhu, Andrea Cherubini, Claire Dune, David Navarro-Alarcon, Farshid Alambeigi, Dmitry Berenson, Fanny Ficuciello, Kensuke Harada, Jens Kober, Xiang Li, et al. Challenges and outlook in robotic manipulation of deformable objects. *IEEE Robotics and Automation Magazine (RA-M)*, 29(3):67–77, 2022.

APPENDIX

A. Detail of Camera Calibration

In this section, we introduce the camera calibration process as part of the overall sensor calibration. Since the tactile sensor is enclosed by an opaque, rounded membrane, conventional calibration board methods cannot be used to determine the pinhole camera's extrinsic parameters. To address this, we designed an indentation setup (as shown in Fig. 10) to capture a sufficient number of spatial points in a known sensor frame, identify their corresponding 2D-pixel coordinates in the image, and establish the mapping between the sensor frame and the image frame. First, the camera's intrinsic parameters K was obtained, either from the camera manufacturer or calibrated using high-precision calibration boards [47]. Next, we define a three-dimensional coordinate system, referred to as the sensor frame (x, y, z) with its origin at the center of the elastomer, as shown in Fig. 10(a). To facilitate the calibration, A custom 3D-printed holder secures the sensor (Fig. 10(b)), while another

3D-printed hemispherical indicator is attached to the holder's groove (Fig. 10(c)). Small pins with a diameter of 1.5mm, serving as indenters, are inserted into pre-defined holes within the indicator for 28 trials. For each trail, the contact positions are recorded both in the camera image as $p_{ij} = (u_{ij}, v_{ij})$ and in the sensor frame as $P_{i,j} = (x_{ij}, y_{ij}, z_{ij})$, where i denotes the trail index and j denotes the contact point index within the trail. The contact positions in the camera image are detected by subtracting the captured image from a reference image without indentation. We use solvePnP [22] to calculate the extrinsic parameters that includes rotation matrix A and translation vector b such that:

$$p_{ij} = K[A \mid b]P_{ij} \tag{8}$$

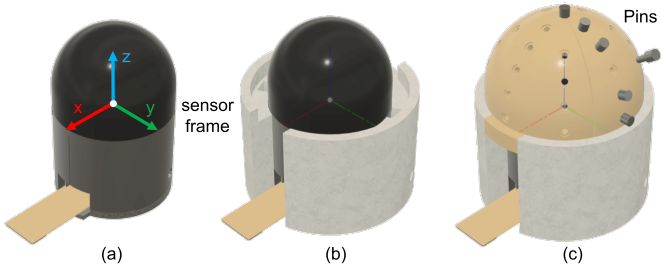

Fig. 10: **Camera calibration using an indentation setup:** The sensor frame is first defined in (a). A holder is designed and 3D-printed to secure the sensor, as shown in (b). A hemispherical indicator is designed and 3D-printed to attach to the sensor holder. Pins are inserted into pre-defined holes to serve as indenters for recording contact locations in the sensor frame, as shown in (c).

After obtaining the intrinsic and extrinsic parameters of the camera, we can project the sensor's curved surface from the sensor frame onto the image frame, obtaining the sensor surface reference projection D (Equation (9)), by which the depth value on the pixel (u, v) can be queried.

$$D(u,v) = \left[Z_c K^{-1} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \right]_{[3,:]}, \tag{9}$$

where $[u\ v\ 1]^T$ and Z_c are given as:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = K \begin{bmatrix} \frac{(A[x,y,z]^T + b)_x}{Z_c} \\ \frac{(A[x,y,z]^T + b)_y}{Z_c} \\ 1 \end{bmatrix}, Z_c = (A[x,y,z]^T + b)_z. \tag{10}$$

B. Detail of Establish Contact

In this section, we detail our approach to generate contact with a flat object using the fingertips. The goal is to control the hand to ensure that at least three fingertips are in contact with the surface. We denote the four fingertips as f_t (thumb), f_i (index), f_m (middle), and f_r (ring). The contact states are represented by two sets: $\boldsymbol{C}$, which includes the fingers in contact, and $\boldsymbol{N}$, which includes the fingers not in contact. The complete process is illustrated in Fig. 11.

1) ***Establish First Contact:*** Starting from status when all fingers are hovering (i.e., $\boldsymbol{C} = \phi, \boldsymbol{N} = \{f_t, f_i, f_m, f_r\}$), the hand is controlled to move downward till one finger touches the surface. For example, if the thumb touches the surface (Fig. 11), the contact state sets are updated to $\boldsymbol{C} = \{f_t\}, \boldsymbol{N} = \{f_i, f_m, f_r\}$

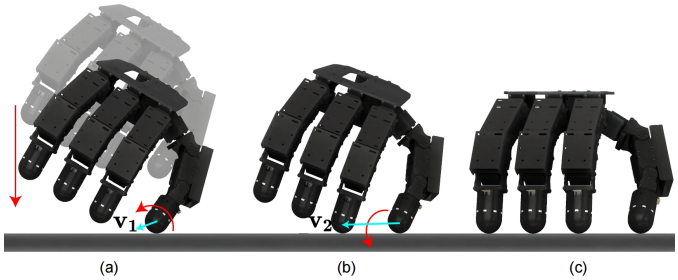

Fig. 11: **Example of establishing contact:** First, the hand descends until a finger makes contact with the surface. A fixed-point rotation is performed around the contacting finger, as shown in (a). The hand then continues to rotate until a second finger makes contact, triggering a fixed-axis rotation around both contacting fingers, as shown in (b). The process is complete when three or more fingers are in contact, as shown in (c).

2) ***Establish Second Contact:*** Once the first contact is made, the hand rotates around the first finger's contact point to create the second contact point. To achieve this, we first obtain the centroid point of the fingertip in contact (denoted as (x_c, y_c, z_c)), and compute the centroid point of fingertip positions in $\boldsymbol{N}$ (denoted as (x_n, y_n, z_n)). This allows us to calculate the rotational axis as:

$$\boldsymbol{v_1} = R_z(90°)(x_n - x_c, y_n - y_c, z_n - z_c)^T, \tag{11}$$

where $R_z(90°)$ is the rotation matrix for a 90-degree rotation around the z-axis. Given $\theta, \boldsymbol{v_1}$ calculated before, robot arm's target end-effector pose ${}^{b}_{ee'}T$ leading to such rotation can be obtained via Rodrigues' rotation formula:

$$\begin{aligned} R(\theta, \boldsymbol{v_1}) =& I + \sin(\theta) \begin{bmatrix} 0 & -v_{1z} & v_{1y} \\ v_{1z} & 0 & -v_{1x} \\ -v_{1y} & v_{1x} & 0 \end{bmatrix} + \\ & (1 - \cos(\theta)) \begin{bmatrix} 0 & -v_{1z} & v_{1y} \\ v_{1z} & 0 & -v_{1x} \\ -v_{1y} & v_{1x} & 0 \end{bmatrix}^2, \end{aligned} \tag{12}$$

The target end effector pose of the robot arm can be calculated as:

$${}^{b}_{ee'}T = {}^{b}_{ee}T\ {}^{ee}_{c}T\ {}^{c}_{c'}\hat{T}\ {}^{c'}_{ee'}T, \tag{13}$$

$${}^{c}_{c'}\hat{T} = \begin{bmatrix} R(\theta, \boldsymbol{v_1}) & 0 \\ 0 & 1 \end{bmatrix}, \tag{14}$$

where b denotes the base of the robot arm, ee and ee' represent the end effector before and after the movement, and c and c' represent the positions (x_c, y_c, z_c) before and after the rotation. The robot arm is then controlled to gradually increase θ until the second fingertip contacts the object surface. Once this occurs, we update the contact states to $\boldsymbol{C} = \{f_t, f_i\}$ and $\boldsymbol{N} = \{f_m, f_r\}$.

*3) **Establishing Third Contact**:* In this step, the hand rotates around an axis defined by the first and second contact points until the third fingertip makes contact. For instance, if the thumb and index finger make contact, the rotation axis is $\boldsymbol{v_2} = \overrightarrow{f_t f_i}$. The arm's target end-effector pose for this rotation is:

$$ {}^{b}_{ee''}T = {}^{b}_{ee'}T \ {}^{ee'}_{c'}T \ {}^{c'}_{c''}\hat{T} \ {}^{c''}_{ee''}T, \tag{15} $$

$$ {}^{c'}_{c''}\hat{T} = \begin{bmatrix} R(\theta', \boldsymbol{v_2}) & 0 \\ 0 & 1 \end{bmatrix}, \tag{16} $$

where c'' and ee'' are c' and ee' after rotation specified by $\boldsymbol{v_2}$. During execution, the angle θ' is gradually increased until a new fingertip contacts the surface, achieving the desired target end-effector pose ${}^{b}_{ee'}T$. Note that these steps may not always be required. In some cases, we observe that the third finger may already be in the contact state when we attempt to establish contact with the second finger.

C. List of Symbols

The definition of symbols can be found in Table II.

TABLE II: Summary of symbols and notations.

Symbols	Descriptions
u, v	Pixel coordinates in VBTS.
X_c, Y_c, Z_c	Camera coordinates in VBTS.
x, y, z	Sensor coordinates in VBTS.
K	The intrinsic parameters of the camera in VBTS.
A, b	The extrinsic parameters of the camera in VBTS.
D	Sensor surface reference projection in VBTS.
M	Depth mapping function in VBTS.
$\boldsymbol{q}$	Rotation angle of controllable hand joints.
$\dot{\boldsymbol{q}}$	Angular velocity of controllable hand joints.
$\boldsymbol{p}$	Positional coordinate of hand joints in arm's base axis.
$\dot{\boldsymbol{p}}$	Linear velocity of hand joints in arm's base axis.
R	Wrist's (end effector of arm) 6D rotation.
Ω	Angular velocity of hand pose.
p_{wrist}	Wrist (end-effector of arm)'s height along arm's z-axis.
$\dot{p}_{wrist}$	Linear velocity of $\boldsymbol{p}_{ee}$.
$\boldsymbol{d}_{tac}$	The deformation depth readings from four fingertip tactile sensors.
$\bar{\boldsymbol{d}}_{tac}$	The target deformation depth.
$\mathcal{D}$	State variable's dimension.
γ	Hand joint angles $\boldsymbol{q}^{1:N_{data}}$, wrist's (end effector of arm) 6D rotation $R^{1:N_{data}}$ and wrist's translation along z-axis $p_{ee}^{1:N_{data}}$ for overall trajectory.
N_{data}	Length of synthesis motion sequence.
N_{pred}	Length of predicted actions.
x^{pred}	Future motion predicted by PP-Tac policy.
N_{prefix}	Length of historical actions.
x^{prefix}	The historical action sequence.
t	Diffusion step.