Robotics: Science and Systems 2025
Los Angeles, CA, USA, June 21–25, 2025

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

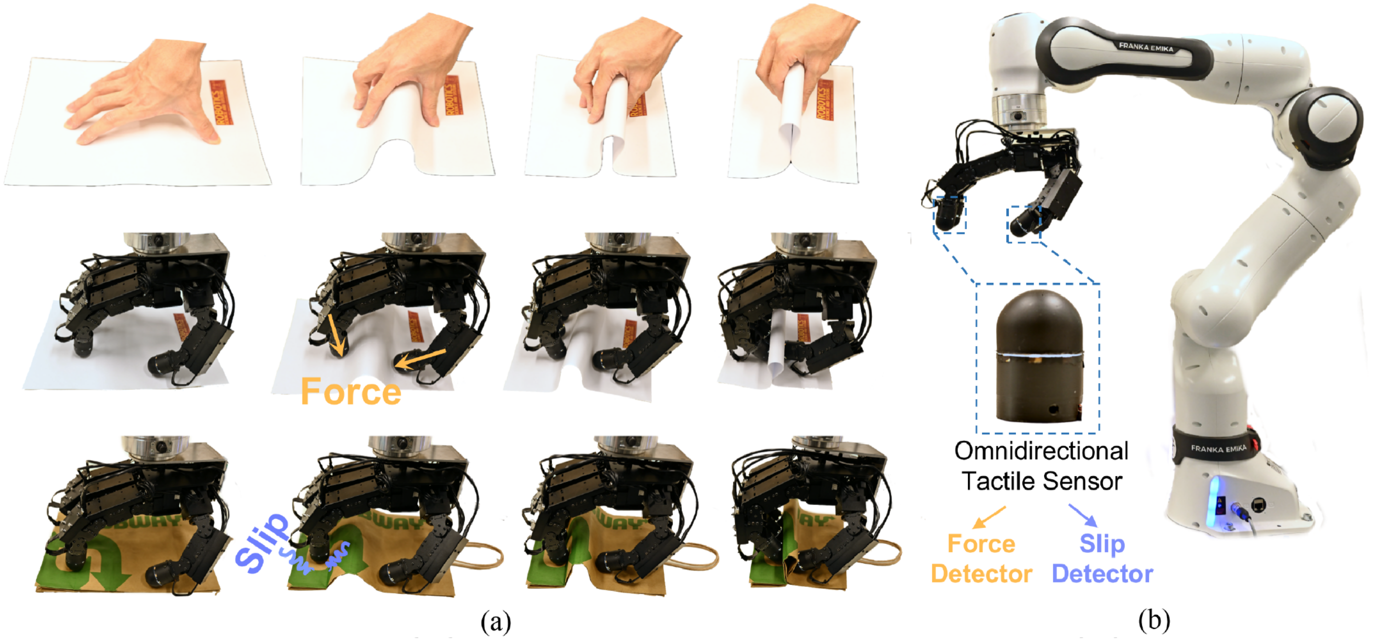

Fig. 1: **Overview of PP-Tac.** The system leverages tactile feedback from the proposed hemispherical sensor (R-Tac), integrated into a dexterous robotic hand, to grasp thin, deformable, paper-like objects. (a) Hand motions are generated by a diffusion-based policy inspired by human strategies, such 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. Despite recent advances in robotic hardware and embodied AI, existing systems continue to struggle with handling thin, flat, and deformable objects such as paper and fabric. These limitations stem from the lack of robust perception techniques for reliable state estimation under diverse visual conditions and the absence of planning methods capable of generating effective grasping motions. To address these limitations, we propose PP-Tac, a robotic system designed to pick up paper-like objects. PP-Tac incorporates a multi-fingered robotic hand equipped with high-resolution, hemispherical tactile sensors (R-Tac) that provide omnidirectional tactile feedback. This hardware configuration enables real-time slip detection and online force control to mitigate slippage during manipulation. Grasp motion generation is accomplished through a trajectory synthesis pipeline, which constructs a dataset of pinching motions and trains a diffusion-based policy to control the hand-finger simultaneously. Experiment results show that PP-Tac successfully grasps paper-like objects with varying material, thickness, and stiffness, achieving an overall success rate of 87.5%. To the best of our knowledge, this is the first system to successfully grasp paper-like deformable objects using a tactile dexterous hand.**

I. Introduction

Robots are increasingly popular as assistive agents in everyday life, particularly within household environments [38]. These robots are designed to perform various domestic tasks, often involving the grasp of thin, deformable objects such as paper and fabric [51]. For instance, clothes-folding tasks [27] require high dexterity and adaptability to accommodate variations in fabric size, texture, and stiffness, while document organization tasks [1] demand precise 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 [51]. 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 [28]; 2) These objects are often flat in shape, lacking salient features for contact points and thus hindering the synthesis of stable grasps [9]; 3) These objects exhibit high appearance variability due to continuous and unpredictable deformations during manipulation, which significantly hinders 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 the 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 consists of two key components: **1) A dexterous robotic hand equipped with hemispherical, high-resolution Vision-Based Tactile Sensors (VBTS) sensors (R-Tac):** Mounted on the fingertips, these tactile sensors provide real-time omnidirectional tactile feedback during grasping. With a hemispherical sensing surface and a high-frame-rate monochrome camera, the design offers faster response and simplified calibration compared to conventional RGB-based tactile sensors. An overview is shown in Fig. 1(b). **2) A diffusion-based motion generation policy (PP-Tac policy):** This policy imitates human picking strategies by first generating expert demonstrations through trajectory optimization that replicate sliding and pinching behaviors. It then trains a diffusion model on these trajectories, enabling the robotic hand to adaptively grasp diverse flat objects using proprioceptive and tactile feedback.

In a series of comprehensive real-world experiments, 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. Examples of successful grasps are shown in Fig. 1(a). The system also demonstrated strong adaptability to previously unseen uneven surfaces. An ablation study further confirmed the significance of each system component, highlighting the essential roles of tactile feedback and the motion generation policy in enabling coordinated hand–finger motion in the dexterous robotic hand.

To the best of our knowledge, this work presents the first demonstration of using a dexterous hand with tactile feedback to pick up thin, flat, deformable paper-like objects. The main contributions of this work are fourfold:

1) We present R-Tac, a novel hemispherical tactile sensor designed for ease of fabrication, calibration, and scalable deployment. R-Tac is integrated into each fingertip of a fully actuated dexterous robotic hand, providing real-time contact feedback during manipulation tasks.
2) We propose a novel trajectory optimization framework for data generation that avoids using computationally expensive tactile or soft-body simulations, while enabling robust sim-to-real transfer.
3) We present the PP-Tac policy, a diffusion-based control strategy that simultaneously generates coordinated hand and finger motions, relying solely on tactile and proprioceptive feedback to manipulate paper-like objects. The policy demonstrates robust generalization across a wide range of materials and surface conditions.
4) We provide a full implementation and systematic evaluation of the proposed algorithms on a physical robotic system. Both the hardware design and code for the PP-Tac system are publicly released to support further research and community development.

II. Related Work

A. Deformable Object Manipulation

Deformable Object Manipulation (DOM) tasks involve manipulating soft objects that deform during interaction, presenting long-standing challenges in robotic research. These difficulties primarily arise from uncertainties in perception and the complex dynamics of soft bodies [16, 3, 31]. Early approaches addressed these issues using vision-based perception for state estimation [51, 37], enabling tasks such as rope handling [33, 37], cloth folding [42, 27], and picking up paper with visual markers [14]. However, vision-based methods often struggle in real-world DOM tasks due to variability in object appearance, unknown physical properties, visual occlusions [25, 6], and inconsistent lighting conditions [48, 22]. These limitations hinder the scalability of vision-based DOM solutions in diverse and unstructured environments.

Tactile sensing, particularly Vision-Based Tactile Sensors (VBTS), has demonstrated significant potential for solving DOM tasks [51]. Leveraging their high-resolution tactile feedback, VBTS has demonstrated high performance in object shape reconstruction [34, 10, 30, 47], localization [20, 26, 7], and slip detection [44, 13]. Prior work has explored VBTS for deformable object manipulation [40], 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 adaptability for objects placed on non-flat surfaces, which highlights the need for integrating dexterous robotic hands with VBTS for robust manipulation [21].

B. Dexterous Robotic Hand with Tactile Sensing

Existing dexterous hands are often equipped with tactile sensors, which commonly utilize capacitive [32], piezoresistive [19], or magnetic-based [15] sensing mechanisms. These designs support a variety of form factors, enabling integration with different robotic fingers. However, their underlying sensing principles often limit spatial resolution and robustness under varying environmental conditions. To improve sensing quality, recent efforts have focused on developing VBTS,

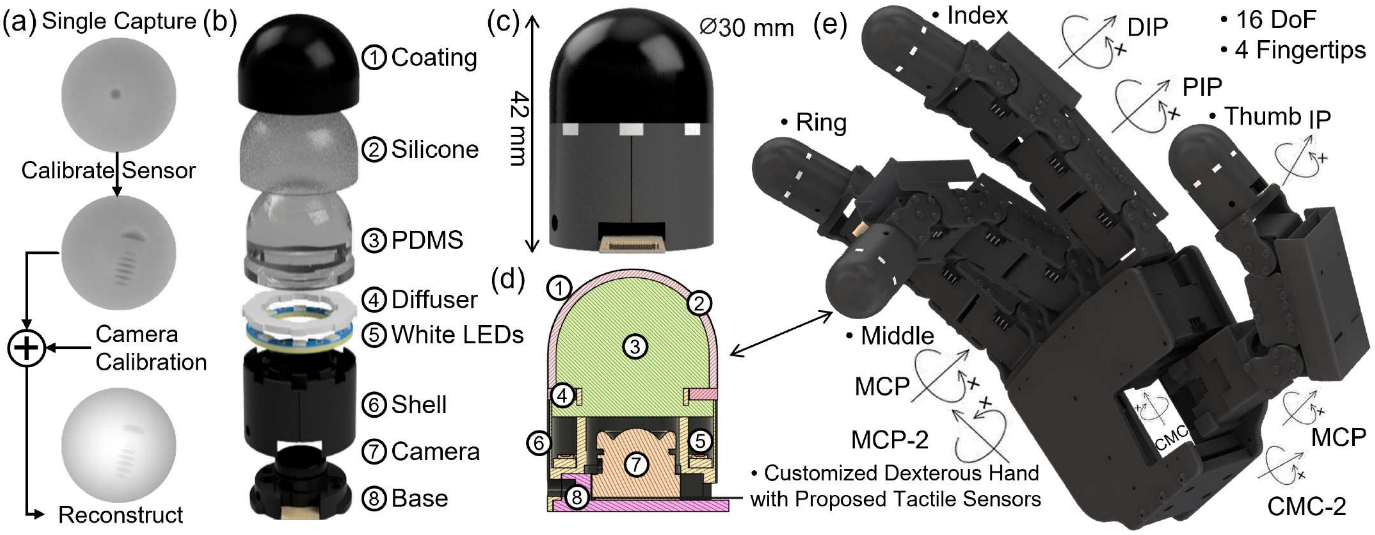

Fig. 2: **The hardware design of the R-Tac and its integration into the four-fingered dexterous robotic hand system.** (a) illustrates the pipeline of depth reconstruction. (b) illustrates the exploded view of the sensor, detailing each component. (c) shows the dimensions of the sensor. (d) shows the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

particularly those with curved elastomer surfaces [10, 46, 23, 2, 4]. Despite promising capabilities, most existing VBTS designs remain non-commercial and are difficult to deploy at scale, especially in dexterous hands. A major challenge lies in sensor calibration. Illumination from RGB chromatic light sources can create uneven light intensity distributions on curved elastomer surfaces, necessitating extensive data collection for accurate calibration. This process often depends on specialized test beds—such as those fabricated via CNC machining—further increasing complexity [10, 46, 23, 2]. Additionally, streaming real-time chromatic video data imposes high bandwidth demands, potentially limiting frame rates in multi-sensor deployments. To address these limitations, we propose R-Tac, a structurally simple, compact, and easily calibratable tactile sensor optimized for scalable deployment.

Dexterous robotic hands equipped with VBTS have been successfully applied to grasping and in-hand manipulation tasks. For example, Do *et al.* [11] utilize DenseTact [10] on an Allegro Hand [36] to grasp and manipulate small screws, while Qi *et al.* [35, 43] integrate fingertip VBTS to perform object reorientation. However, using VBTS-equipped dexterous robotic hands for picking up thin, flat, deformable paper-like objects, such as paper sheets, remains unexplored.

III. Hardware Design

To meet the dexterity requirements of paper-picking tasks, we designed and fabricated a set of hemispherical VBTS sensors, named R-Tac, which were custom-integrated into the Allegro Hand.

A. Fingertip-shaped Tactile Sensing

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

- **Round-shaped:** The hemispherical contact module enables omnidirectional tactile perception.
- **High resolution:** Supports accurate depth reconstruction and reliable slip detection during interaction.
- **Low-cost and easy fabrication:** Comprised of off-the-shelf or easily fabricated components, with a total cost of approximately $60.
- **Efficient calibration:** The monochrome sensing principle simplifies lighting control and minimizes manual calibration effort, making it well-suited for multi-fingered deployments.
- **Lightweight data transmission:** The monochrome camera produces less data volume per frame, enabling high-speed data transfer.

Following these principles, the sensor's design and integration into the dexterous hand are shown in Fig. 2. We now detail the sensor components 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 [29], 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

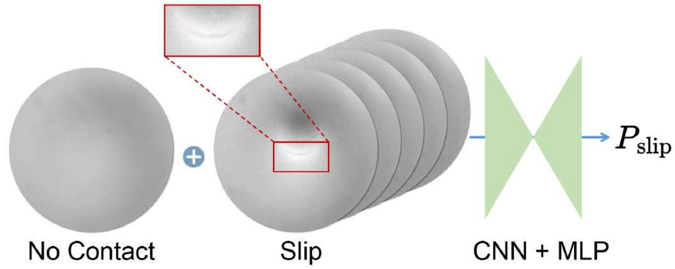

Fig. 3: **Slip detection.** The left tactile image represents a no-contact state, while the middle captures wrinkle features indicative of slip. The network estimates the slip probability P_{slip} by processing the no-contact reference image with the five most recent tactile frames.

Paint) is airbrushed onto the perception layer to form the opaque protective layer. The entire manufacturing process can be completed within three days.

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 [29]. 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 is modeled based on elasticity theory, where it is linearly proportional to the deformation depth. The slip detection module operates as follows:

- **Detection Model:** As shown in Fig. 3, slip events are characterized by the appearance of distinct wrinkle patterns in the tactile images. We employ a lightweight neural network composed of a convolutional neural network (CNN) followed by a multilayer perceptron (MLP) to detect these events. The network takes as input a temporal sequence of five tactile frames along with a reference no-contact image. The CNN extracts features from each frame, and the resulting feature maps are concatenated and passed through the MLP to estimate the slip probability P_{slip}.
- **Training:**The network is trained on approximately 20 minutes of tactile data collected from four sensors. The dataset includes 40% slip and 60% non-slip samples, with each frame manually annotated. Binary cross-entropy is used as the loss function.
- **Inference:** During inference, a threshold is applied to P_{slip} to determine slip events. Empirical evaluation shows that a threshold of 0.75 offers a good balance between sensitivity and precision, achieving a slip detection accuracy of 86%.

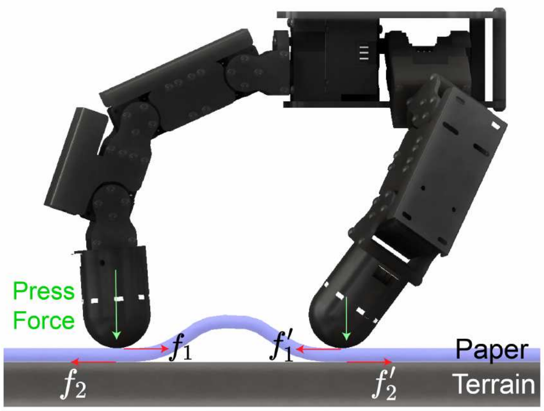

Fig. 4: **Force analysis during grasping flat objects.** The grasping process relies on three key forces: 1) The contact normal force exerted by the sensor on the object. 2) the static friction force (f_1, f_1') between fingers and the object, and 3) the dynamic friction force (f_2, f_2') between the object and the supporting terrain. A successful grasp occurs when the static friction (f_1, f_1') exceeds the critical buckling resistance of the paper, causing the sheet to deform and form a stable pinch region.

B. Robotic Hand System

We integrated the proposed R-Tac sensors into a fully actuated dexterous robotic hand, with each sensor mounted at the distal end of the fingertips to enable contact sensing for subsequent paper-picking tasks. The hand comprises 16 controllable DoFs, including the DIP, PIP, MCP, and MCP-2 joints for the index, middle, and ring fingers, as well as the CMC, CMC-2, MCP, and IP joints for the thumb. Actuation is provided by Dynamixel XC330-M288-T motors, multiplexed through a U2D2 Hub. Each tactile sensor communicates with the PC via a USB interface. The robotic hand is mounted on a Franka Emika Research 3 robotic arm, which interfaces with the PC through a high-speed Ethernet connection.

IV. 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 the 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.

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 terrains. 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 finger joints and the hand pose (*i.e.*, the end-effector pose of the manipulator). Intuitively, this resembles human grasping behavior: when picking up a sheet of paper from a flat surface, the hand must first elevate and then lower to establish stable contact. However, a hand-finger coupling issue arises: the motion of one finger requires a specific hand 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 via trajectory optimization in simulation, eliminating the need for complex teleoperation interfaces. Although reinforcement learning (RL) presents an alternative, it typically requires soft-body simulation to capture the dynamics of deformable objects and the behavior of VBTS elastomers, often involving additional real-to-sim procedures to ensure fidelity. In contrast, our method relies on rigid-body dynamics and demonstrates direct sim-to-real transfer, as validated through physical experiments. The grasping procedure begins by initiating contact between the fingertips and the object's surface (see Appendix B for implementation details). Upon contact, the fingers close gradually to pinch the object, each following an independently optimized trajectory while applying a target normal force (Figs. 4 and 5).

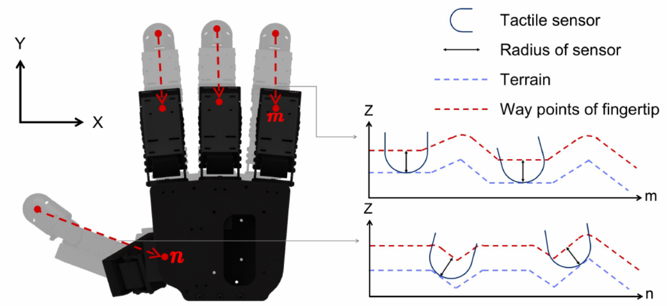

Fig. 5: **Fingertip trajectories from data synthesis.** Trajectories ensure fingertip sliding along the terrain surface. Adjusting the distance between waypoints and terrain affects sensor deformation. The right figure projects trajectories of two fingers onto the m-z and n-z planes, where m and n are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

To generate diverse fingertip trajectories, we first constructed randomized terrain profiles and manually designed initial pinching motions that emulate natural, intuitive grasping behaviors. As illustrated in Figs. 4 and 5, the (x, y) coordinates of the fingertip paths were extracted from these handcrafted sequences as target trajectories. The corresponding z-coordinates were obtained by projecting these points onto the terrain surface and sampling the local height at each location. This process yields the final target fingertip trajectories, denoted as $\boldsymbol{ee}_{target} \in \mathbb{R}^{4\times 3}$, for all four fingers.

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

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

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

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

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

where $\boldsymbol{\gamma}$ is the optimization variables consisting of N_{data} frames' finger joint angles $\boldsymbol{q}$, hand (*i.e.*, wrist, mounted to the end-effector of the arm) rotation $\boldsymbol{R}$ and hand 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 $L_{\triangle}$ regularizes the motion to remain close to the initial pose. Furthermore, $L_{\boldsymbol{R},\boldsymbol{p}_{\text{wrist}}}$ penalizes excessive wrist movement, helping to keep the arm within its reachable workspace. We use stochastic gradient descent (SGD) for optimization. After filtering out collision-prone sequences, we obtained a dataset of 500,000 grasp samples, each consisting of $N_{\text{data}} = 100$ 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 [17, 18, 8, 41], which predicts N_{pred} future states conditioned on N_{prefix} historical states. The state variables are written as:

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

where $\boldsymbol{p} \in \mathbb{R}^{17\times3}$ is hand joints' position in world coordinate, $\dot{\boldsymbol{p}} \in \mathbb{R}^{17\times3}$ 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 [50]) 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 an over-parameterized input allows the network to extract more robust and expressive latent features for the diffusion policy.

The overall pipeline is illustrated in Fig. 6, with the right figure depicting a single denoising diffusion step in detail. We apply an encoder-only transformer to predict future robot motion $\boldsymbol{x}_0^{\text{pred}}$ given prefix motion $\boldsymbol{x}^{\text{prefix}}$, diffused future motion $\boldsymbol{x}_t^{\text{pred}}$, 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 $\mathbb{R}^{(1+N_{\text{prefix}}+N_{\text{pred}})\times\mathcal{D}}$, 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) $N_{\text{prefix}} \times \mathcal{D}$ dimensions corresponding to the prefix states of N_{prefix} time steps. 3) $N_{\text{pred}} \times \mathcal{D}$ dimensions for the predicted states of N_{pred} time steps. Instead of predicting ϵ_t (formulated by [18]), we follow [45] to predict the state sequence itself $\hat{\boldsymbol{x}}_0^{\text{pred}}$. Predicting $\hat{\boldsymbol{x}}_0^{\text{pred}}$ 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 follows:

$$L = \|\hat{\boldsymbol{x}}_0^{\text{pred}} - \boldsymbol{x}_0^{\text{pred}}\|_2^2 + \lambda_{consist} L_{consist}, \quad (6)$$

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

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 $\boldsymbol{x}_{1000}^{\text{pred}} \sim \mathcal{N}(0, I)$ and iteratively denoise it to produce $\boldsymbol{x}_0^{\text{pred}}$. To ensure real-time performance, we reduce denoising steps to 10 and set $N_{\text{pred}} = N_{\text{prefix}} = 5$, 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 $\boldsymbol{x}^{\text{prefix}}$ during training:

- Add random Gaussian noise to γ 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 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).

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 $\boldsymbol{x}_0^{\text{pred}}$. Subsequently, noise is added to the $t = 1000 - 100N_i$ level and denoised again to $\boldsymbol{x}_0^{\text{pred}}$, 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.

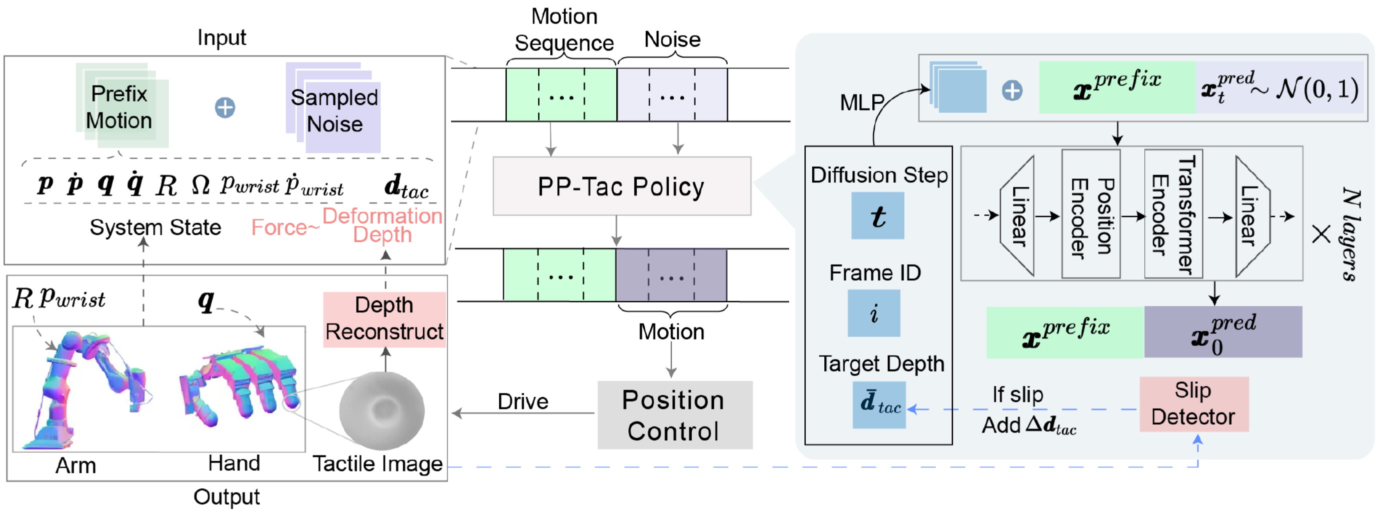

Fig. 6: **Inference pipeline of the proposed PP-Tac policy.** Conditioned on robot proprioception and the target force that needs to be exerted, PP-Tac can infer the action of the next steps. If a slip is detected between the finger and the flat object underneath, an incremental amount of force will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

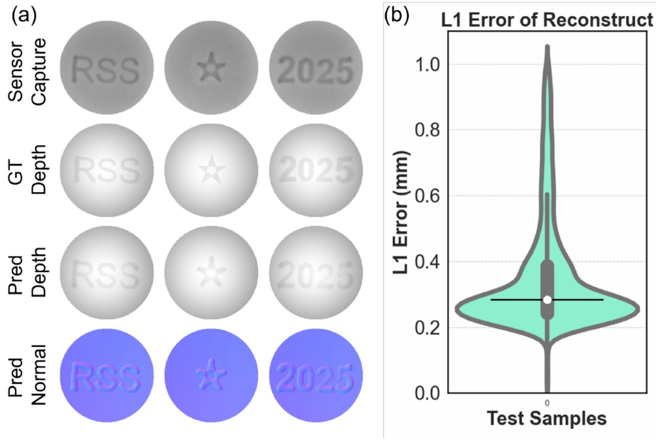

Fig. 7: **Reconstruction results.** (a) Gallery of reconstructed depth and normal maps from tactile images. (b) Depth reconstruction error of the indentation test.

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 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

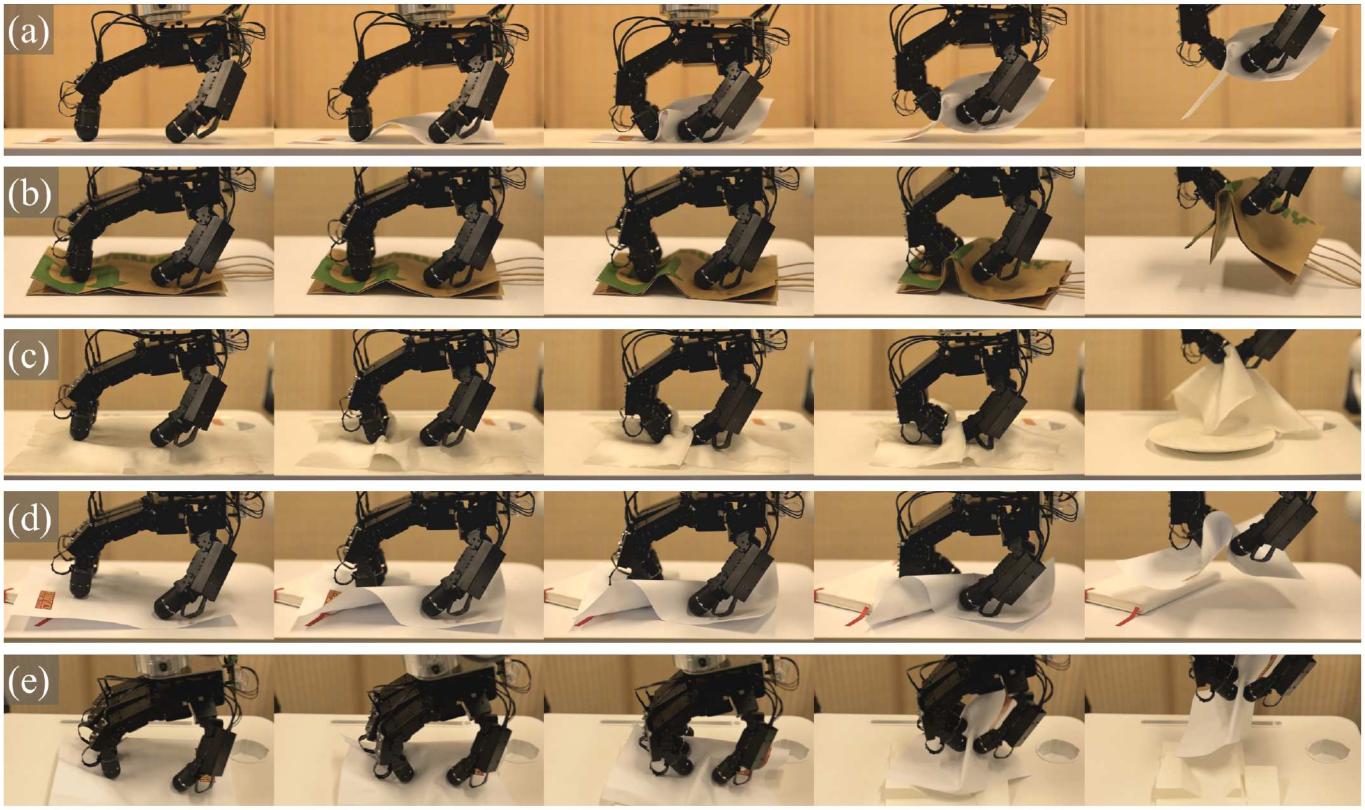

Fig. 8: **Gallery of Grasping Different Objects in Real-World Evaluations.** This figure demonstrates successful grasps of five flat objects on four different types of terrains, highlighting the effectiveness of our hardware and the PP-Tac algorithm. (a) A paper on a flat desktop. (b) A stiff Kraft paper 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.

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 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

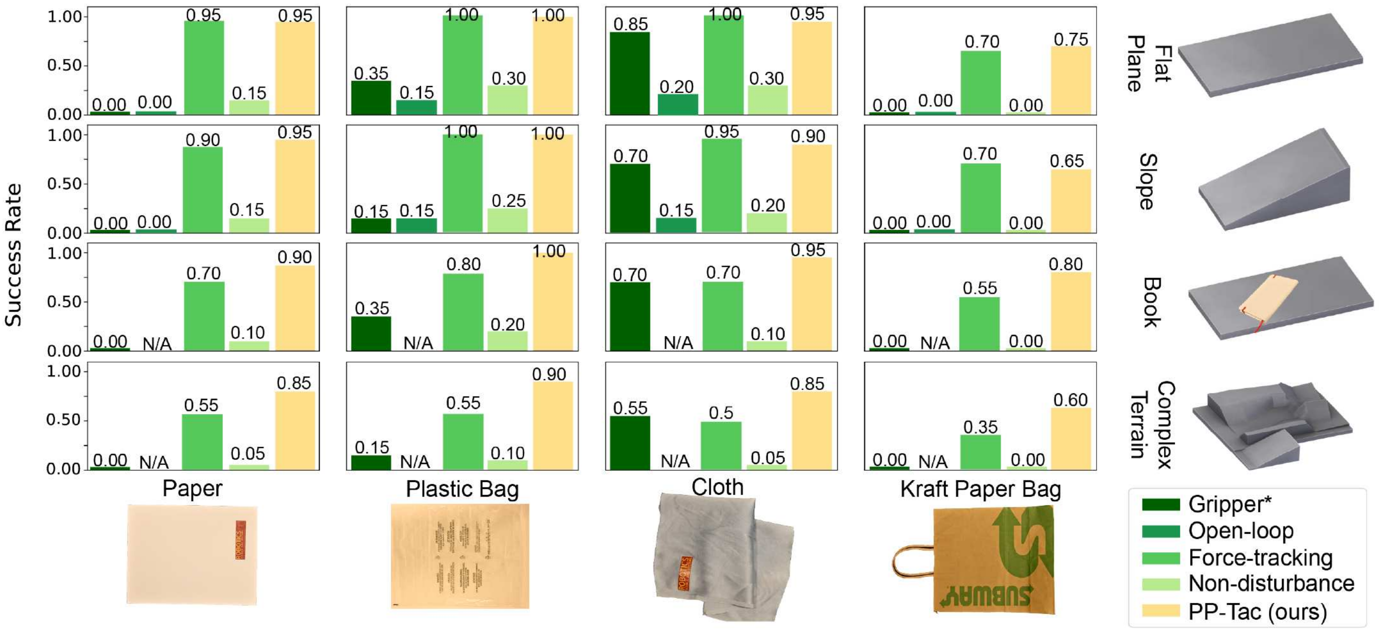

Fig. 9: **Experiment results.** Evaluations were conducted to quantify the success rate of grasping four different flat objects (paper, plastic bag, cloth, and paper bag) across four terrain setups (plane, slope, book placement, and randomly generated complex terrain). Baseline conditions included: (1) 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; these scenarios are marked as ‘N/A’.

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 [39], Allegro Hand 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 $\boldsymbol{x}^{\text{prefix}}$. 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] Christof Elbrechter, Robert Haschke, and Helge Ritter. Bi-manual robotic paper manipulation based on real-time marker tracking and physical modelling. In *International Conference on Intelligent Robots and Systems (IROS)*, pages 1427–1432. IEEE, 2011.

[15] 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.

[16] 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.

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:6840–6851, 2020.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] Pei Lin. Handdiffuse: generative controllers for two-hand interactions via diffusion models. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, pages 5280–5288, 2025.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

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

[37] 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.

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

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

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

[50] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5745–5753, 2019.

[51] 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 [49]. 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 [24] to calculate the extrinsic parameters that includes rotation matrix A and translation vector b such that:

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

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

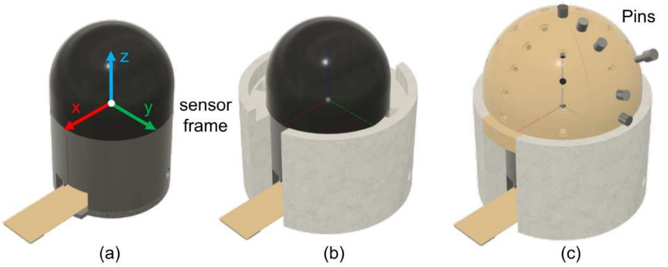

Fig. 10: **Camera calibration using an indentation setup:** The sensor frame is first defined in (a). A holder is designed and 3D-printed to secure the sensor, as shown in (b). A hemispherical indicator is designed and 3D-printed to attach to the sensor holder. Pins are inserted into pre-defined holes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

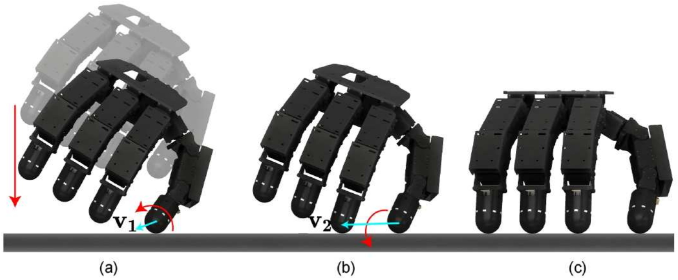

Fig. 11: **Example of establishing contact:** First, the hand descends until a finger makes contact with the surface. A fixed-point rotation is performed around the contacting finger, as shown in (a). The hand then continues to rotate until a second finger makes contact, triggering a fixed-axis rotation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

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 $\boldsymbol{C} = \{f_t\}, \boldsymbol{N} = \{f_i, f_m, f_r\}$

*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:

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

where $R_z(90^\circ)$ 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.