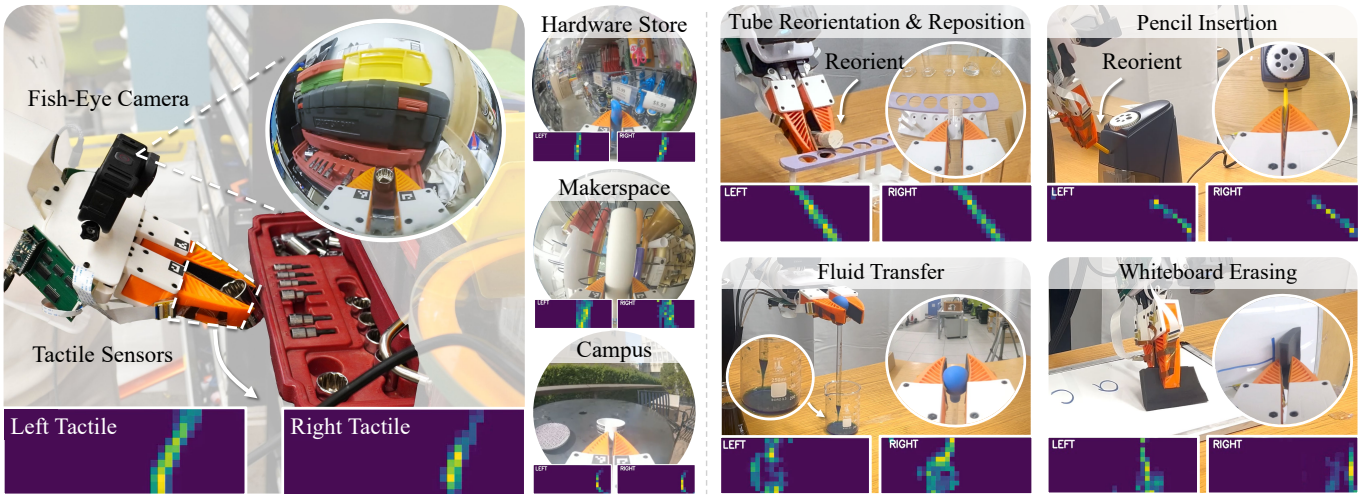
Touch in the Wild: Learning Fine-Grained Manipulation with a Portable Visuo-Tactile Gripper

Xinyue Zhu[*,1] Binghao Huang[*,1] Yunzhu Li[1]
Columbia University[1]
[*]Equal contribution



(a) In-the-Wild Visuo-Tactile Data Collection (b) Visuo-Tactile Policy Rollout

Fig. 1: (a) Our portable handheld gripper enables synchronized collection of visual and tactile data through integrated multimodal sensing, supporting large-scale data collection in the wild. (b) We introduce a multimodal representation learning framework that fuses visual and tactile inputs to support fine-grained downstream manipulation tasks.

***Abstract*—Handheld grippers are increasingly used to collect human demonstrations due to their ease of deployment and versatility. However, most existing designs lack tactile sensing, despite the critical role of tactile feedback in precise manipulation. We present a portable, lightweight gripper with integrated tactile sensors that enables synchronized collection of visual and tactile data in diverse, real-world, and in-the-wild settings. Building on this hardware, we propose a cross-modal representation learning framework that integrates visual and tactile signals while preserving their distinct characteristics. The learned representations are interpretable and consistently emphasize contact regions during physical interactions. When used for downstream manipulation tasks, these representations enable more efficient and effective policy learning, supporting precise robotic manipulation based on multimodal feedback. We validate our approach on fine-grained tasks such as test tube insertion and pipette-based fluid transfer, demonstrating improved accuracy and robustness under external disturbances. Our project page is available at https://binghao-huang.github.io/touch_in_the_wild/**

I. Introduction

Humans naturally rely on both vision and touch when interacting with the physical world. Whether inserting a key into a lock or adjusting a pipette during a lab experiment, tactile feedback plays a critical role in guiding precise motor actions, especially in situations where visual information may be unreliable due to occlusion, poor lighting, or dynamic backgrounds. While vision provides global, semantic context, touch offers local, high-resolution feedback about contact and force. The integration of these two modalities is fundamental to effective manipulation in everyday environments.

Recent advances in handheld grippers have made it easier to collect human demonstrations "in the wild." However, most existing systems focus exclusively on visual sensing, largely neglecting tactile feedback. This gap limits their usefulness for capturing the fine-grained, contact-rich strategies humans use in real-world tasks. Moreover, relying on vision alone makes these systems vulnerable to environmental variability, whereas tactile sensing offers a complementary and robust signal that is invariant to lighting and viewpoint.

Two key challenges have prevented widespread visuo-tactile data collection in natural settings: *(i) Portable tactile hardware.* Most existing tactile sensors are bulky, rigid, or resource-intensive. For example, soft bubble sensors [26] are too large for handheld use, while other optical-based sensors [49] often require external computing resources (e.g., a dedicated PC for image streaming), making them unsuitable for mobile or outdoor deployment. *(ii) Learning from multimodal data.* Tactile and visual signals differ significantly in scale and semantics. Tactile inputs are local and physical, while vision encodes broader spatial context. Learning effective

representations that integrate both modalities–particularly from large-scale, unstructured datasets–remains an open challenge.

To address these issues, we present a Portable Visuo-Tactile System for large-scale data collection and multimodal policy learning in real-world settings. Our contributions are threefold: *(1) A lightweight, handheld visuo-tactile gripper.* We integrate flexible piezoresistive tactile sensors into a soft, handheld gripper to enable portable, high-frequency tactile data collection. The system captures human manipulation strategies across a wide range of environments, indoors and out. *(2) Robust tactile sensing for in-the-wild deployment.* Our choice of piezoresistive sensors enables consistent performance across sensor units and environmental conditions. Unlike optical-based alternatives, our sensors require minimal processing and are unaffected by lighting, making them suitable for uncontrolled environments. The collected data transfers well to robotic platforms, bridging human demonstrations and autonomous execution. *(3) A multimodal representation learning framework.* We propose a masked autoencoding approach that jointly learns from tactile and visual inputs while preserving modality-specific information. This enables policies to leverage fine-grained tactile feedback in a human-like manner, improving both sample efficiency and manipulation accuracy.

We validate our system on fine-grained robotic manipulation tasks in the real world, such as test tube insertion and pipette-based fluid transfer, demonstrating successful policy transfer and robustness to environmental disturbances. Our results underscore the promise of portable visuo-tactile platforms in bridging the gap between human demonstrations and robot learning in complex, real-world settings.

II. Related Wroks

Scalable multi-sensory data and learning. Reinforcement and imitation learning have become central to robotic manipulation [48, 21, 37, 43, 18, 6, 40], but progress is often bottlenecked by the lack of large-scale, high-quality multimodal datasets—especially those involving tactile signals. While simulation can mitigate data scarcity, simulated tactile signals tend to diverge significantly from their real-world counterparts, limiting transferability [1, 37, 42, 35]. This makes scalable real-world visuo-tactile data collection increasingly important. However, acquiring multi-sensory data at scale remains challenging. Beyond RGB streaming, each additional modality [15, 47, 12, 46, 28]–such as audio [31], force, or tactile–introduces added hardware complexity, synchronization overhead, and environmental constraints. As a result, most prior efforts have limited data collection to structured indoor environments [22].

However, tactile sensing is most valuable in uncontrolled, real-world settings, where vision may degrade due to poor lighting or background clutter [41, 14], while contact forces remain stable. Prior "in-the-wild" systems have demonstrated strong visual-only performance on simple tasks [8], but they overlook the complementary benefits of touch. To address this gap, we present a portable, handheld visuo-tactile system that pairs a fisheye RGB camera [8] with a lightweight, flexible tactile array. The device enables synchronized, large-scale collection of visual and tactile data across diverse, unstructured environments. Using this setup, we construct a visuo-tactile dataset and train a masked reconstruction encoder that: (i) accurately reconstructs tactile signals, and (ii) improves downstream visuo-tactile policy learning. Unlike prior work using flexible sensors for narrow tasks (e.g., pose or trajectory prediction) without pretraining, our approach supports scalable, cross-modal representation learning for general-purpose manipulation.

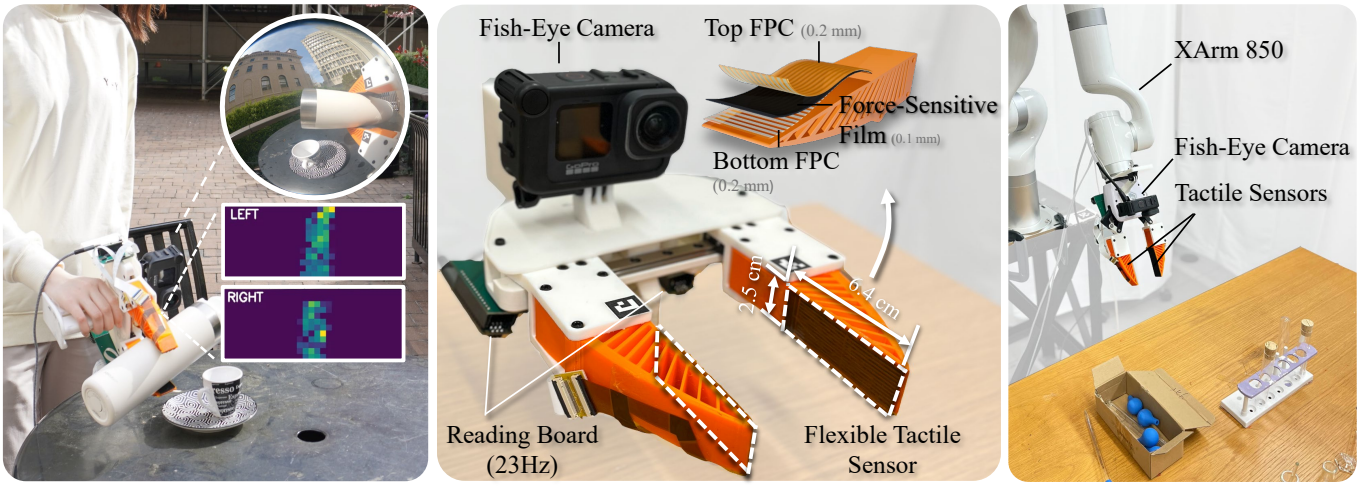
Visuo-tactile manipulation. Tactile feedback plays a critical role in human manipulation, particularly when visual input is occluded or ambiguous [24, 23]. Similarly, tactile sensing enables robots to interact more reliably with objects and environments that are complex or require precision. As a result, there is growing interest in integrating vision and touch to improve robotic manipulation [22, 39, 11, 27, 40, 5, 19, 51, 32, 5, 9, 16, 45]. Much of this prior work relies on optical tactile sensors, which image surface deformations to infer contact geometry and texture [3, 50, 51, 25, 4, 44, 12]. While these sensors provide rich signals, they are typically rigid and bulky, limiting their applicability in portable or unstructured settings.

In contrast, we use thin, flexible tactile sensors[22, 38, 13] that directly measure force distributions over the contact surface. Embedded in soft robotic fingers, these sensors provide consistent, object-agnostic representations that are easier to generalize and better suited for scalable, real-world learning. Although such sensors have been explored in lab settings[22, 30, 2], large-scale, in-the-wild visuo-tactile datasets with synchronized RGB and tactile data capturing physical interactions have not previously been available. Our work fills this gap by collecting–and publicly releasing–a diverse, real-world visuo-tactile dataset. It spans a wide range of tasks and environments, laying the foundation for scalable multimodal learning and robust policy development.

III. Visuo-Tactile Data Collection System

A. Scalable Flexible Tactile Sensors

As shown in Fig. 2 (a), we embed thin, matrix-based tactile pads into the soft, fin-shaped fingers of our handheld gripper. The sensor architecture builds on the triple-layer design from 3D-ViTac[22], adapted to fit the geometry of the adaptive fin-shaped gripper [8]. Each tactile pad consists of a piezoresistive sensing layer sandwiched between two flexible printed circuits (FPCs). To accommodate the elongated, flexible fins, we introduce two key modifications: *(1) Higher spatial resolution.* Capturing contact patterns along the finger's length requires denser spatial sampling. The stainless-steel electrodes used in 3D-ViTac limit both resolution and signal stability. By replacing them with FPC electrodes, we achieve uniform trace pitch, improved robustness, and a per-pad resolution of 12×32 taxels, each measuring a $2 \times 2mm^2$ area. This allows us to capture fine-grained, dynamic contact signals. *(2) Rapid, scalable fabrication.* The use of FPCs enables tool-free assembly. Each pad can be fabricated in under five minutes



(a) Data Collection in the Wild (b) Portable Visuo-Tactile Data Collection Gripper (c) Robot Setup

Fig. 2: (a) Multimodal data collection in the wild using our portable visuo-tactile system, with example tactile signals from both fingers. (b) Close-up of the handheld gripper, equipped with flexible tactile sensors and a fisheye camera for synchronized visual-tactile capture. (c) Robotic setup for downstream tasks, featuring an xArm 850 with the same sensor configuration.

and mounted on the gripper in an additional two, supporting scalable deployment for large-scale tactile data collection.

B. Portable Multi-Modal Sensing System

In-the-wild large-scale data collection. To enable real-world visuo-tactile data collection at scale, we design a compact and ergonomic handheld gripper that integrates both sensing modalities. Each tactile pad connects to a custom Arduino-based PCB, with two boards neatly housed beneath the gripper's palm (Fig. 2 (a)). The full handheld unit–including batteries–weighs approximately 962 g, making it comfortable for prolonged use.

At the firmware level, we optimize the serial protocol to stream each 12×32 pad at 23 Hz, providing frame-accurate synchrony with the camera information. Tactile frames are timestamped directly on the microcontroller and transmitted over USB to a host device (e.g., a laptop or NVIDIA Jetson Orin Nano). The battery-powered, handbag-sized system is easily deployed in grocery stores, outdoor markets, and other unstructured, in-the-world environments (Fig. 2 (a)), enabling high-throughput and scalable visuo-tactile data collection.

Multi-modal data synchronization. Precise alignment between vision and touch is essential for learning effective visuo-tactile representations. Although both the tactile system and the GoPro camera operate above 20 Hz (e.g., tactile sensors at 23 Hz, GoPro 9 at 60 Hz), aligning their data streams poses challenges due to clock drift and limited timestamp precision on the camera.

We address this with a hardware-free synchronization strategy: *(i) Video stream.* Before each demonstration, a QR code displaying the current host time is shown to the camera, refreshed at 30 Hz. *(ii) Tactile stream.* Tactile data is published via ROS2 at 23 Hz, with each packet carrying a host-clock timestamp. *(iii) Post-processing.* During offline processing, we decode the QR code sequence from the video, recover exact host timestamps for each frame, and align them with the tactile data using the shared clock reference. This procedure yields tightly aligned visual and tactile recordings–without the need for external synchronization hardware–enabling accurate multimodal supervision for downstream learning.

IV. Visuo-Tactile Representation and Policy Learning

To perform precise manipulation, robots must integrate visual and tactile signals that differ substantially in both content and structure. RGB images provide global, semantic context—such as object identity and workspace layout—while tactile signals offer local, contact-rich feedback that is often occluded in vision [29]. These complementary modalities follow different statistical distributions, making it non-trivial to learn unified representations that preserve modality-specific information while enabling effective cross-modal reasoning. We propose a two-stage learning framework (Fig. 3) that first learns a joint visuo-tactile representation via masked tactile reconstruction, and then integrates the learned representation into a diffusion policy [7] for downstream manipulation tasks.

A. Problem Formulation

Let $\mathcal{D}_{\text{pretrain}} = \{(I, T)\}$ be a large-scale dataset of synchronized RGB-tactile frame pairs, where $I \in \mathbb{R}^{3\times224\times224}$ is an RGB image captured from a wrist-mounted fisheye camera, and $T \in \mathbb{R}^{3\times24\times32}$ is a colormapped tactile image composed of vertically stacked fingertip sensor readings. The goal is to learn a multimodal encoder E_ϕ that fuses these two modalities into a joint representation $\mathbf{z}_{\text{fusion}} = E_\phi(I, T)$ that preserves modality-specific structure and supports both self-supervised pertaining and downstream manipulation tasks.

We divide the learning into two stages: *(1) Stage 1:* Pretrain E_ϕ using a masked autoencoding objective that reconstructs full tactile images from partially observed tactile input and corresponding visual frames. *Stage 2:* Use the pretrained encoder within a diffusion policy to learn manipulation behaviors from demonstrations.

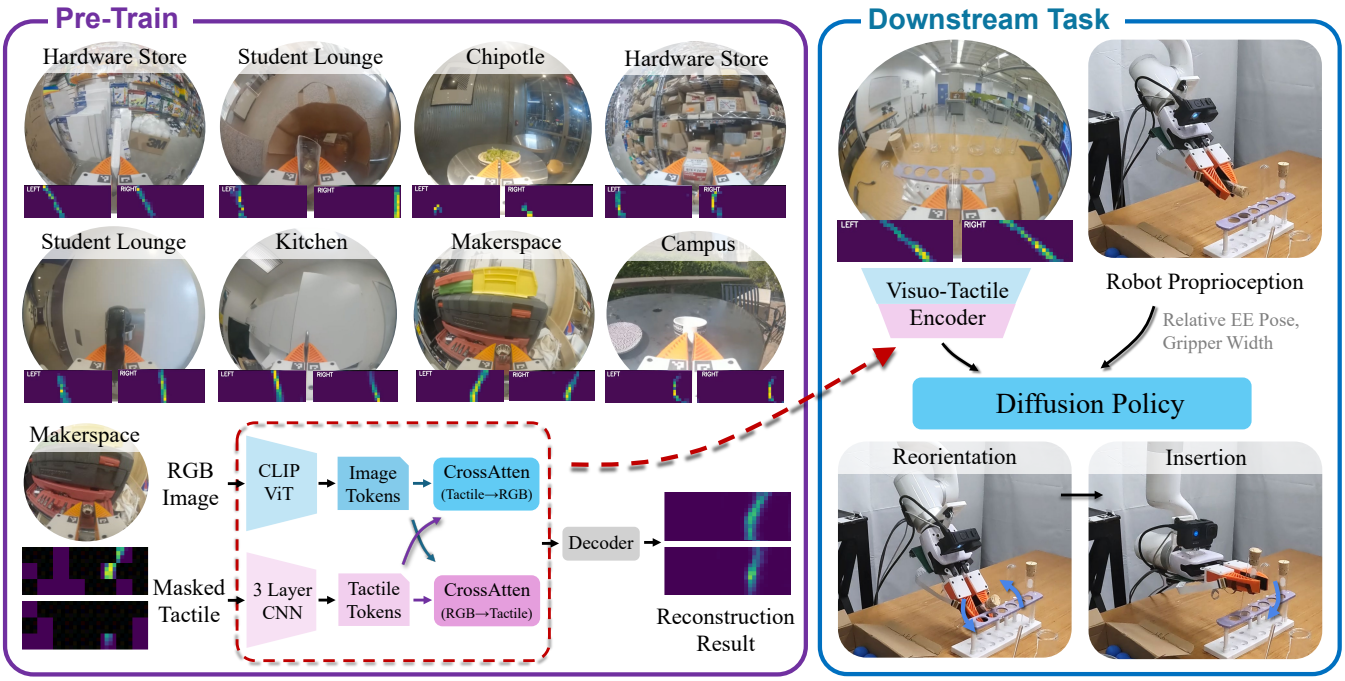


Fig. 3: **Method Overview** of our two-stage pipeline. *Stage 1:* We pretrain a visuo-tactile encoder via cross-modal reconstruction using a large-scale dataset collected across diverse indoor and outdoor environments. *Stage 2:* The pretrained encoder is combined with robot proprioception to condition a diffusion policy for downstream tasks such as object reorientation and insertion.

B. Stage 1: Visuo-Tactile Representation Learning

While prior work often relies on contrastive learning to align embeddings from different modalities[10, 17], such objectives tend to suppress the fine-grained, geometry-sensitive signals captured by tactile sensors. Instead, we adopt a *masked autoencoding objective* [20], which reconstructs missing tactile regions conditioned on partially observed tactile input and visual context. This formulation encourages the encoder to retain tactile-specific information while leveraging vision for inference.

Formally, we optimize the encoder E_ϕ to minimize the expected tactile reconstruction loss:

$$E_\phi = \arg\min_\phi \ \mathbb{E}_{(I,T)\sim\mathcal{D}_{\text{pretrain}}} \left[\left\| T - \hat{T} \right\|_2^2 \right], \tag{1}$$

where $\hat{T} = \text{Dec}(E_\phi(I, T_{\text{vis}}))$ is the predicted tactile image from the fused visuo-tactile embedding.

Tactile encoder. Each tactile reading consists of two fingertip arrays, each of shape $3 \times 12 \times 32$, stacked vertically to form a $3 \times 24 \times 32$ RGB tactile image. This image is divided into non-overlapping 4×4 patches, resulting in a 6×8 patch grid. During training, we randomly mask 60–80% of the patches in 95% of samples using a learnable token T_{mask}; the remaining 5% are shown in full. The masked tactile input is defined as:

$$T_{\text{vis}} = M \odot T + (1 - M) \odot T_{\text{mask}}, \tag{2}$$

where $M \in \{0,1\}^{6\times8}$ is a binary patch mask. The visible tactile input T_{vis} is processed by a 3-layer CNN to produce a 768-dimensional embedding $\mathbf{z}_{\text{tac}}$.

Vision encoder. The RGB image I is processed by a ViT-B/16 encoder initialized from CLIP [33]. We fine-tune all layers with a learning rate of 3×10^{-5}, and extract the final [CLS] token as the 768-dimensional visual embedding $\mathbf{z}_{\text{img}}$.

Cross-modal fusion. To integrate the tactile and visual features, we apply two rounds of multi-head cross-attention (MHAttn):

$$\mathbf{z}'_{\text{tac}} = \text{MHAttn}(Q = \mathbf{z}_{\text{tac}},\ K = \mathbf{z}_{\text{img}},\ V = \mathbf{z}_{\text{img}}) \xrightarrow{\text{LayerNorm}} \mathbf{z}''_{\text{tac}}, \tag{3}$$

$$\mathbf{z}'_{\text{img}} = \text{MHAttn}(Q = \mathbf{z}_{\text{img}},\ K = \mathbf{z}''_{\text{tac}},\ V = \mathbf{z}''_{\text{tac}}) \xrightarrow{\text{LayerNorm}} \mathbf{z}''_{\text{img}}. \tag{4}$$

We concatenate the updated embeddings to obtain the fused representation:

$$\mathbf{z}_{\text{fusion}} = \left[\mathbf{z}''_{\text{tac}};\ \mathbf{z}''_{\text{img}}\right] \in \mathbb{R}^{2d}. \tag{5}$$

Tactile reconstruction decoder. The fused feature $\mathbf{z}_{\text{fusion}}$ is passed through a two-layer MLP followed by a sigmoid activation to produce the reconstructed tactile image $\hat{T} \in (0,1)^{1\times24\times32}$, where $\hat{T} = \text{Dec}(\mathbf{z}_{\text{fusion}})$. We use a full-image reconstruction loss:

$$\mathcal{L}_{\text{stage1}} = \left\| T - \hat{T} \right\|_2^2, \tag{6}$$

which encourages both local contact inference and global structural understanding.

Stabilization via EMA. We maintain a target encoder updated via exponential moving average (EMA) of the online weights with a decay factor of 0.995. The tactile CNN and cross-attention layers are optimized with a learning rate of 1×10^{-4}, while the CLIP backbone is fine-tuned with 3×10^{-5}. The EMA encoder is used for checkpointing and attention visualization.

C. Stage 2: Policy Learning via Behavior Cloning

Once pretrained, the visuo-tactile encoder E_ϕ is integrated into a conditional diffusion policy π_θ for downstream robotic manipulation tasks.

Observation space. At each timestep t, the robot observes: $\mathbf{o}_t = (I_t, T_t, p_t)$, where p_t denotes the robot's proprioceptive state (e.g., end-effector pose, gripper width). The encoder produces the fused visuo-tactile embedding:

$$\mathbf{z}_t = E_\phi(I_t, T_t). \tag{7}$$

Diffusion policy. We adopt a conditional diffusion policy π_θ [7], which iteratively refines a noisy action sequence over K denoising steps. At each step, the model conditions on $\mathbf{z}_t$ and p_t to generate the next action $\hat{\mathbf{a}}_t$. Given a demonstration dataset $\mathcal{D}_{\text{demo}} = \{(\mathbf{o}_t, \mathbf{a}_t)\}$, we train the policy with a behavior cloning loss:

$$\mathcal{L}_{\text{stage2}}(\theta, \phi) = \sum_t \|\hat{\mathbf{a}}_t - \mathbf{a}_t\|_2^2, \tag{8}$$

where $\hat{\mathbf{a}}_t = \pi_\theta(\mathbf{z}_t, p_t)$.

Training details. We use Diffusion Policy's convolutional U-Net [34] with DDIM inference [36]. The model is conditioned on the fused visuo-tactile embedding and two consecutive proprioceptive observations. All encoder components—including the CLIP backbone, tactile CNN, and cross-attention layers—are fine-tuned during this stage using a learning rate of 3×10^{-5}.

V. Experiments

In this section, we address several key questions regarding the role of touch in fine-grained manipulation and the impact of different pretraining strategies on downstream tasks. Specifically, we investigate: (1) How does touch improve fine-grained manipulation? (2) How does a visuo-tactile encoder trained on extensive data aid policy learning? (3) How do variations in pretraining - such as the number of demonstrations or training epochs - affect downstream task performance?

A. Large-Scale Visuo-Tactile Data and Pre-Training

To enable effective visuo-tactile pretraining, we curated a diverse dataset consisting of over 2.6 million visuo-tactile pairs collected from 12 indoor and outdoor environments. The dataset includes more than 2,700 demonstrations and covers 43 manipulation tasks. The data can be divided into three categories: (1) the four main tasks presented in this paper, (2) additional unstructured indoor tasks to enhance diversity, and (3) over 30 in-the-wild tasks that capture complex real-world scenarios.

We assess the quality of the representation learned by our pretrained visuo-tactile encoder using two complementary analyses. First, we input masked tactile and RGB images into the encoder to assess its ability to reconstruct the missing tactile images. This tests whether the encoder has learned meaningful cross-modal associations and can generalize to both in-distribution scenarios—environments seen during training—and out-of-distribution scenarios involving entirely novel backgrounds. Second, we visualize the encoder's attention by extracting self-attention maps from the final layer of the ViT module. This allows us to examine whether the model consistently attends to relevant contact regions in the RGB images across different environments. Figure 4 presents qualitative examples from four tasks: two from in-distribution environments and two from out-of-distribution test settings.

B. Experiments Setup

We evaluate our multi-modal sensing and learning system on four challenging real-world robotic tasks. Below are the basic descriptions and evaluation metrics for all tasks:

(1) Task Requiring In-Hand State Information

Transparent Tube Collection. The robot must pick up a test tube from a box, reorient it in-hand using the test tube rack, and precisely insert it into the test tube rack. *Evaluation Metric:* The task is considered successful if the test tube is fully inserted into the test tube rack without breaking.

Pencil Insertion. The robot needs to insert a pencil into a sharpener. Since the pencil is initially grasped upright, the robot must first reorient it before performing a precise insertion. *Evaluation Metric:* The task is considered successful if the pencil is accurately inserted into the sharpener.

(2) Task Requiring Fine-Grained Force Information

Fluid Transfer. The robot uses a pipette to transfer water between containers. It must grasp the pipette firmly, apply just enough pressure to extract liquid without dropping it. Then the robot needs to move to the top of other container and gently squeeze to release the water into it. This task demands continuous and sensitive force modulation. *Evaluation Metric:* The task is considered successful if the water is transferred into the second container without spilling.

Whiteboard Erasing. The robot uses a soft eraser to remove two strokes of text from the whiteboard. It must apply the right amount of pressure to erase the marker ink without exceeding force limits that could damage the system. The task requires consistent and controlled force application throughout. *Evaluation Metric:* The task is considered successful if all visible marker ink is removed from the whiteboard.

In the experiments, we compare our methods with the following baselines.

(1) Vision Only. This method feeds one RGB image as input to CLIP, and extracts a 768-dimensional CLIP embedding. This embedding, along with the robot's proprioceptive information, is then fed into the image-based diffusion policy. We follow the same implementation as outlined in [8].

(2) Vision w/ CNN Tactile Fusion. This method processes two tactile images—one from the left gripper and one from the right gripper—through a 3-layer CNN. The features extracted by the CNN form a 512-dimensional vector, which is then concatenated with the visual information and used as conditioning input along with other proprioceptive states for diffusion policy.

(3) Ours w/o Pretraining. This method uses the visual-tactile encoder proposed in the paper, with the vision backbone initialized from CLIP and the other parts of the joint encoder

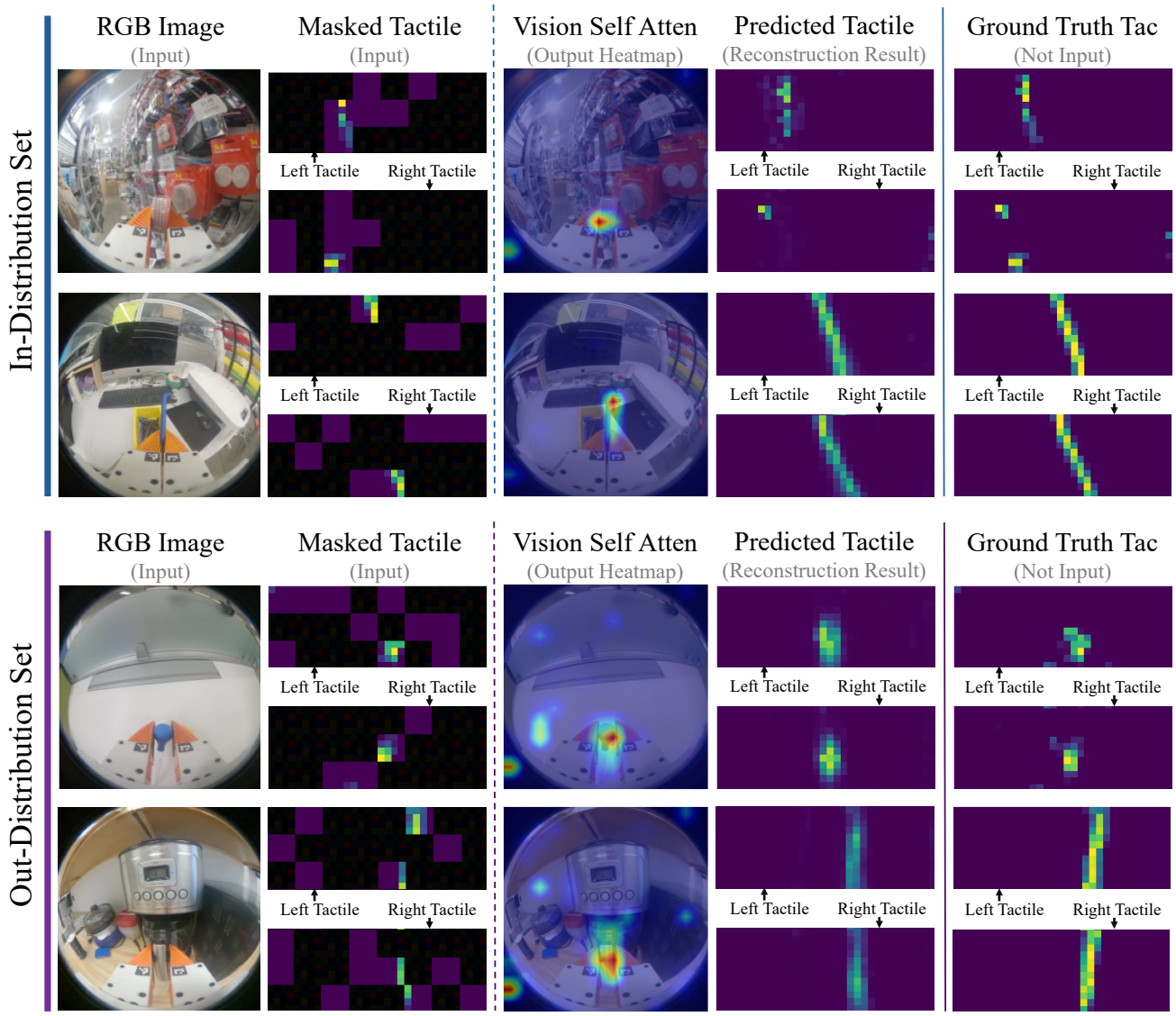


Fig. 4: **Qualitative Results of Pre-Training.** We present four examples showcasing the results of our visuo-tactile pretrained encoder, highlighting both tactile image reconstruction and ViT self-attention heatmaps. For tactile image reconstruction, the encoder successfully reconstructs tactile images for both in-distribution and out-of-distribution inputs. Additionally, we observe that the vision encoder consistently focuses on the gripper contact region, regardless of the background or whether the object is seen or unseen.

initialized from scratch. The output embedding from visual-tactile encoder is concatenated with other robot proprioceptive states as conditioning for the diffusion policy.

(4) Ours w/ Pretraining. This method uses the visual-tactile encoder proposed in the paper with pretrained weights. Similarly, the output embedding from the visual-tactile encoder is concatenated with other robot proprioceptive states as conditioning for the diffusion policy.

For each of the four tasks, we conduct 20 trials with slight randomization in the initial robot position and environmental conditions. All policies are trained for 60 epochs, at which point they have converged. The results are presented in Table I. Our pretrained policies consistently outperform all baseline policies across the four evaluated tasks.

C. Qualitative Analysis

Our system enhances a handheld gripper by integrating tactile sensing and training a large-scale visuo-tactile encoder to further improve manipulation policies. We observe three key benefits from incorporating touch and leveraging pretrained representations. (1) *Tactile feedback provides explicit in-hand pose information.* In a single-camera handheld setup, visual inputs often suffer from severe occlusions. For example, in the test tube insertion task, the vision-only policy relied heavily on the color of the wooden cork to infer orientation. A minor change—such as switching to a lighter cork with less distinct features—confused the vision model and degraded performance in reorientation. The tactile policy, however, remained unaffected by such variations. (2) *Tactile feedback improves detection of critical state transitions.* In fine-grained

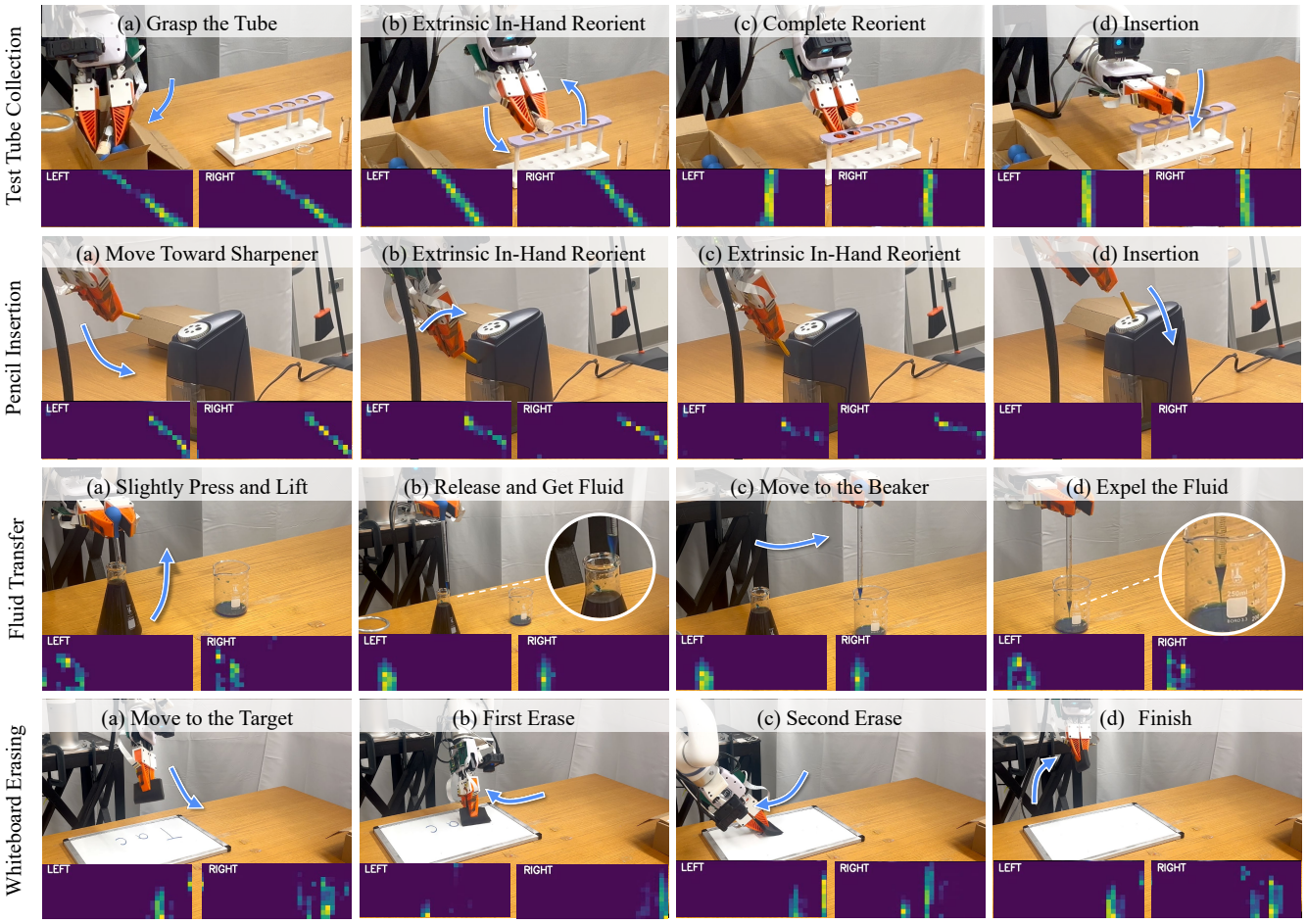


Fig. 5: **Quantitative Results.** We evaluate our visuo-tactile policy across four fine-grained manipulation tasks. Detailed descriptions and metrics can be found in Sec. V

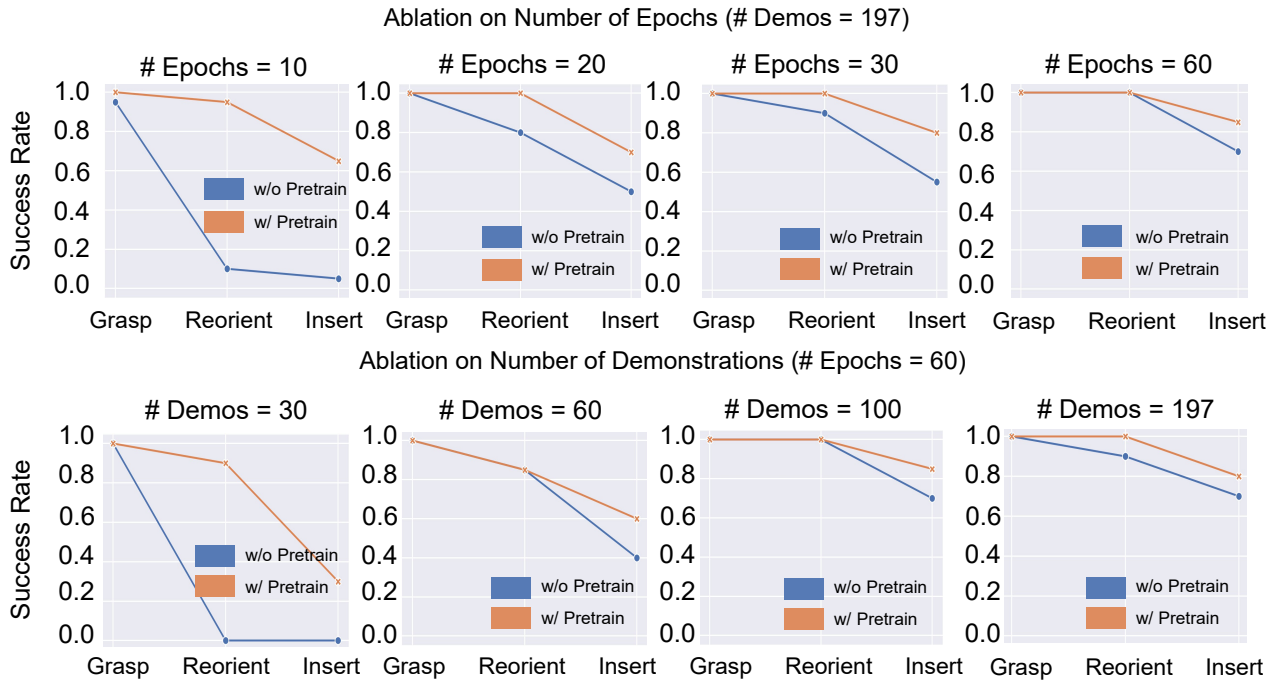


Fig. 6: Ablation results for varying numbers of demonstrations and epochs for Transparent Tube Collection Task. Our findings show that the policy with pretraining consistently outperforms the policy without pretraining, both in low-epoch and low-demonstration regimes.

Tasks Requiring In-Hand State Information

Modalities	Transparent Tube Collection				Pencil Insertion		
	Grasp	**Reorient**	**Insert**	**Whole Task**	**Reorient**	**Insert**	**Whole Task**
Vision Only	**1.00**	0.25	0.25	0.25	0.50	0.65	0.45
CNN Tactile Fusion	**1.00**	**1.00**	0.50	0.50	0.80	0.75	0.70
Ours w/o Pretraining	**1.00**	**1.00**	0.70	0.70	0.80	0.80	0.75
Ours w/ Pretraining	**1.00**	**1.00**	**0.85**	**0.85**	**0.95**	**0.90**	**0.85**

Tasks Requiring Fine-Grained Force Information

Modalities	Fluid Transfer				Whiteboard Erasing		
	Acquire	**Transfer**	**Expel**	**Whole Task**	**First Erase**	**Second Erase**	**Whole Task**
Vision Only	0.95	0.85	0.55	0.55	0.65	0.65	0.55
CNN Tactile Fusion	0.90	0.75	0.75	0.70	**0.70**	0.50	0.45
Ours w/o Pretraining	**1.00**	0.90	0.80	0.80	0.60	**0.75**	0.60
Ours w/ Pretraining	**1.00**	**1.00**	**0.90**	**0.90**	**0.70**	**0.75**	**0.70**

TABLE I: **Comparison with Baselines.** We evaluate our policy over 20 episodes and the best performance for each task is bolded.

force-controlled tasks like fluid transfer, accurately identifying when one action phase ends and another begins is essential. The vision-only policy struggles with this because visual features—such as the gripper's width—remain similar before and after contact, making it difficult to determine if the gripper is still squeezing or has completed the action. As a result, the policy often skips the "expel water" phase prematurely, incorrectly assuming the transfer is finished. In contrast, the tactile policy receives direct feedback from pressure changes, enabling it to detect subtle shifts in force and correctly infer when the water has been fully expelled. This reliable tactile signal helps the policy transition smoothly between task stages, improving both accuracy and robustness.

(3) *Joint visuo-tactile encoders enable more coordinated use of vision and touch.* Naive fusion approaches that simply concatenate visual and tactile features often fail to meaningfully combine the two modalities. As a result, the policy may over-rely on one input while neglecting the other. This imbalance was evident in the whiteboard erasing task: the CNN tactile fusion policy applied excessive force to maximize tactile signal changes, which triggered safety warnings and caused the task to fail. In contrast, our jointly trained visuo-tactile encoder learns to coordinate both modalities, allowing the policy to modulate force more appropriately based on visual context and tactile feedback. This balanced integration reduces failure cases caused by overuse or misuse of either modality.

Pretraining Ablations: Varying Number of Demonstrations and Epochs.

To evaluate the effectiveness of pretraining, we assess the performance of the Transparent Tube Collection task under varying number of demonstrations and epochs. The results are shown in Fig. 6. We found that pretraining provides significant benefits, particularly in low-data and low-epoch training settings.

(1) Low-Data Regime (Fewer Than 60 Demonstrations). When only 30 or 60 demonstrations are available, the policy initialized without pretraining often hesitates after the grasping stage to uncertainty about the next step. In contrast, the policy pretrained for just 30 demonstrations follows smoother trajectories and typically only encounters failure during the final insertion phase. We believe that pretraining helps the joint encoder learn visual-tactile patterns early on, which enables the downstream policy to focus on learning effective action trajectories.

(2) Low-Epoch Regime (Fewer Than 60 Epochs). In the low-epoch regime, we observed that the policy without pretraining was more sensitive to initial environmental configurations. For instance, when the test tube was placed at a steep incline and was difficult to grasp, an imperfect grasp position had a considerable impact on the execution of the reorientation task. The policy without pretraining sometimes over- or under-reoriented, resulting in failure. We believe that in the low-epoch regime, the pretrained policy benefits from prior knowledge that emphasizes tactile cues, which makes it less reliant on noisy environmental factors for decision-making.

VI. Conclusion and Limitations

We present a handheld gripper enhanced with tactile sensing and accompanied by a large-scale visuo-tactile dataset. To demonstrate its utility, we pre-train a joint encoder on this data and evaluate it across several fine-grained manipulation tasks. Our study uses a single-arm robot with a parallel gripper—an inherently versatile but relatively simple end-effector. Because tactile feedback can unlock even richer behaviors on multi-finger hands, future work will extend our approach to dexterous grippers and explore more intricate manipulation skills.

References

[1] Iretiayo Akinola, Jie Xu, Jan Carius, Dieter Fox, and Yashraj Narang. Tacsl: A library for visuotactile sensor simulation and learning. *arXiv*, 2024.

[2] A. Bicchi and V. Kumar. Robotic grasping and contact: a review. In *Proceedings 2000 ICRA. Millennium Conference. IEEE International Conference on Robotics and Automation. Symposia Proceedings (Cat. No.00CH37065)*, volume 1, pages 348–353 vol.1, 2000. doi: 10.1109/ROBOT.2000.844081.

[3] Roberto Calandra, Andrew Owens, Manu Upadhyaya, Wenzhen Yuan, Justin Lin, Edward H Adelson, and Sergey

Levine. The feeling of success: Does touch sensing help predict grasp outcomes? *arXiv preprint arXiv:1710.05512*, 2017.

[4] Roberto Calandra, Andrew Owens, Dinesh Jayaraman, Justin Lin, Wenzhen Yuan, Jitendra Malik, Edward H Adelson, and Sergey Levine. More than a feeling: Learning to grasp and regrasp using vision and touch. *IEEE Robotics and Automation Letters*, 3(4):3300–3307, 2018.

[5] Yizhou Chen, Mark Van der Merwe, Andrea Sipos, and Nima Fazeli. Visuo-tactile transformers for manipulation. In *6th Annual Conference on Robot Learning*, 2022.

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

[7] 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*, 2024.

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

[9] Vedant Dave, Fotios Lygerakis, and Elmar Rueckert. Multimodal visual-tactile representation learning through self-supervised contrastive pre-training. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8013–8020, 2024. doi: 10.1109/ICRA57147.2024.10610228.

[10] Vedant Dave, Fotios Lygerakis, and Elmar Rueckert. Multimodal visual-tactile representation learning through self-supervised contrastive pre-training. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8013–8020, 2024. doi: 10.1109/ICRA57147.2024.10610228.

[11] Zihao Ding, Guodong Chen, Zhenhua Wang, and Lining Sun. Adaptive visual–tactile fusion recognition for robotic operation of multi-material system. *Frontiers in Neurorobotics*, 17, 2023. ISSN 1662-5218. doi: 10.3389/fnbot.2023.1181383. URL https://www.frontiersin.org/articles/10.3389/fnbot.2023.1181383.

[12] Yiming Dou, Fengyu Yang, Yi Liu, Antonio Loquercio, and Andrew Owens. Tactile-augmented radiance fields. *arXiv preprint arXiv:2405.04534*, 2024.

[13] Ruoxuan Feng, Jiangyu Hu, Wenke Xia, Tianci Gao, Ao Shen, Yuhao Sun, Bin Fang, and Di Hu. Anytouch: Learning unified static-dynamic representation across multiple visuo-tactile sensors. *arXiv preprint arXiv:2502.12191*, 2025.

[14] Chelsea Finn, Tianhe Yu, Tianhao Zhang, Pieter Abbeel, and Sergey Levine. One-shot visual imitation learning via meta-learning. In *Conference on robot learning*, pages 357–368. PMLR, 2017.

[15] Ruohan Gao, Zilin Si, Yen-Yu Chang, Samuel Clarke, Jeannette Bohg, Li Fei-Fei, Wenzhen Yuan, and Jiajun Wu. Objectfolder 2.0: A multisensory object dataset for sim2real transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10598–10608, 2022.

[16] Abraham George, Selam Gano, Pranav Katragadda, and Amir Barati Farimani. Vital pretraining: Visuo-tactile pretraining for tactile and non-tactile manipulation policies. *arXiv preprint arXiv:2403.11898*, 2024.

[17] Irmak Guzey, Ben Evans, Soumith Chintala, and Lerrel Pinto. Dexterity from touch: Self-supervised pre-training of tactile representations with robotic play, 2023.

[18] Huy Ha, Yihuai Gao, Zipeng Fu, Jie Tan, and Shuran Song. UMI on legs: Making manipulation policies mobile with manipulation-centric whole-body controllers. In *Proceedings of the 2024 Conference on Robot Learning*, 2024.

[19] Johanna Hansen, Francois Hogan, Dmitriy Rivkin, David Meger, Michael Jenkin, and Gregory Dudek. Visuotactile-rl: Learning multimodal manipulation policies with deep reinforcement learning. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 8298–8304, 2022. doi: 10.1109/ICRA46639.2022.9812019.

[20] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[21] Binghao Huang, Yuanpei Chen, Tianyu Wang, Yuzhe Qin, Yaodong Yang, Nikolay Atanasov, and Xiaolong Wang. Dynamic handover: Throw and catch with bimanual hands. *arXiv preprint arXiv:2309.05655*, 2023.

[22] Binghao Huang, Yixuan Wang, Xinyi Yang, Yiyue Luo, and Yunzhu Li. 3d-vitac: Learning fine-grained manipulation with visuo-tactile sensing. *arXiv preprint arXiv:2410.24091*, 2024.

[23] Per Jenmalm and Roland S. Johansson. Visual and somatosensory information about object shape control manipulative fingertip forces. *Journal of Neuroscience*, 17(11):4486–4499, 1997. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.17-11-04486.1997. URL https://www.jneurosci.org/content/17/11/4486.

[24] Roland S. Johansson and Göran Westling. Roles of glabrous skin receptors and sensorimotor memory in automatic control of precision grip when lifting rougher or more slippery objects. *Experimental Brain Research*, 56:550–564, 2004. URL https://api.semanticscholar.org/CorpusID:16631166.

[25] Raj Kolamuri, Zilin Si, Yufan Zhang, Arpit Agarwal, and Wenzhen Yuan. Improving grasp stability with rotation measurement from tactile sensing. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6809–6816. IEEE, 2021.

[26] Naveen Kuppuswamy, Alex Alspach, Avinash Uttamchandani, Sam Creasey, Takuya Ikeda, and Russ Tedrake. Soft-

bubble grippers for robust and perceptive manipulation. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9917–9924. IEEE, 2020.

[27] Michelle A Lee, Yuke Zhu, Krishnan Srinivasan, Parth Shah, Silvio Savarese, Li Fei-Fei, Animesh Garg, and Jeannette Bohg. Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8943–8950. IEEE, 2019.

[28] Hao Li, Yizhi Zhang, Junzhe Zhu, Shaoxiong Wang, Michelle A Lee, Huazhe Xu, Edward Adelson, Li Fei-Fei, Ruohan Gao, and Jiajun Wu. See, hear, and feel: Smart sensory fusion for robotic manipulation. *arXiv preprint arXiv:2212.03858*, 2022.

[29] Yunzhu Li, Jun-Yan Zhu, Russ Tedrake, and Antonio Torralba. Connecting touch and vision via cross-modal prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10609–10618, 2019.

[30] Fangchen Liu, Chuanyu Li, Yihua Qin, Ankit Shaw, Jing Xu, Pieter Abbeel, and Rui Chen. Vitamin: Learning contact-rich tasks through robot-free visuo-tactile manipulation interface. *arXiv preprint arXiv:2504.06156*, 2025.

[31] Zeyi Liu, Cheng Chi, Eric Cousineau, Naveen Kuppuswamy, Benjamin Burchfiel, and Shuran Song. Maniwav: Learning robot manipulation from in-the-wild audio-visual data. *arXiv preprint arXiv:2406.19464*, 2024.

[32] Yuzhe Qin, Binghao Huang, Zhao-Heng Yin, Hao Su, and Xiaolong Wang. Dexpoint: Generalizable point cloud reinforcement learning for sim-to-real dexterous manipulation. In *Conference on Robot Learning*, pages 594–605. PMLR, 2023.

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[34] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[35] Zilin Si and Wenzhen Yuan. Taxim: An example-based simulation model for gelsight tactile sensors. *IEEE Robotics and Automation Letters*, 7(2):2361–2368, 2022.

[36] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv:2010.02502*, October 2020. URL https://arxiv.org/abs/2010.02502.

[37] Entong Su, Chengzhe Jia, Yuzhe Qin, Wenxuan Zhou, Annabella Macaluso, Binghao Huang, and Xiaolong Wang. Sim2real manipulation on unknown objects with tactile-based reinforcement learning. *arXiv preprint arXiv:2403.12170*, 2024.

[38] Subramanian Meenakshi Sundaram, Petr Kellnhofer, Yunzhu Li, Jun-Yan Zhu, Antonio Torralba, and Wojciech Matusik. Learning the signatures of the human grasp using a scalable tactile glove. *Nature*, 569:698 – 702, 2019. URL https://api.semanticscholar.org/CorpusID:169033286.

[39] Sudharshan Suresh, Haozhi Qi, Tingfan Wu, Taosha Fan, Luis Pineda, Mike Lambeta, Jitendra Malik, Mrinal Kalakrishnan, Roberto Calandra, Michael Kaess, et al. Neural feels with neural fields: Visuo-tactile perception for in-hand manipulation. *arXiv preprint arXiv:2312.13469*, 2023.

[40] Herke van Hoof, Nutan Chen, Maximilian Karl, Patrick van der Smagt, and Jan Peters. Stable reinforcement learning with autoencoders for tactile and visual data. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3928–3934, 2016. doi: 10.1109/IROS.2016.7759578.

[41] Chen Wang, Linxi Fan, Jiankai Sun, Ruohan Zhang, Li Fei-Fei, Danfei Xu, Yuke Zhu, and Anima Anandkumar. Mimicplay: Long-horizon imitation learning by watching human play. *arXiv preprint arXiv:2302.12422*, 2023.

[42] Shaoxiong Wang, Mike Lambeta, Po-Wei Chou, and Roberto Calandra. Tacto: A fast, flexible, and open-source simulator for high-resolution vision-based tactile sensors. *IEEE Robotics and Automation Letters*, 7(2):3930–3937, 2022.

[43] Yixuan Wang, Guang Yin, Binghao Huang, Tarik Kelestemur, Jiuguang Wang, and Yunzhu Li. Gendp: 3d semantic fields for category-level generalizable diffusion policy. In *8th Annual Conference on Robot Learning*, 2024.

[44] 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.

[45] Fengyu Yang, Chenyang Ma, Jiacheng Zhang, Jing Zhu, Wenzhen Yuan, and Andrew Owens. Touch and go: Learning from human-collected vision and touch. *arXiv preprint arXiv:2211.12498*, 2022.

[46] Fengyu Yang, Jiacheng Zhang, and Andrew Owens. Generating visual scenes from touch. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22070–22080, 2023.

[47] Fengyu Yang, Chao Feng, Ziyang Chen, Hyoungseob Park, Daniel Wang, Yiming Dou, Ziyao Zeng, Xien Chen, Rit Gangopadhyay, Andrew Owens, et al. Binding touch to everything: Learning unified multimodal tactile representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26340–26353, 2024.

[48] Zhao-Heng Yin, Binghao Huang, Yuzhe Qin, Qifeng Chen, and Xiaolong Wang. Rotating without seeing: Towards in-hand dexterity through touch. *arXiv preprint arXiv:2303.10880*, 2023.

[49] Wenzhen Yuan, Siyuan Dong, and Edward H Adelson.

Gelsight: High-resolution robot tactile sensors for estimating geometry and force. *Sensors*, 17(12):2762, 2017.

[50] Wenzhen Yuan, Yuchen Mo, Shaoxiong Wang, and Edward H Adelson. Active clothing material perception using tactile sensing and deep learning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4842–4849. IEEE, 2018.

[51] Ying Yuan, Haichuan Che, Yuzhe Qin, Binghao Huang, Zhao-Heng Yin, Kang-Won Lee, Yi Wu, Soo-Chul Lim, and Xiaolong Wang. Robot synesthesia: In-hand manipulation with visuotactile sensing. *arXiv preprint arXiv:2312.01853*, 2023.