# CRAFT: Contact-Rich Affordance via Mechanoreceptor-inspired Framework for Tactile Manipulation

**Hojin Jung (2025313307)**[1] **and Minjae Kim (2025321692)**[1]
[1]HAN Lab

## 1. Introduction

Contact-rich manipulation means that a robot performs tasks by taking into account physical interactions with the environment—such as continuous contact, friction, and collisions—while manipulating objects. As environments become more complex and tasks more difficult, imitation learning is being actively researched as a solution to these challenges.

Imitation learning is a learning-based method for transferring the ability of a human to perform complex and difficult tasks proficiently to a robot [11, 13]. A demonstration dataset is constructed from an expert, and the robot's policy is trained to generate actions identical to those performed by the expert for a given observation. Action generation has evolved from action chunking-based policies [15] and diffusion-based policies [1] to, more recently, flow matching-based policies that learn the action distribution as a velocity field to generate precise, multimodal actions with fewer inference steps [2, 7, 14]. These imitation learning policies have proven their effectiveness in various contact-rich tasks, such as battery insertion, cable routing, and peg-in-hole assembly [1, 8].

It is important to fully incorporate the sensory information that a person uses while performing a task when constructing a dataset with the help of experts. Specifically, all factors that influence judgment and decision-making during task execution—such as the scene as seen from the expert's perspective, the manner of movement, and the amount of force applied—must be identified and quantified. Consequently, approaches that go beyond relying solely on the robot's motor and vision data to also incorporate contact tactile information have emerged [4, 5]. This enables the robot to perform tasks—such as precise insertion, slip compensation, and adjusting the grip force for fragile objects—that would be difficult to achieve using vision alone [12]. However, when performing a task, an expert does not rely on a single tactile signal but utilizes all four types of tactile information detected by mechanoreceptors.

For a robot to perform dexterous manipulation while assessing situations and making decisions like a human, it must be able to detect tactile information as richly as a human does. In humans, tactile sensations at the fingertips are detected by four types of mechanoreceptors [6]. The slowly adapting (SA-I) Merkel's disk and Ruffini endings (SA-II) detect sustained pressure and shape, respectively, as well as skin stretch, while the rapidly adapting (RA1) Meissner corpuscles and (RA2) Pacinian corpuscles detect low- and high-frequency vibrations, respectively, capturing contact and slip. These four receptors are functionally classified into the slowly adapting (SA) series, which responds to sustained stimulation, and the rapidly adapting (RA) series, which responds to vibrations. Through this complementary integration of spatial pressure information and temporal vibration information, humans can precisely perceive the progress of a task even in contact situations where vision is limited. This advantage is particularly evident in tasks that rely more heavily on touch than on vision—such as loosening or tightening screws with a screwdriver—where vibration signals at the fingertips serve as key cues for determining the success or failure of the task.

In this project, we propose CRAFT, a mechanoreceptor-inspired hierarchical tactile perception framework for contact-rich bilateral manipulation, drawing inspiration from the structure of human tactile perception. The proposed framework is validated through an unscrewing task using a driver. The main contributions of this paper are as follows:

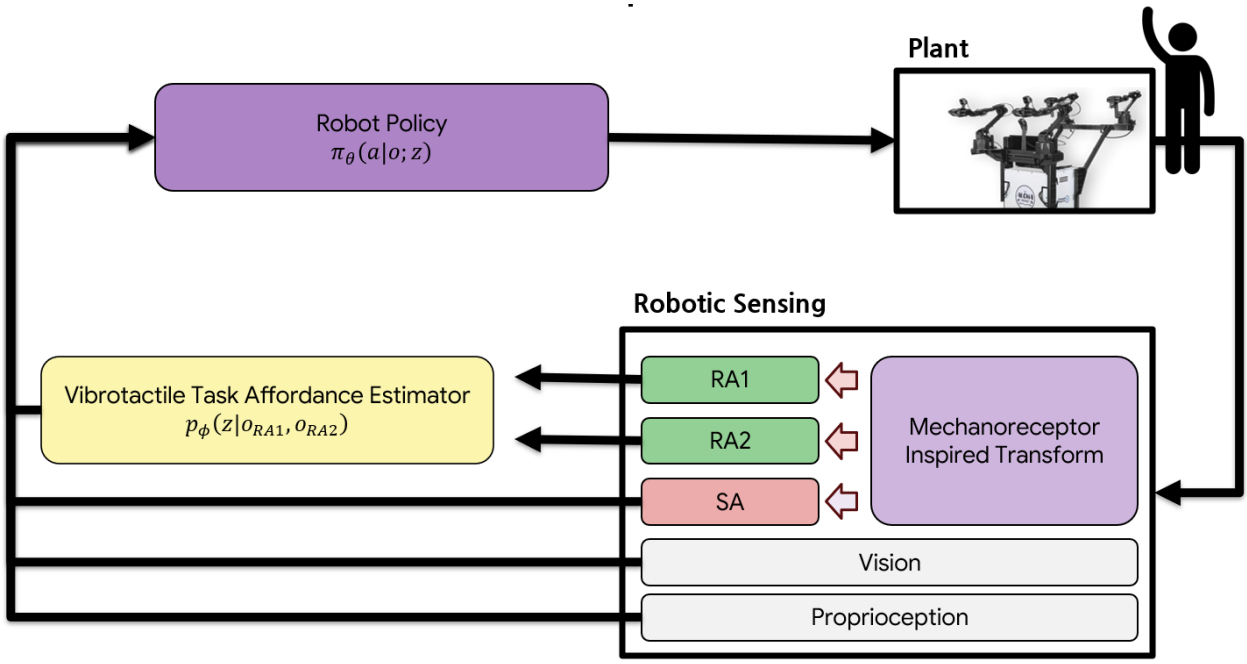



**Figure 1:** Proposed dual-arm robot framework

- We propose a mechanoreceptor-inspired tactile representation that decomposes magnetometer-based tactile signals into SA (KDE-based spatial encoding) and RA1 and RA2 ( CWT-based time-frequency encoding).
- We propose a vibrotactile task affordance estimator that estimates task affordance in real time based on RA1 and RA2 vibrations, and use it to construct a closed-loop structure that switches to a recovery policy when a failure is detected.
- We validate the framework by performing a unscrew task on actual dual-arm robot and demonstrate, through estimator ablation, time-frequency representations achieve higher state estimation performance (particularly in failure detection) than time-series representations.

## 2. Method

### 2.1. The Proposed Dual-Arm Robot Framework

Figure 1 shows the proposed robot framework. The robot uses a dual-arm robot with each arm having 7-DoF. Each arm is equipped with either a visual sensor or a tactile sensor, and at every time step, data from the robot's motors, vision, and tactile sensors are observed as *robotic sensing*. Among these, the vibrotactile data (RA1, RA2) are used as inputs to the vibrotactile task affordance estimator $p_\phi(z \mid o_{RA1}, o_{RA2})$ to estimate the task affordance $z$ of the currently performed task. The robot policy $\pi_\theta(a \mid o; z)$ generates an action based on the estimated $z$ and the robotic sensing observations, and the generated action is applied to the plant, forming a closed-loop system.

### 2.2. Robotic Sensing

Each arm is equipped with two types of sensors. The first is an RGB camera which corresponds to the human eye. RGB images $I_R, I_L, I_H \in \mathbb{R}^{3\times H\times W}$ are acquired from the wrist camera on each arm and a global camera that surveys the entire workspace. Second are tactile sensors which correspond to human skin; magnetometer-based tactile sensors are mounted on the left and right fingers of the left arm's gripper, respectively. Magnetometers return 3-axis magnetic field data. In this project, we use the magnitude of these sensors. Each tactile sensor contains five magnetometers, so at every time step, tactile time-series data $M \in \mathbb{R}^{2\times 5}$ is observed (2: left and right sensors; 5: number of magnetometers per sensor). Therefore, the observation $o_t$ fed back to the dual-arm robot at each time step is as follows:

$$o_t = \{I_{R,t}, I_{L,t}, I_{H,t}, M_t, \mathbf{q}_t, \dot{\mathbf{q}}_t\}. \tag{1}$$

Where $\mathbf{q}$ and $\dot{\mathbf{q}} \in \mathbb{R}^{14}$ are the joint angles and angular velocities of the robot, respectively.

### 2.3. Mechanoreceptor Inspired Transform

The tactile sensor used is the customizable cut-cell microstructured magnetic touch sensor proposed in [10]. A magnet is embedded within the microstructure and each sensor circuit is equipped with five magnetometers. When the microstructure is pressed by external contact, the magnetic field changes, and this change is used to represent SA (compression/extension) and vibration (RA1, RA2) information.

*SA representation.* The five magnetometers arranged on a single sensor measure local deformations caused by contact at different spatial locations. To represent these sparse measurements as a continuous two-dimensional pressure distribution, we apply a kernel density estimation-based spatial encoding method—which places a Gaussian kernel at each magnetometer location and synthesizes the signal by weighting it with the measured values—to generate an SA spatial pressure image.

*RA representation.* Humans detect vibrations through the Meissner corpuscles (RA1) and Pacinian corpuscles (RA2) in their mechanoreceptors; the detection bands of these two types of corpuscles are 5–150 Hz and 20–1,000 Hz, respectively [9]. To represent the time-series data acquired from the magnetometer in accordance with the detection bands of each body, the continuous wavelet transform (CWT) is applied to generate a time-frequency scalogram image. However, within the sensor's effective bandwidth, the CWT is performed in the range of 5–150 Hz for RA1 and 20–200 Hz for RA2. Through this process, vibration information corresponding to RA1 and RA2 is obtained from a single tactile sensor in the form of scalogram images.

### 2.4. Vibrotactile Task Affordance Estimator

Screwdriving tasks using a screwdriver involve dexterous contact-rich manipulation that relies more on tactile feedback than on vision; the progress of the task and whether it succeeds or fails can be determined based on the vibration of the screwdriver detected at the fingertips. Therefore, we train an estimator $p_\phi(z \mid o_{RA1}, o_{RA2})$ that estimates the task affordance $z$ from the RA1 and RA2 time-frequency representations $o_{RA1}$ and $o_{RA2}$ with a window size of $k$ where $z \in \{\textit{no-contact}, \textit{success}, \textit{failure}\}$ is a categorical variable representing the current contact state. The estimator consists of a CNN that takes four consecutive scalogram frames stacked along the channel dimension as input and outputs probabilities for the three classes at each time step. To prevent unstable mode transitions caused by momentary misclassifications, an exponential moving average (EMA) is applied to the predicted probabilities. A duration filter is also implemented to finalize the state only when the same class is maintained for at least a minimum duration. The estimated $z$ is used by the robot policy to recognize task progress and switch to recovery actions in the event of failure (Sec. 2.5).

### 2.5. Imitation Learning based Robot Policy

While performing a task, the robot agent observes robot vision, SA tactile, RA1/RA2 tactile, motor angle, and motor angular velocity (Sec. 2.2). Additionally, the task affordance $z$, defined in Sec. 2.4, is given. Figure 2 shows the proposed model architecture which integrates and utilizes visual-tactile multimodal data. The backbone utilizes action chunking with transformer (ACT), proposed by [15]. ACT is a transformer-based imitation learning approach that generates the next action using a temporal ensemble when robot joint and corresponding scene are provided.

Unlike existing ACT methods, the key feature of Cutaneous Action Chunk Transformer (CACT) is its active utilization of mechanoreceptor-inspired cutaneous information. The operation sequence of CACT is as follows:

1. The robot observes sensory information: visual, tactile, and motor joint.
2. Tactile data is preprocessed according to Sec. 2.3 to obtain image-based tactile representations (SA, RA1, RA2).
3. Individual ResNet feature extraction yields feature vectors for vision and SA, respectively; these are concatenated with motor angle and angular velocity to form embedded sensory vectors.

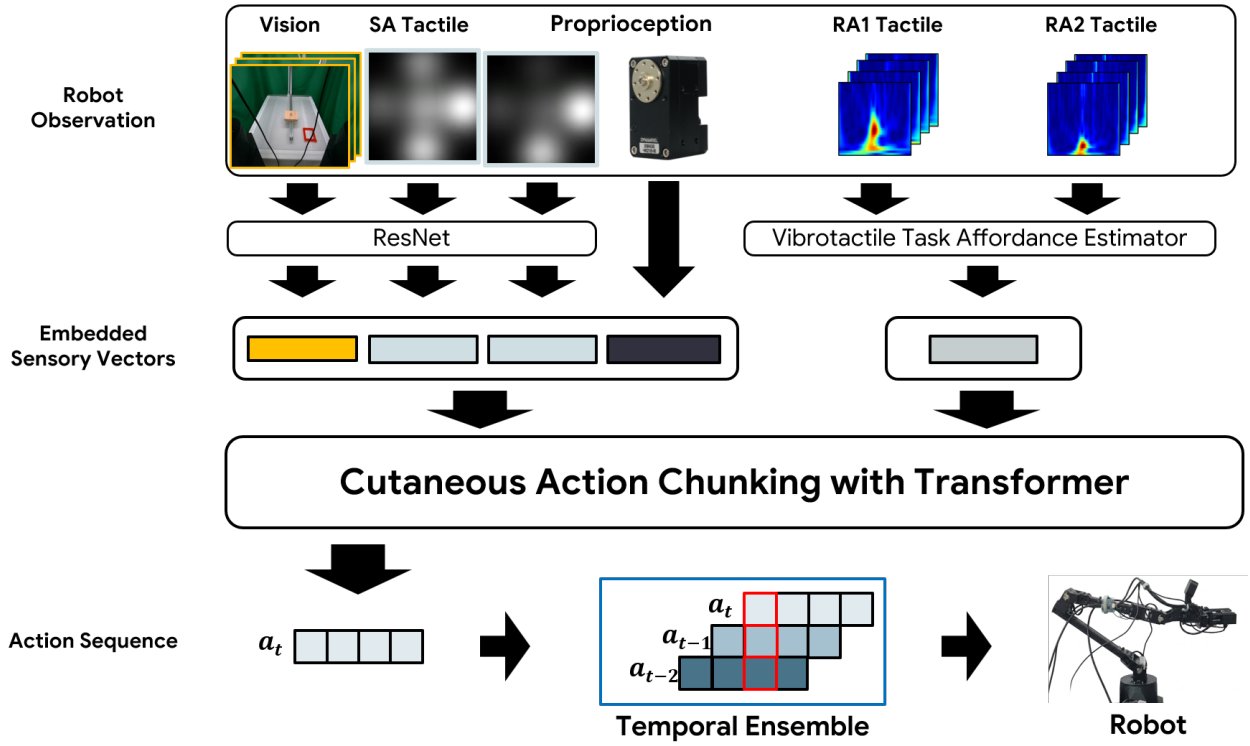



**Figure 2:** Imitation Learning based Robot Agent Figure

4. RA1 and RA2 are used to estimate the task affordance $z$ (Sec. 2.4).
5. CACT has a hierarchical structure that switches operating modes based on $z$. While $z$ is *no-contact* or *success*, the CACT policy generates action sequences using the embedded sensory vectors as input.
6. Conversely, once $z$ is confirmed to be *failure*, the system switches to the recovery policy and plays back a pre-recorded trajectory that raises the arm holding the driver to a safe position. In this state, the human resets the driver's torque after which the robot returns to the CACT policy and continues performing the task.

## 3. Experiments

### 3.1. Experiments Setup

To validate the proposed visual-tactile multimodal sensing robot platform, we perform the task of unscrewing, which is a form of dexterous, contact-rich manipulation. The goal of the task is for one arm to unscrew a screw using an electric screwdriver, while the other arm moves the removed screw to a designated location. If the robot fails to unscrew the screw due to an inappropriate torque setting on the screwdriver, it detects the task failure via a vibrotactile task affordance estimator. In this case, following a recovery policy, the robot raises the arm holding the screwdriver to transition to a safe state; due to hardware limitations, a human then resets the screwdriver's torque in this state. The robot subsequently resumes the unscrewing task.

The dual-arm robot platform utilizes the master-slave architecture proposed in [3]. The movements of the master, controlled by a human via teleoperation, are replicated by the slave. The left slave robot is equipped with both visual and tactile sensors; the tactile sensors are located on the left and right grippers of the end-effector. The left slave robot is the arm that holds the screwdriver and interacts directly with the screw; to ensure safe experimentation, the screwdriver is securely fastened to a jig to prevent it from slipping out. In contrast, the right slave robot is equipped with only visual sensors. Additionally, a global visual sensor is positioned centrally to provide a view of the entire workspace.

**Table 1:** Confusion matrices of the task-state estimator (held-out validation, episode-wise 8:2 split). Rows are ground truth, columns are predictions; diagonal (correct) entries are bold. (a) seq-MLP (2.1 M params), overall accuracy 74.4%. (b) CWT+CNN (245 K params), overall accuracy 94.9%.

| True | Predicted | | |
|---|---|---|---|
| | None | Success | Failure |
| None | **110** | 12 | 21 |
| Success | 3 | **127** | 7 |
| Failure | 27 | 55 | **126** |

(a) seq-MLP

| True | Predicted | | |
|---|---|---|---|
| | None | Success | Failure |
| None | **134** | 5 | 4 |
| Success | 7 | **129** | 1 |
| Failure | 8 | 0 | **200** |

(b) CWT+CNN

### 3.2. Data Collection

One demonstrator (male, age 30) and one operator (male, age 27) participated in the data acquisition for the demonstration. Prior to the actual data acquisition, the demonstrator was clearly briefed on the experimental protocol and was given a two-week practice period to become fully proficient in operating the robot and navigating the sensor feedback environment.

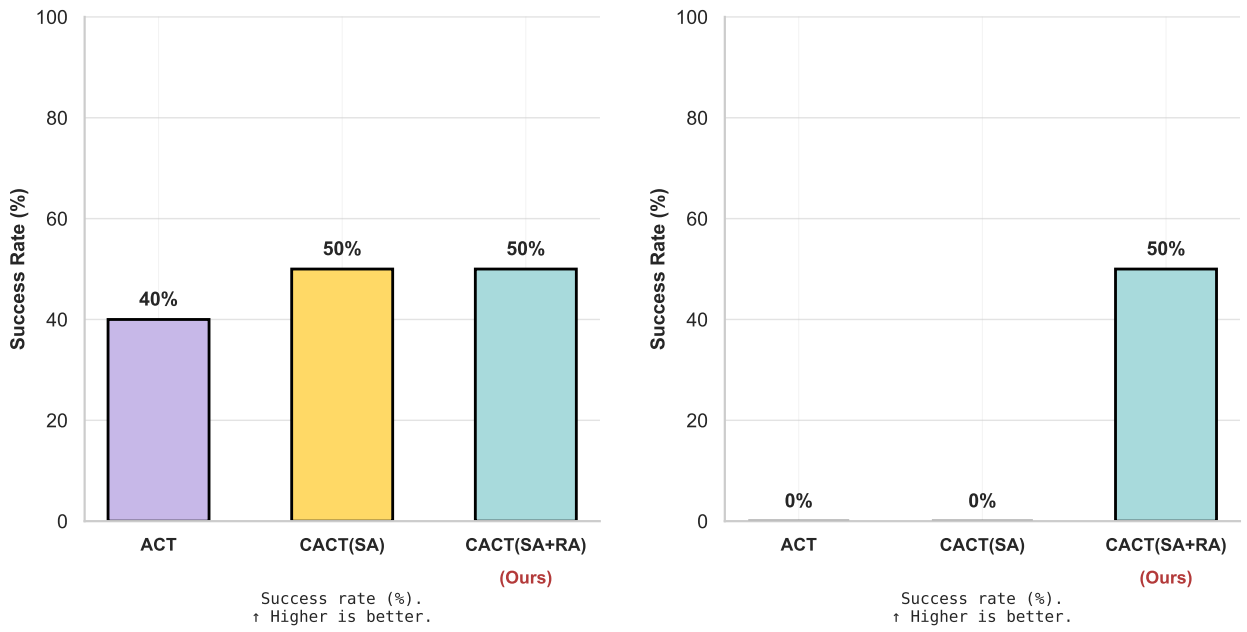
The data acquisition protocol was as follows. First, the demonstrator positioned himself in front of the dual-arm robot, while the operator checked the PC and the robot's status. The operator then sequentially activated the unscrewing task environment, the sensors mounted on the robot, and the robot manipulator. Once preparations were complete, the demonstrator performed the task of unscrewing a screw via teleoperation. During the experiment, the screwdriver's torque was set to two levels depending on whether the screw could be unscrewed. The success condition occurs when sufficient torque allows the screw to be unscrewed normally, while the failure condition occurs when insufficient torque prevents the screw from being unscrewed, requiring recovery. For each torque condition, a dataset was constructed consisting of 100 demonstration episodes, each lasting at least 10 seconds, for a total of 200 episodes.

## 4. Result and Discussion

### 4.1. Vibrotactile Task Affordance Estimator

Table 1 presents the results of an analysis of the performance of the proposed vibrotactile task affordance estimator. The estimator was trained on a total of 3,000 data pairs (approximately 1,000 per label) and validated on 488 held-out data pairs. Due to the episode-based splitting, the class distribution in the validation set is not uniform (none/success/failure = 143/137/208). We compared the two models based on the format of the vibration and haptic data provided as input to the estimators. The first model (seq-MLP) is an MLP that takes as input time series data obtained by bandpass filtering the sensor time series into the RA1 and RA2 bands, while the second model (CWT+CNN) is a CNN that takes as input scalogram images obtained by transforming the same time series using CWT.

As shown in Table 1, CWT+CNN achieved an overall accuracy of 94.9%, which is 20.5%p higher than the 74.4% achieved by seq-MLP. The gap between the two models is greatest in the *failure* class, which is the most critical for safety. The recall of seq-MLP for the *failure* class is only 60.6% (126/208), meaning it misses approximately 40% of failures; in particular, there were 55 cases where a failure was misclassified as a success, which was the most common error. In other words, bandpass time-series representations alone cannot sufficiently distinguish the vibration patterns during failure from those during success. In contrast, CWT+CNN achieved a failure recall of 96.2% (200/208) and a precision of 97.6% (200/205), accurately detecting nearly all failures. This is because the time-frequency representation reveals spectral signatures unique to failures—such as slippage and missteps—allowing the CNN to learn these as distinguishable patterns. From a closed-loop recovery perspective, the recall of failure detection directly determines whether recovery is triggered; therefore, the high recall of CWT+CNN is crucial to the reliability of the proposed framework.




**(a)** Sufficient torque

**(b)** Insufficient torque

**Figure 3:** Task success rate (↑ higher is better) under (a) sufficient and (b) insufficient driver torque.

Examples of multimodal observations over time, RA1 and RA2 inputs, and qualitative rollouts of estimated task affordance are presented in Appendix A (Figure 4).

### 4.2. Performance Comparative Analysis for the Trained Agent

Fig. 3 compares the task success ratios of the trained agent under conditions of sufficient and insufficient torque, respectively. Three methods are compared: the first is ACT which uses only vision and proprioception information; the second is CACT(SA) which adds SA tactile feedback to ACT; and the third is CACT(SA+RA) which incorporates an RA-based vibrotactile task affordance estimator and closed-loop recovery. Each method was evaluated through 10 rollouts.

Fig. 3 shows that, on the left, the success rates for ACT, CACT(SA), and CACT(SA+RA) are similar at 40% and 50%, respectively. This indicates that, if the environment provided to the robot is ideal (i.e., if the torque is set appropriately), the robot's policy successfully performs the task. In contrast, in the case shown on the right side of Fig. 3, where torque is insufficient, the results diverge significantly. ACT and CACT(SA), which lack failure detection and recovery capabilities, terminate the task without loosening the screw, resulting in a 0% success rate. In contrast, CACT(SA+RA) detects failure via vibration haptics and completes the task by following a recovery procedure (raising the arm and resetting the human torque), achieving a 50% success rate. This demonstrates that RA-based estimators and closed-loop recovery are crucial for handling failure situations that cannot be resolved by simple tactile augmentation alone.

## 5. Conclusion

In this project, we propose CRAFT, a contact-rich dual-arm manipulation hierarchical tactile perception framework inspired by the structure of human mechanoreceptors. We decomposed the signals from magnetic tactile sensors into SA (KDE-based spatial encoding) and RA1 · RA2 (CWT-based time-frequency encoding), and constructed a closed-loop structure that switches to recovery mode upon failure detection via a vibrotactile task affordance estimator, which estimates the task state in real time based on RA vibrotactile feedback. In a unscrew task using an electric screwdriver, the time-frequency representation-based estimator achieved an accuracy 20.5 % higher than that of the time-series representation and a failure reproduction rate of 96.2%. Furthermore, under conditions where insufficient torque required recovery, the proposed CACT(SA+RA) was the only method to achieve a 50% success rate in situations where both the vision-based ACT and the

CACT(SA)—which uses only SA haptic feedback—failed (0%), thereby demonstrating the utility of the RA-based estimator and closed-loop recovery.

Future research directions are as follows. First, since the current recovery procedure includes a step where a human manually resets the driver torque due to hardware constraints, efforts should be made to make this process fully autonomous. Second, the framework can be extended beyond the single unscrew task to various contact-rich tasks, and absolute success rates and generalization performance can be improved through additional demonstration data and evaluation rollouts.

## A. Qualitative Rollout of the Task Affordance Estimator

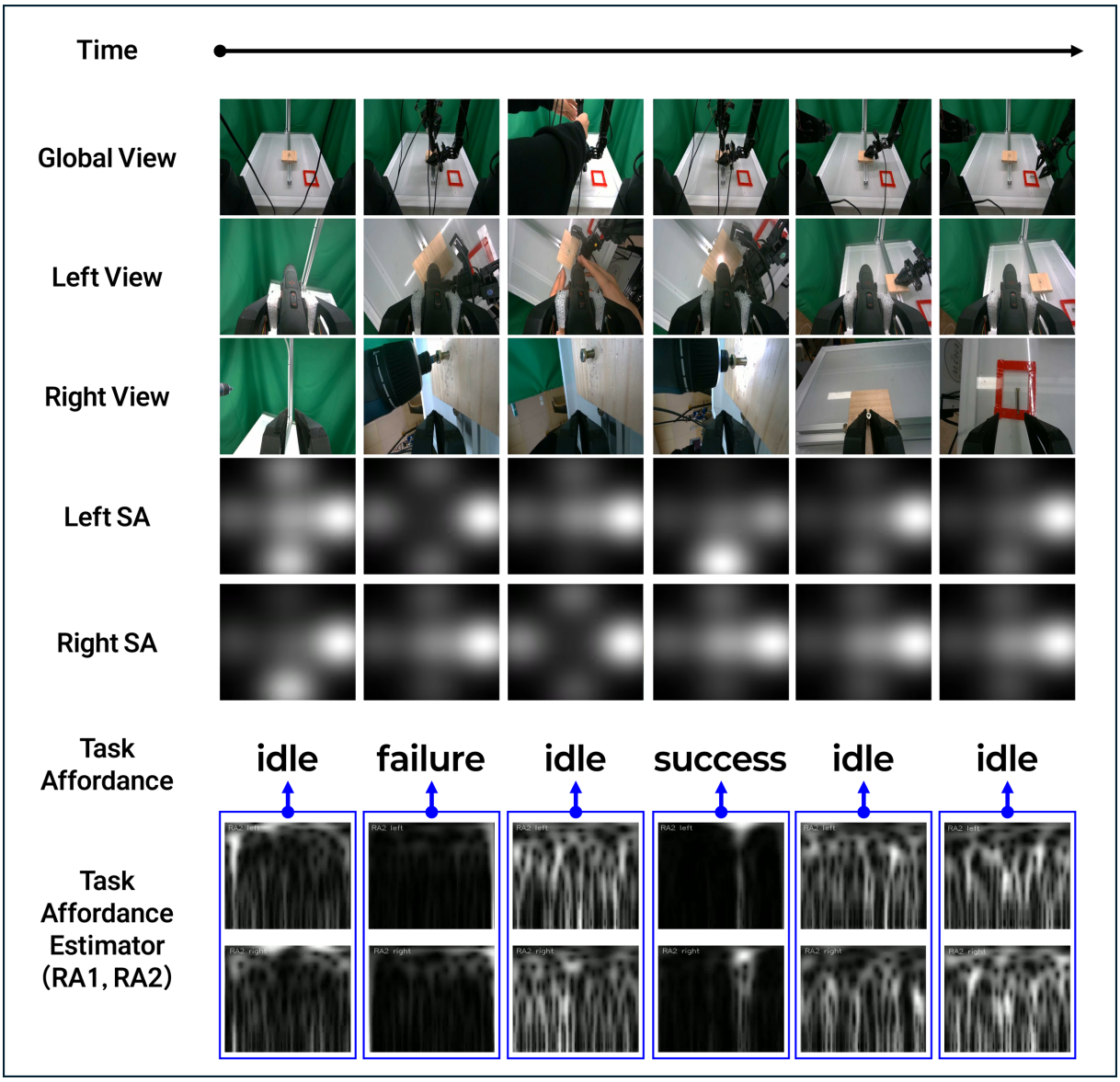



**Figure 4:** Qualitative rollout of the unscrewing task over time. Top to bottom: global, left-, and right-arm camera views; left and right SA (spatial pressure) tactile maps; the task affordance estimated by the vibrotactile estimator (idle / failure / success); and the RA1/RA2 time-frequency inputs (left and right grippers) fed to the estimator. The estimator infers the contact state directly from the RA scalograms.

## References

[1] 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. *The International Journal of Robotics Research*, 44(10–11):1684–1704, 2025. 1

[2] Eugenio Chisari, Nick Heppert, Max Argus, Tim Welschehold, Thomas Brox, and Abhinav Valada. Learning robotic manipulation policies from point clouds with conditional flow matching. *arXiv preprint arXiv:2409.07343*, 2024. 1

[3] Zipeng Fu, Tony Z Zhao, and Chelsea Finn. Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation. *arXiv preprint arXiv:2401.02117*, 2024. 4

[4] Ningquan Gu, Kazuhiro Kosuge, and Mitsuhiro Hayashibe. Tactilealoha: Learning bimanual manipulation with tactile sensing. *IEEE Robotics and Automation Letters*, 10(8):8348–8355, 2025. doi: 10.1109/LRA.2025.3585396. 1

[5] Binghao Huang, Yixuan Wang, Xinyi Yang, Yiyue Luo, and Yunzhu Li. 3d-vitac: Learning fine-grained manipulation with visuo-tactile sensing. In *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pages 2557–2578. PMLR, 2025. URL https://proceedings.mlr.press/v270/huang25e.html. 1

[6] Roland S. Johansson and J. Randall Flanagan. Coding and use of tactile signals from the fingertips in object manipulation tasks. *Nature Reviews Neuroscience*, 10(5):345–359, 2009. 1

[7] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 1

[8] Jianlan Luo, Charles Xu, Xinyang Geng, Gilbert Feng, Kuan Fang, Liam Tan, Stefan Schaal, and Sergey Levine. Multistage cable routing through hierarchical imitation learning. *IEEE Transactions on Robotics*, 40:1476–1491, 2024. 1

[9] Tamás Oroszi, Marieke J. G. van Heuvelen, Csaba Nyakas, and Eddy A. van der Zee. Vibration detection: Its function and recent advances in medical applications. *F1000Research*, 9:F1000–Faculty, 2020. 3

[10] Venkatesh Pattabiraman, Zizhou Huang, Daniele Panozzo, Denis Zorin, Lerrel Pinto, and Raunaq Bhirangi. eflesh: Highly customizable magnetic touch sensing using cut-cell microstructures, 2025. URL https://arxiv.org/abs/2506.09994. 3

[11] Edgar Welte and Rania Rayyes. Interactive imitation learning for dexterous robotic manipulation: Challenges and perspectives—a survey. *Frontiers in Robotics and AI*, 12:1682437, 2025. 1

[12] Han Xue, Jieji Ren, Wendi Chen, Gu Zhang, Yuan Fang, Guoying Gu, Huazhe Xu, and Cewu Lu. Reactive diffusion policy: Slow-fast visual-tactile policy learning for contact-rich manipulation. *arXiv preprint arXiv:2503.02881*, 2025. 1

[13] Maryam Zare, Parham M. Kebria, Abbas Khosravi, and Saeid Nahavandi. A survey of imitation learning: Algorithms, recent developments, and challenges. *IEEE Transactions on Cybernetics*, 54(12):7173–7186, 2024. 1

[14] Qinglun Zhang, Zhen Liu, Haoqiang Fan, Guanghui Liu, Bing Zeng, and Shuaicheng Liu. Flowpolicy: Enabling fast and robust 3d flow-based policy via consistency flow matching for robot manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 14754–14762, 2025. 1

[15] Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *Proceedings of Robotics: Science and Systems*, 2023. doi: 10.15607/RSS.2023.XIX.016. 1, 3