

TouchGuide: Inference-Time Steering of Visuomotor Policies via Touch Guidance

Zhemeng Zhang^{1,2*} Jiahua Ma^{3*} Xincheng Yang^{2*}
Xin Wen^{3†} Yuzhi Zhang^{3†} Boyan Li^{1†} Yiran Qin^{4‡} Jin Liu^{1,2} Can Zhao¹
Li Kang^{1,5} Haoqin Hong⁶ Zhenfei Yin⁴ Philip Torr⁴ Hao Su⁷ Ruimao Zhang³ Daolin Ma^{1,2,✉}
¹Shanghai Jiao Tong University ²Xense Robotics ³Sun Yat-sen University ⁴Oxford
⁵Shanghai AI Laboratory ⁶University of Science and Technology of China ⁷UCSD
*Equal contribution [†]Equal contribution [‡]Project leader [✉]Corresponding author
martelzhang@gmail.com martelzhang.github.io/touchguide

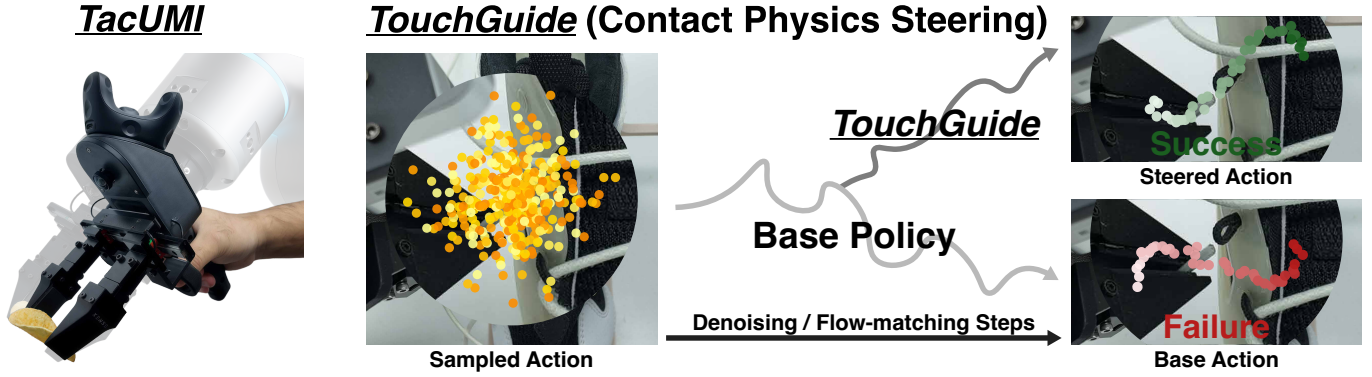


Fig. 1: **TacUMI** is a low-cost yet high-precision handheld data collection system that provides direct tactile feedback through a rigid mechanical coupling. **TouchGuide** is a multi-modal fusion paradigm that steers a visuomotor policy via touch guidance during denoising or flow matching, producing actions that better adhere to contact physics without retraining the base policy.

Abstract—Fine-grained and contact-rich manipulation remain challenging for robots, largely due to the underutilization of tactile feedback. To address this, we introduce TouchGuide, a novel cross-policy visuo-tactile fusion paradigm that fuses modalities within a low-dimensional action space. Specifically, TouchGuide operates in two stages to guide a pre-trained diffusion or flow-matching visuomotor policy at inference time. First, the policy produces a coarse, visually-plausible action using only visual inputs during early sampling. Second, a task-specific Contact Physical Model (CPM) provides touch guidance to steer and refine the action, ensuring it aligns with realistic physical contact conditions. Trained through contrastive learning on limited expert demonstrations, the CPM provides a tactile-informed feasibility score to steer the sampling process toward refined actions that satisfy physical contact constraints. Furthermore, to facilitate TouchGuide training with high-quality and cost-effective data, we introduce TacUMI, a data collection system. TacUMI achieves a favorable trade-off between precision and affordability; by leveraging rigid fingertips, it obtains direct tactile feedback, thereby enabling the collection of reliable tactile data. Extensive experiments on five challenging contact-rich tasks, such as shoe lacing and chip handover, show that TouchGuide consistently and significantly outperforms state-of-the-art visuo-tactile policies. Videos and other visualizations can be found on the project website: <https://martelzhang.github.io/touchguide>.

I. INTRODUCTION

Imitation learning [75, 69, 15, 44, 56] has enabled robots to perform complex manipulation tasks [47], such as folding clothes and preparing coffee [31, 45, 5, 3]. However, many intuitive fine-grained and contact-rich skills (e.g., shoe lacing or handing over fragile chips) remain challenging. These tasks demand a policy that effectively integration of visual and tactile information: visual information provides global scene context and task understanding, while tactile feedback supplies localized, contact-specific information essential for fine-grained control. This motivates a central question: *How should visual and tactile information be fused to improve policy planning in fine-grained and contact-rich manipulation?*

Existing visuo-tactile policy models generally follow one of two dominant fusion strategies: feature-level [65, 57, 13, 4] or policy-level [9, 8] fusion. While policy-level methods can mitigate the modality-dominance issue in feature-level concatenation caused by sparse features, they typically require training multiple single-modality policies, which makes it difficult to capture the cross-modal correlations that are critical for tasks demanding tight multimodal synergy.

To address this fundamental limitation, we introduce **TouchGuide**, a novel cross-policy visuo-tactile fusion paradigm. Departing from conventional computer vision ap-

proaches that focus on “fusing multi-modal data per se”, we rethink the problem from a robotic control perspective. Our **core idea** is to extract task-relevant semantics from each modality and perform fusion within the low-dimensional action space in a way that can benefit robotic manipulation. Concretely, TouchGuide steers a pre-trained visuomotor policy based on diffusion or flow matching (e.g., the action expert in $\pi_{0.5}$) during inference. The process begins by generating a visually feasible, coarse action from visual observation alone in the early sampling phase. Then, in the later phase, a task-specific Contact Physical Model (CPM) injects touch guidance, refining the action to better satisfy real-world physical contact constraints. The CPM is trained via contrastive learning on limited expert demonstrations to directly output a feasibility score. This score assesses whether the action complies with the contact physical constraints implied by the current tactile observations in task-specific environment, thereby generating a more robust steering signal. Notably, CPM can be applied across multiple policies, including VLAs [32, 5, 50, 58], diffusion policies [15, 69, 56, 44], and flow-matching policies [34, 71].

Moreover, in fine-grained and contact-rich manipulation, policy execution performance depends heavily on the quality of expert demonstrations. Teleoperation, due to high latency and indirect haptic feedback, hinders intuitive control. While handheld devices such as UMI [14] provide a platform-decoupled solution for data collection, existing systems relying on SLAM-based localization or motion capture (MoCap) cannot simultaneously achieve both high accuracy and low cost. To this end, we introduce **TacUMI**, designed according to the following principles: (1) Cost-effective (base cost \$720, tactile sensors are additional, using Xense sensors) and lightweight (540 g), while providing the high-precision data required for fine-grained manipulation. (2) Direct tactile feedback enabled by rigid fingertips. (3) A unified interface for synchronized acquisition of high-resolution visual and tactile signals.

Our main contributions are threefold:

- We propose **TouchGuide**, a novel cross-policy fusion paradigm for fine-grained and contact-rich manipulation.
- We develop **TacUMI**, a cost-effective yet high-precision data collection system with direct tactile feedback.
- Experiments on five challenging fine-grained and contact-rich tasks show that our **TouchGuide** can outperform other SOTA visuo-tactile methods and generalize across policies and tactile modalities effectively.

II. RELATED WORK

A. Hardware and Data Collection System

1) *Tactile Sensors*: Optical tactile sensors [68, 53, 37, 49, 41, 40, 36, 61] have advanced rapidly in recent years and have been widely adopted, owing to their ability to provide rich information (i.e., force and torque can be indirectly inferred from tactile images [74, 73]). Considering cost-effectiveness and ease of deployment, we selected the Xense tactile sensor [61]. The sensor offers software force estimation interfaces [62],

making it possible to evaluate our system algorithms across multiple tactile modalities.

2) *Data Collection System*: Teleoperation is a common choice for data collection, from ALOHA-style leader-follower systems [75, 46, 1, 10, 60, 25, 6] to peripheral-driven control such as 3D mice [19], VR controllers [65, 33], and hand tracking [48]. However, contact-rich and precise manipulation remains challenging due to missing tactile feedback; existing solutions mainly provide indirect haptic cues via VR visualization [65, 35, 20].

Some UMI-like [14, 76] handheld gripper data collection devices offer a data collection paradigm that operates independently of the physical robot [79, 43, 76, 63]. This type of handheld data-collection gripper, which operates without the physical robot, can provide haptic feedback directly, and many tactile data collection systems use this paradigm [64, 42, 79, 59, 67, 28, 39, 16]. ViTaMIn [42] and Touch in the Wild [79] adopt SLAM-based localization; however, the resulting accuracy is insufficient for many precise tasks. Some works (e.g., exUMI [64], ViTaMIn-B [39]) attach a VR Motion Tracker (e.g., VR controller of Meta Quest 3) to the gripper for localization, but this also introduces a notable drawback: it requires wearing a VR headset, which can impose a substantial burden during long-duration data collection. FreeTacMan [59] and FARM [28] use motion-capture systems, achieving high accuracy while remaining lightweight. However, they impose stricter requirements on the capture environment, and the high cost of motion-capture equipment makes large-scale data collection difficult.

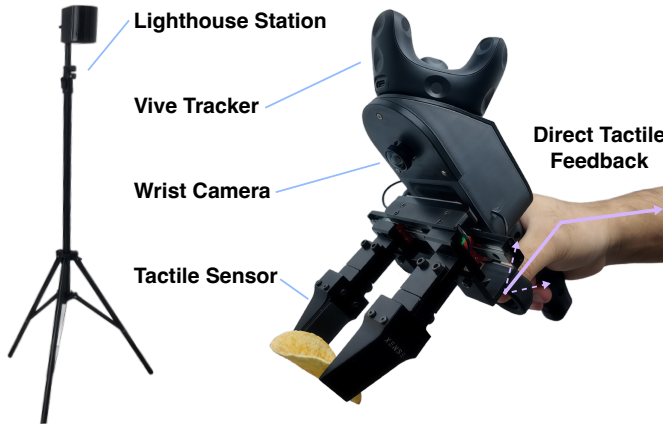
In contrast, our **TacUMI** uses Vive Trackers [54] for localization, reducing cost as much as possible while still meeting the accuracy requirements of precise manipulation tasks, thereby enabling large-scale data collection. Meanwhile, we minimize the linkage structure between the fingers and the gripper to obtain more direct haptic feedback, improving the quality of the tactile data.

B. Visuo-tactile Policies

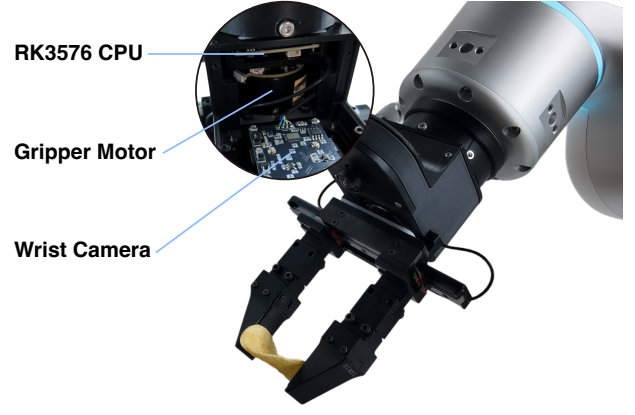
Visuo-tactile policies are crucial for precise manipulation, and prior works [29, 65, 70, 13, 27, 30, 4, 52, 21, 38, 12, 26, 66, 17, 9] have continuously explored this direction. Most of these policies (e.g., 3D-ViTac [29], VTLa [70], FreeTacMan [59]) rely on simple feature-level concatenation, which can cause the policy to overemphasize a single modality [9]. Policy-level composition methods such as PolicyConsensus [9] and stage-wise fusion approaches like RDP [65] alleviate this issue to some extent, but they do not perform multi-modal fusion in the action space. In contrast, our method, **TouchGuide**, introduces Contact Physical Model (CPM) to efficiently fuse modalities in the action space, steering the policy distribution toward the real distribution (see Sec. IV-C).

C. Inference-time Steering

Inspired by classifier guidance [18], some recent works [22, 51, 55, 72] have begun to adopt inference-time steering for robotic manipulation, which is an effective way to guide



(a) Data Collection with TacUMI



(b) Policy Execution on Flexiv with TacUMI

Fig. 2: Overview of **TacUMI** data collection system. (a) TacUMI (Collection-side) uses a Vive tracker for localization to obtain accurate end-effector poses, while the operator receives direct tactile feedback. (b) During policy inference, we use an execution-side device that is structurally identical to the collection-side TacUMI, coupled to different robot arms via an adapter.

actions toward a target distribution during inference. DynaGuide [22] and LPB [51] guide visuomotor policies via a dynamics model; however, this paradigm is ineffective for fine-grained manipulation, because future observations can become entirely different due to small errors in the action. In contrast, TouchGuide focuses on the current contact physics and can be integrated with a wide range of visuomotor policies (*e.g.*, DP [15] and $\pi_{0.5}$ [32]), yielding clear improvements over dynamics-based approaches (results are presented in Sec. V).

III. DATA COLLECTION SYSTEM: TACUMI

A. System Overview

As shown in Fig. 2, TacUMI is a UMI-like [14] handheld gripper that collects tactile data without relying on proprioception. Its development follows three key characteristics:

Accuracy and Low-cost. TacUMI is localized using a Vive tracker with two Lighthouse base stations, providing reliable and accurate pose estimates. In contrast, UMI [14] relies on SLAM-based tracking; for fine-grained tactile manipulation tasks which requires high precision, its accuracy is insufficient and tracking can be easily lost in feature-sparse scenes. Notably, a setup with one Vive tracker and two Lighthouse base stations has a base cost of only about **\$720**, maintaining low cost while providing accurate localization. A detailed comparison of data collection quality across different methods is provided in the experiments (see Sec. V-E).

Direct Tactile Feedback. Given that collecting tactile data for fine-grained manipulation critically depends on real-time tactile feedback, teleoperation presents a fundamental challenge by introducing a delay that impedes the operator’s ability to perceive and respond to touch signals promptly. Although some teleoperation methods, such as TactAR [65], provide visualized tactile feedback, this relies on human learning. After becoming accustomed to the interface, operators often ignore the visual tactile cues and instead act based on experience [65]. In contrast, this gripper-finger design enables direct tactile feedback, facilitating the collection of high-quality data.

Usability and Synchronization. TacUMI adopts a unified hardware-software interface. With a single aviation cable, it streams all observations in real time, including visual and tactile signals and proprioceptive states. Our data collection is compatible with the LeRobot [7] format and uses one software pipeline to acquire all modalities in a unified timestamped stream, enabling effective software-level synchronization. TacUMI captures visual $\mathbf{V}_t \in \mathbb{R}^{640 \times 480 \times 3}$ and tactile images $\mathbf{T}_t^{\text{image}} \in \mathbb{R}^{200 \times 350 \times 3}$ at 30Hz, sufficient for most fine-grained manipulation tasks. We also provide force observations $\mathbf{T}_t^{\text{force}} \in \mathbb{R}^{20 \times 35 \times 3}$ as a dense 3D force field; in our algorithm, it is handled the same as tactile images, except for the input dimensionality. An onboard RK3576 CPU runs the XenseSDK [62] to compute this force field online from the tactile images, without requiring external processing, as data are being captured. Since tactile sensing is modality-specific (*i.e.*, tactile images are not shared across sensors for training), this capability is crucial for scalable tactile data collection.

More details on hardware selection, design choices, and system parameters are provided in the Appendix E.

B. Data Processing for Training

We define $\{W\}$ as the world frame provided by Vive-tracker Lighthouse, $\{T\}$ as the Vive-tracker frame, $\{E\}$ as the end-effector (ee) frame and $\{B\}$ as the robot base frame. Before data collection, we align the robot base frame with the world frame, *i.e.*, the transformation matrix ${}^W T_B \in SE(3)$ should be obtained, and ${}^T T_E$ can be computed from the 3D model file. The data sequence collected by the Vive tracker after the coordinate transformation is $\{{}^B T_E^{(0)} \dots {}^B T_E^{(i)}\}$, *i.e.*, the end-effector pose in the robot base frame. Similar to FastUMI [76], the data sequence can be expressed as $\{\mathbf{p}_{ee}^{(i)}, \mathbf{R}_{ee}^{(i)}\}$, where $\mathbf{p}_{ee}^{(i)}$ and $\mathbf{R}_{ee}^{(i)}$ are the end-effector position and orientation in the robot base frame, respectively.

For policy training, we use relative ee pose $\{\mathbf{p}_{rel}^{(i)}, \mathbf{R}_{rel}^{(i)}\}$ as the dataset format, which is considered more effective than

using absolute ee pose [14]. This can be expressed as follows:

$$\mathbf{p}_{\text{rel}}^{(i)} = \mathbf{R}_{\text{ee}}^{(0)\top} (\mathbf{p}_{\text{ee}}^{(i)} - \mathbf{p}_{\text{ee}}^{(0)}), \quad (1)$$

$$\mathbf{R}_{\text{rel}}^{(i)} = \mathbf{R}_{\text{ee}}^{(0)\top} \mathbf{R}_{\text{ee}}^{(i)}, \quad (2)$$

where $\{\mathbf{p}_{\text{ee}}^{(0)}, \mathbf{R}_{\text{ee}}^{(0)}\}$ denotes the last observation state in an action chunk. Unlike UMI [14], we use the absolute ee pose $\{\mathbf{p}_{\text{ee}}^{(i)}, \mathbf{R}_{\text{ee}}^{(i)}\}$ as the robot state, with the goal of enabling the policy to perceive its pose in the real workspace. Moreover, since our model is designed to provide inference-time steering for both DP [15] and $\pi_{0.5}$ [32], and $\pi_{0.5}$ typically conditions on a single observation step, using relative state would make the state always the identity matrix, which would render proprioception ineffective.

IV. METHOD: TOUCHGUIDE

In this section, we introduce TouchGuide, a two-stage cross-policy multimodal fusion paradigm based on classifier guidance [18] (Sec. IV-A), inspired by recent advances in inference-time steering [51, 22, 55]. TouchGuide performs inference-time steering on a pre-trained diffusion-based or flow-matching-based (e.g., action expert of $\pi_{0.5}$) visuomotor policy $\pi_{\theta}(\mathbf{A}_t | \mathbf{V}_t)$. During the early phase of the denoising or flow-matching process, the policy generates an initial coarse action based solely on visual observation, ensuring visual feasibility without yet enforcing tactile constraints. In the later phase, a task-specific Contact Physical Model (CPM, detailed in Sec. IV-B) provides touch guidance via a feasibility score, which steers the action toward better satisfaction of real-world physical contact. The complete inference-time guidance process of TouchGuide is detailed in Sec. IV-C.

A. Preliminaries: Classifier Guidance

Diffusion Model. Following the notation in prior work [18], classifier guidance can be written as follows:

$$\hat{\epsilon}_{\theta}(x_t) = \epsilon_{\theta}(x_t) - \eta \sqrt{1 - \bar{\alpha}_t} \nabla_{x_t} \log p_{\phi}(y|x_t). \quad (3)$$

Flow Matching. Since prior work has not explicitly derived classifier guidance for flow matching [77, 23], we provide a brief introduction to the corresponding flow-matching formulation of classifier guidance.

Proposition 1. *The classifier guidance for flow matching can be formulated as follows:*

$$\hat{u}_{\theta}(x_t) = u_{\theta}(x_t) - \eta \frac{t}{1-t} \nabla_{x_t} \log p_{\phi}(y|x_t), \quad (4)$$

where $u_{\theta}(x_t)$ is the unconditional velocity field, ϕ parameterizes the auxiliary noise-dependent classifier, y is the target class label (e.g., additional condition), and η is the guidance scale. The term $t/(1-t)$ acts as a time-dependent weighting coefficient derived from the specific flow definition. See the detailed derivation in Appendix A.

Remark 1. *For VLAs, we can achieve tactile modality fusion by steering their action expert. Specifically, since the action expert of $\pi_{0.5}$ is flow-matching based, we can apply Eq. 4 to steer the action sampling process (see Eq. 12 in Sec. IV-C).*

B. Contact Physical Model

At time step t , Contact Physical Model (CPM, see Fig. 3(a)) outputs a task-specific feasibility score conditioned on the tactile feedback $\mathbf{T}_t$, the visual observation $\mathbf{V}_t$ and the coarse action $\mathbf{A}_t$ generated by the early phase of pre-trained policy $\pi_{\theta}(\mathbf{A}_t | \mathbf{V}_t)$. Specifically, we use pretrained DINOv2 as both the tactile and visual encoder to obtain the tactile and visual embeddings, i.e., $\mathbf{T}_t^{\text{emb}}$ and $\mathbf{V}_t^{\text{emb}}$ respectively, as it provides effective representations of robotic observations [22, 78] (see Appendix C for encoder comparison experiments). Then, we fuse these two embeddings using N-layer Transformer encoder $\text{TranEnc}(\cdot)_{\times N}$ and finally, yield the latent observation $\mathbf{O}_t$:

$$\mathbf{O}_t^{\text{emb}} = \text{TranEnc}(\text{concat}[\mathbf{T}_t^{\text{emb}}, \mathbf{V}_t^{\text{emb}}])_{\times N} \xrightarrow{\text{L2 Norm}} \mathbf{O}_t. \quad (5)$$

Similarly, we can obtain the latent action $\mathbf{a}_t$ by:

$$\mathbf{A}_t^{\text{emb}} = \mathcal{E}_{\mathbf{A}}(\mathbf{A}_t) \xrightarrow{\text{L2 Norm}} \mathbf{a}_t, \quad (6)$$

where $\mathcal{E}_{\mathbf{A}}$ is the action encoder (1D CNN + MLP) and $\mathbf{A}_t^{\text{emb}}$ is the action embedding. The feasibility score s represents the cosine similarity between the observation and the action in the latent space [11], and is defined as follows:

$$s = s_{\phi}(\mathbf{V}_t, \mathbf{T}_t, \mathbf{A}_t) = \mathbf{O}_t^{\top} \mathbf{a}_t. \quad (7)$$

We provide a derivation of our design for the feasibility score s , along with its physical interpretation (see Appendix G).

Training Strategy. We use the limited expert demonstrations to train our task-specific CPM. Given the visual-tactile observation (i.e., $\mathbf{V}_t$ and $\mathbf{T}_t$) and an ground-truth action $\mathbf{A}_t$, we can yield the latent observation $\mathbf{O}_t$ and latent action $\mathbf{a}_t$. Subsequently, we establish the positive samples using latent observation-action pairs from the same time step $\{(\mathbf{O}_t, \mathbf{a}_t)\}$, and negative samples using pairs from different time steps $\{(\mathbf{O}_{t+\Delta t}, \mathbf{a}_t), (\mathbf{O}_t, \mathbf{a}_{t+\Delta t})\}$. Finally, a standard contrastive learning loss $\mathcal{L}_{\text{CPM}}$ is adopted:

$$\mathcal{L}_{\text{CPM}}(\mathbf{O}_t, \mathbf{a}_t) = \frac{1}{2} (\mathcal{L}_{\mathbf{O} \rightarrow \mathbf{a}} + \mathcal{L}_{\mathbf{a} \rightarrow \mathbf{O}}), \quad (8)$$

$$\mathcal{L}_{\mathbf{O} \rightarrow \mathbf{a}} = -\frac{1}{M} \sum_{i=1}^M \log \frac{\exp(\mathbf{O}_i^{\top} \mathbf{a}_i / \tau)}{\sum_{j=1}^M \exp(\mathbf{O}_i^{\top} \mathbf{a}_j / \tau)}, \quad (9)$$

$$\mathcal{L}_{\mathbf{a} \rightarrow \mathbf{O}} = -\frac{1}{M} \sum_{i=1}^M \log \frac{\exp(\mathbf{a}_i^{\top} \mathbf{O}_i / \tau)}{\sum_{j=1}^M \exp(\mathbf{a}_i^{\top} \mathbf{O}_j / \tau)}, \quad (10)$$

where τ is a learnable temperature parameter and M denotes the number of samples used during training. In order to enable the CPM to effectively operate in a noisy action space, which is generated by the early phase of the pretrained policy during inference, we enhance the ground-truth action $\mathbf{A}_t$ by adding noise following a geometric distribution using the same noise scheduler as DP [22]; for flow-matching policies, we instead apply linear interpolation.

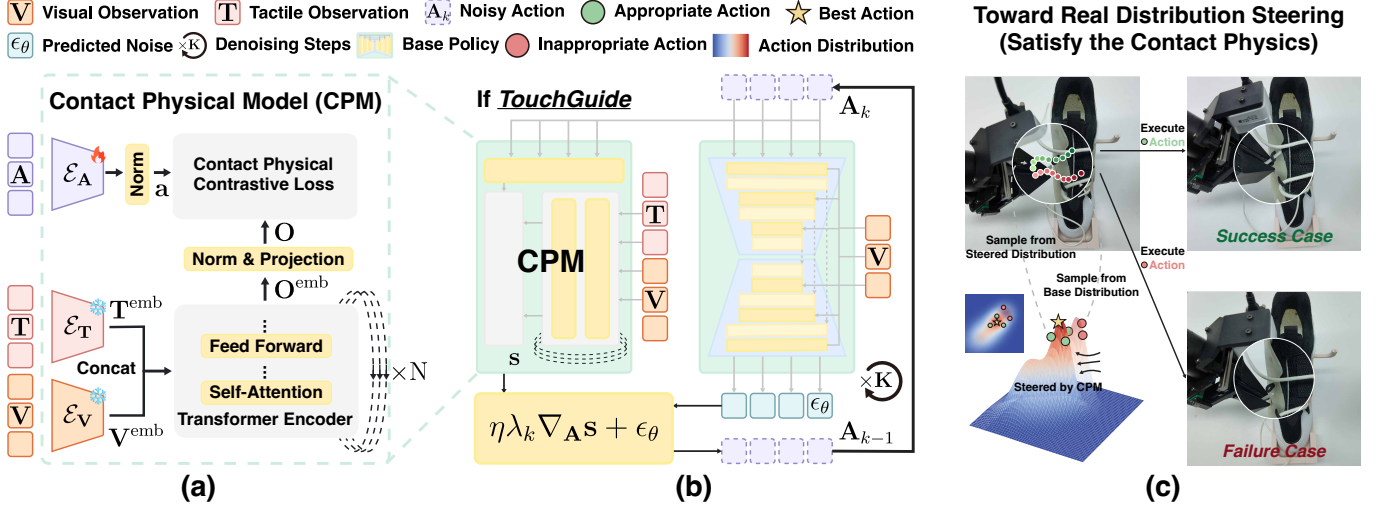


Fig. 3: Overview of **TouchGuide** framework. (a) The architecture of the task-specific Contact Physical Model (CPM). (b) During inference, the CPM serves as an external model that steers the base policy’s action generation within the sampling process using a feasibility score. (c) In action space, TouchGuide can be viewed as a form of contact-physics steering that steers the policy distribution toward the real distribution.

Algorithm 1 TouchGuide (Inference Time)

Require: Base policy π_θ , Contact Physical Model s_ϕ , total steps K , guidance steps $K_{\text{TouchGuide}}$, guidance scale η , visual observation V_t , tactile observation T_t

- 1: $A_t^K \leftarrow \text{Sample from } \mathcal{N}(0, I)$
- 2: **for** $k = K, \dots, 1$ **do**
- 3: **if** $\pi_\theta \in \text{Diffusion-based}$ **then** $\triangleright e.g., \text{DP, DP3}$
- 4: $\epsilon \leftarrow \pi_\theta(V)$
- 5: **if** $k \leq K_{\text{TouchGuide}}$ **then**
- 6: $s \leftarrow s_\phi(V_t, T_t, A_t^k)$
- 7: $\hat{\epsilon} \leftarrow \epsilon - \eta\sqrt{1 - \bar{\alpha}_k} \nabla_{A_t^k} s$
- 8: $A_t^{k-1} \leftarrow \text{Update } A_t^k \text{ using } \hat{\epsilon}$
- 9: **else**
- 10: $A_t^{k-1} \leftarrow \text{Update } A_t^k \text{ using } \epsilon$
- 11: **end if**
- 12: **else if** $\pi_\theta \in \text{Flow-matching}$ **then** $\triangleright e.g., \pi_{0.5}$
- 13: $k \leftarrow \text{Scale } [K, 0] \text{ to } [1, 0]$
- 14: $u \leftarrow \pi_\theta(V_t)$
- 15: **if** $k \leq K_{\text{TouchGuide}}$ **then**
- 16: $s \leftarrow s_\phi(V_t, T_t, A_t^k)$
- 17: $\hat{u} \leftarrow u - \eta\frac{k}{1-k} \nabla_{A_t^k} s$
- 18: $A_t^{k-1} \leftarrow \text{Update } A_t^k \text{ using } \hat{u}$
- 19: **else**
- 20: $A_t^{k-1} \leftarrow \text{Update } A_t^k \text{ using } u$
- 21: **end if**
- 22: **end if**
- 23: **end for**
- 24: **return** A_t^0

C. Steering the Policies via Touch Guidance

Fig. 3(b) illustrates the entire TouchGuide pipeline, where touch guidance is used to steer the sampling process of

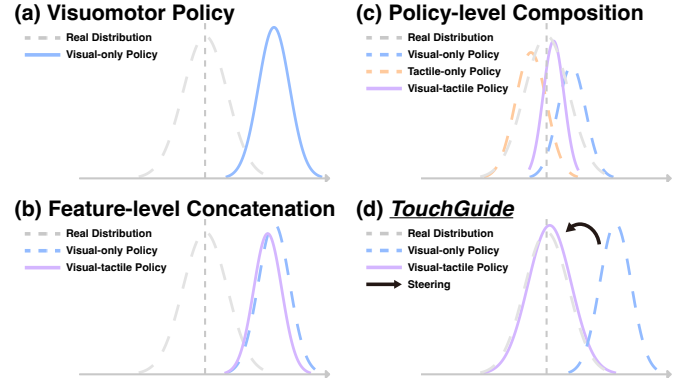


Fig. 4: Comparison of policy distributions in action space.

base policy to generate actions that better adhere to the contact physics. Following [22, 51], both diffusion models and flow-matching models can be guided by the gradients $\nabla_{x_t} \log p_\phi(y|x_t)$ provided by a classifier toward a particular set of samples that maximizes $p_\phi(y|x_t)$ (see Eq. 3, 4) [18, 77, 23]. In our case, $\log p_\phi(y|x_t)$ corresponds to the feasibility score s produced by the CPM. With this formulation, touch guidance for flow-matching (*e.g.*, action expert of $\pi_{0.5}$) and diffusion policies can be respectively expressed by the following two equations:

$$\hat{\epsilon}_\theta = \epsilon_\theta(A_t^k, V_t) - \eta\sqrt{1 - \bar{\alpha}_k} \nabla_{A_t^k} s_\phi(V_t, T_t, A_t^k), \quad (11)$$

$$\hat{u}_\theta = u_\theta(A_t^k, V_t) - \eta\frac{k}{1-k} \nabla_{A_t^k} s_\phi(V_t, T_t, A_t^k), \quad (12)$$

where η is the guidance scale, $\epsilon_\theta(\cdot)$ and $u_\theta(\cdot)$ are the noise predictor in diffusion policy and velocity predictor in flow-matching policy, respectively. The detailed **TouchGuide** procedure is provided in Alg. 1.

As shown in Fig. 3(c), our TouchGuide can be viewed as using the CPM to steer the policy distribution $\mathcal{Q}_{\text{policy}}$ toward the real distribution $\mathcal{Q}_{\text{real}}$, which refers to the optimal distribution to accomplish the task in the real world. Therefore, sampling actions from the CPM-steered $\mathcal{Q}_{\text{policy}}$ is more likely to produce actions that better adhere to the contact physics. In Fig. 4, we illustrate the policy distributions of TouchGuide and previous methods (including feature-level concatenation [65, 57, 13, 4] and policy-level composition [8, 9]) in the action space. For feature-level concatenation, although tactile sensing provides additional information, the model may still overly attend to the visual modality (since tactile signals often exhibit limited variation in many cases), and thus its policy distribution $\mathcal{Q}_{\text{policy}}$ cannot be explicitly driven toward the real distribution $\mathcal{Q}_{\text{real}}$. Policy-level composition, though an improvement over feature-level methods, still falls short in enabling effective multi-modal fusion within the action space $\mathcal{A}$. Its distribution operates multiplicatively in $\mathcal{A}$, and it lacks an explicit mechanism to steer the $\mathcal{Q}_{\text{policy}}$ toward $\mathcal{Q}_{\text{real}}$. To address these limitations, we introduce a CPM. Our method assesses the physical feasibility of both tactile observations and noisy actions, thereby providing explicit guidance for steering in $\mathcal{A}$. Further physical interpretations of how CPM steers the action distribution are provided in the Appendix G.

V. EXPERIMENTS

In this section, we evaluate the proposed TouchGuide framework and the TacUMI data-collection system across five challenging tasks. Specifically, our experiments are designed to answer the following questions: **Q1.** How does tactile feedback improve fine-grained robotic manipulation? **Q2.** How does TouchGuide perform compared to prior visuo-tactile policies? **Q3.** Does TouchGuide generalize across different robots, tactile sensors, and policies? **Q4.** Is the TouchGuide fusion framework robust and generalizable under visual domain shifts and occlusions? **Q5.** Can TacUMI efficiently provide high-quality data for precise contact-rich tasks?

A. Experimental Setup

1) *Hardware:* For dual-arm tasks, we employ the Bi-ARX5 [2], while the Flexiv Rizon4 [24] integrated with TacUMI is used for single-arm tasks. Given the high symmetry of the gripper’s tactile sensing, we attach tactile sensors only to the outer finger of each arm during dual-arm experiments. This design reduces dataset size and enhances training efficiency.

2) *Tasks:* As shown in Fig. 5, we design five challenging contact-rich tasks. To thoroughly evaluate the performance of our paradigm, these tasks include long-horizon, precise manipulation (e.g., Shoe Lacing), collaborative handover of fragile objects (e.g., Chip Handover), and low-data (only 20 demos), high-precision manipulation (e.g., Lock Opening). More detailed task descriptions and examples of the evaluation metrics are provided in Appendix K.

Shoe Lacing. The robot needs to grasp the shoelace and sequentially thread it through four eyelets (two on each side).

The task is considered successful if the robot threads the shoelace through all four eyelets within the allotted time.

Chip Handover. The robot needs to grasp a potato chip, execute a collaborative in-air handover, and finally place it into a designated container. The task is considered successful if, within the allotted time, the chip is placed in the container without any breaking.

Cucumber Peeling. The robot must grasp both a peeler and a cucumber, after which the two robotic arms collaborate to peel the cucumber. Task score is assessed based on the ratio of the length of peel removed by the robot to that removed by a human expert. The resulting score is quantized into one of seven discrete values: 0, 0.1, 0.3, 0.5, 0.7, 0.9, or 1.

Vase Wiping. The robot needs to pick up a blackboard eraser from the table, approach the vase held by a human, and erase the marker traces on its surface. The task score is defined similarly to Cucumber Peeling.

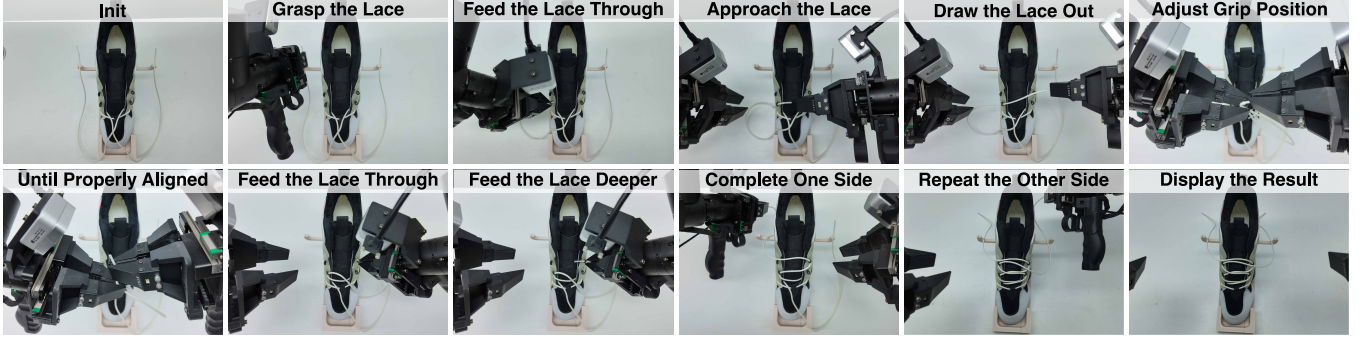
Lock Opening. The robot executes a sequential manipulation task involving a human partner: grasping a key held by the human, inserting it into a lock also held by the human, and finally turning the key to open the lock. The task is considered successful if the robot maintains hold of the key without dropping it and opens the lock within the allotted time.

3) *Baselines:* We adopt two state-of-the-art (SOTA) visuomotor policies (i.e., DP [15] and $\pi_{0.5}$ [32]) as the base policy for TouchGuide, and use several visuo-tactile policies (e.g., RDP [65], Policy Consensus [9] and SafeDiff [57]) as the baselines for comparison. Specifically, we include Tactile Dynamics as one of our baselines. This method is derived from Latent Policy Barrier [51] and employs a dynamics model to jointly predict tactile and visual observations. Details on the touch guidance setup of Tactile Dynamics for $\pi_{0.5}$ can be found in Sec. IV-A. For TouchGuide, we use two tactile modalities (i.e., force and tactile image) to thoroughly validate the cross-modality capability of our paradigm. The specific baseline configurations are detailed in Appendix J. For these five tasks, each baseline is evaluated over 20 trials with moderate randomization of the initial environment conditions, and the results are reported in Table I.

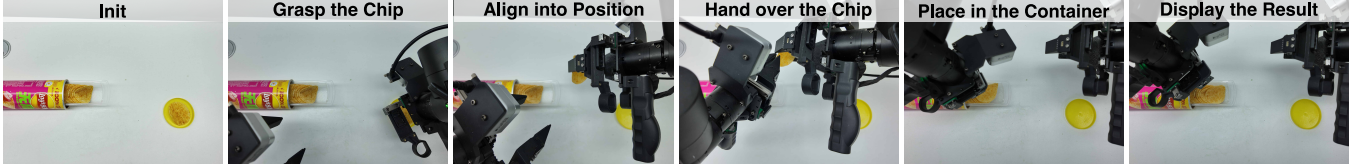
B. Main Results and Analysis

Tactile sensing provides rich and accurate in-hand state information, which is crucial for fine-grained manipulation (Q1). Tactile sensing not only provides rich force information but also yields accurate in-hand state information. For fine-grained manipulation tasks such as shoe lacing, chip handover, and lock opening, this in-hand state information is particularly critical, as it determines the end-effector pose required for threading through an eyelet or executing a reliable handover. In contrast, such in-hand states are difficult to infer accurately from vision alone and are often only weakly observable through specular highlights. Table I shows that TouchGuide yields clear performance gains over DP and $\pi_{0.5}$ (16.3% $\rightarrow$ 36.2%, 35.9% $\rightarrow$ 58.0%). This highlights the importance of accurate in-hand state estimation for precise manipulation.

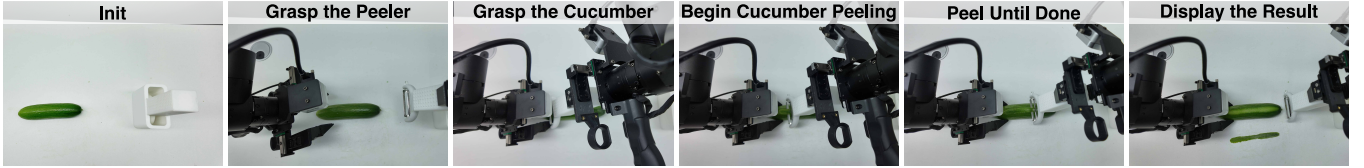
Task1: Shoe Lacing



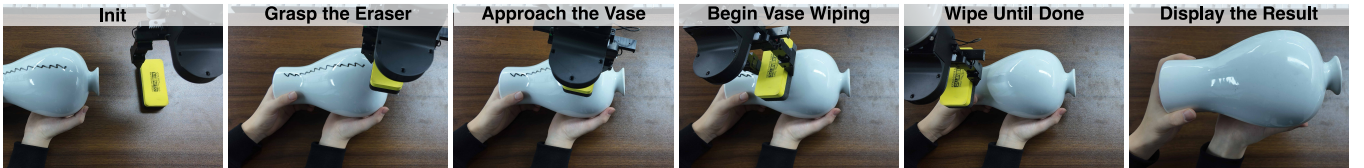
Task2: Chip Handover



Task3: Cucumber Peeling



Task4: Vase Wiping



Task5: Lock Opening

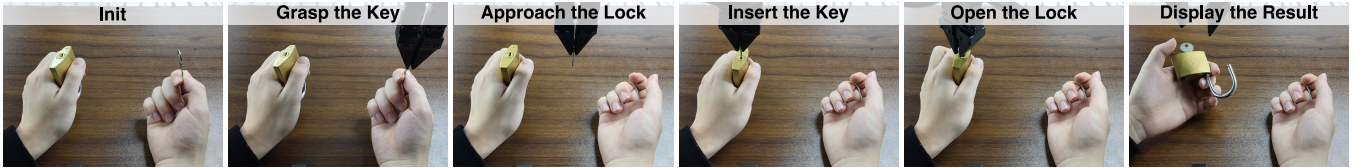


Fig. 5: Five experiment tasks including Shoe Lacing, Chip Handover, Cucumber Peeling, Vase Wiping, and Lock Opening.

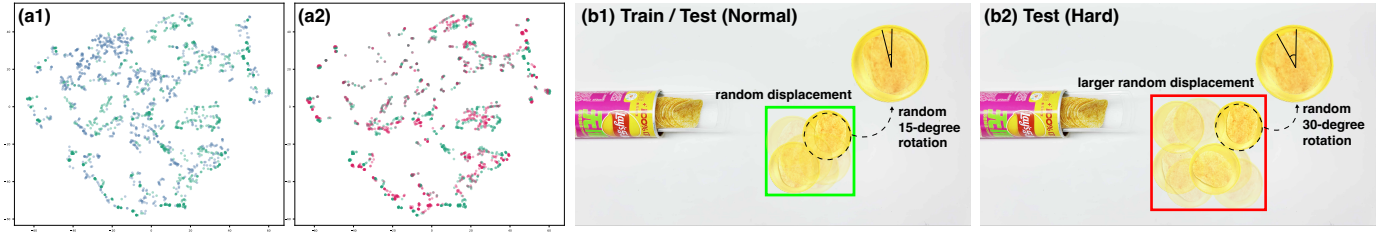
TouchGuide explicitly steers the base policy in action space via touch guidance, yielding performance gains over current SOTA visuo-tactile policies (Q2). Existing methods mainly target rapid reactions to tactile changes (*e.g.*, RDP [65] and PolicyConsensus [9]); however, RDP relies on low-dim tactile embeddings that discard in-hand state cues, hindering fine-grained manipulation. In many tasks, initial contact is decisive: for Lock Opening and Shoe Lacing, small errors at first contact can drop the key or severely misalign it, pushing the system into out-of-distribution states that are especially problematic for imitation learning. In contrast, our method uses a CPM to assess the feasibility of noisy actions proposed by the base policy and steers the policy at inference time using the resulting feasibility scores. Evaluating physical consistency between actions and tactile signals is crucial for producing appropriate contacts. Table I supports this claim:

TouchGuide (36.2%) improves over other SOTA visuo-tactile policies, including RDP [65] (30.3%) and PolicyConsensus [9] (24.7%). Meanwhile, TouchGuide adds negligible inference time overhead (see Appendix F3).

TouchGuide demonstrates strong generalization, and can be applied across different robots, tactile modalities, and policies (Q3). Unlike prior visuo-tactile policies, TouchGuide augments a visuomotor policy with touch guidance in a fully decoupled manner, requiring no retraining to the base policy. As a result, it can inherit the strong capability of the base policy. For example, TouchGuide built on $\pi_{0.5}$ achieves solid performance on a high-precision task such as Lock Opening with only **20 demonstrations**. Moreover, TouchGuide achieves strong performance across different robots (*e.g.*, Bi-ARX5 and Flexiv Rizon4), different policies (*e.g.*, DP and $\pi_{0.5}$), and different tactile modalities (*e.g.*, force

TABLE I: Policy performance comparison on five challenging fine-grained, contact-rich tasks.

		Bi-Arx5 (Dual-arm)			Flexiv Rizon4 (Single-arm)		Average
		Shoe Lacing (100 Demos)	Chip Handover (50 Demos)	Cucum. Peeling (50 Demos)	Vase Wiping (30 Demos)	Lock Opening (20 Demos)	
DP	Diffusion Policy	0%	5%	0.500	0.265	0%	16.3%
	DP w/ Tactile Observation	0%	15%	0.520	0.240	5%	19.2%
	SafeDiff	0%	10%	0.635	0.325	5%	22.2%
	RDP	0%	20%	0.740	0.475	10%	30.3%
	Policy Consensus	0%	15%	0.630	0.405	5%	24.7%
	Tactile Dynamics	0%	15%	0.695	0.335	0%	23.6%
	TouchGuide (Force)	0%	30%	0.805	0.510	15%	35.3%
	TouchGuide (Tactile Img.)	0%	25%	0.810	0.550	20%	36.2%
$\pi_{0.5}$	$\pi_{0.5}$	20%	25%	0.785	0.360	20%	35.9%
	$\pi_{0.5}$ w/ Tactile Observation	20%	15%	0.760	0.305	10%	30.3%
	Tactile Dynamics	10%	30%	0.815	0.395	20%	36.2%
	TouchGuide (Force)	25%	40%	0.955	0.590	30%	49.9%
	TouchGuide (Tactile Img.)	35%	60%	0.975	0.675	30%	58.0%


 Fig. 6: Action distribution visualization and evaluation setups on Chip Handover. (a1,a2) t-SNE of $\mathbf{a}_{\text{base}}$, $\mathbf{a}_{\text{real}}$, $\mathbf{a}_{\text{steered}}$. (b1) Training / Normal setup. (b2) Hard setup with larger displacement and rotation.

and tactile image). Detailed results are reported in Table I. More baseline failure cases are provided in Appendix D.

C. Ablation Studies

1) *Noise Pretraining Ablation*: Since the task-specific CPM takes noisy actions as input at inference time, we add noise into the action using the same noise scheduler for diffusion models, or via linear interpolation for flow-matching models. The number of noising steps is sampled from a geometric distribution. As shown in Table II, this key setting can effectively improve the task success rate for TouchGuide (39.17% $\rightarrow$ 62.50%).

TABLE II: Ablation study on noise pretraining.

	Chip Handover	Cucumber Peeling	Lock Opening	Average
Ours w/o Noise Pretraining	30%	0.725	15%	39.17%
Ours w/ Noise Pretraining	60%	0.975	30%	62.50%

2) *CPM Modality Ablation*: For most fine-grained manipulation tasks, tactile and visual signals are sparse at different time periods. For example, in the Cucumber Peeling task, tactile feedback is sparse when the robot holds the peeler and moves toward the cucumber. During peeling, however, small motion variations lead to pronounced changes in tactile feedback, while visual information becomes relatively sparse. Therefore, we argue that for task-specific CPM, the visual and

tactile modalities are equally important (see Table III). The results show without vision, the success rate drops from 62.50% to 43.33%, and without touch, it decreases to 43.50%.

TABLE III: Ablation study on CPM input modalities.

	Chip Handover	Cucumber Peeling	Lock Opening	Average
Ours w/o Vision	30%	0.850	15%	43.33%
Ours w/o Touch	35%	0.755	20%	43.50%
Ours w/ Vision & Touch	60%	0.975	30%	62.50%

Also, we provide ablation studies on the guidance scale η and guidance steps $K_{\text{TouchGuide}}$ in Alg. 1 (see Appendix B).

D. Generalization and Robustness

TouchGuide is robust under visual domain shifts and occlusions (Q4). As shown in Fig. 7, TouchGuide consistently improves the base policy across unseen objects, unseen scenes, and their combination (Fig. 8(a)), indicating that touch guidance compensates for visual distribution shifts by providing complementary contact information. Table IV further demonstrates robustness to transient visual occlusion (Fig. 8(b)). Fig. 6(a1,a2) confirms that $\mathbf{a}_{\text{steered}}$ aligns more closely with $\mathbf{a}_{\text{real}}$ than $\mathbf{a}_{\text{base}}$ (Additional details are provided in Appendix I).

E. System Comparison and User Study

TacUMI provides accurate and reliable data that is well suited for robot learning (Q5). We find that although VR

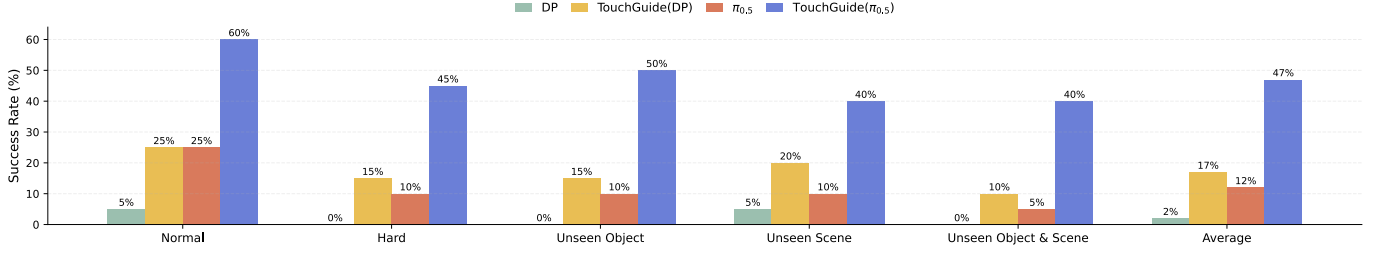


Fig. 7: Generalization evaluation on Chip Handover across Normal, Hard, Unseen Object, Unseen Scene, and Unseen Object & Scene conditions. TouchGuide consistently and substantially improves the base policy under all settings.

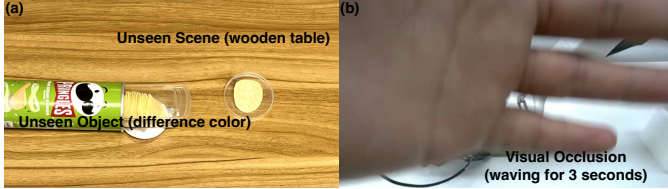


Fig. 8: Additional experimental settings. (a) Unseen Object and Unseen Scene setups. (b) Visual Occlusion setup.

TABLE IV: Robustness evaluation under visual occlusion in the Cucumber Peeling task.

	None	Before	During	After	Average
TouchGuide($\pi_{0.5}$)	0.975	0.955	0.910	0.960	0.950

TABLE V: Success rate comparison of data collection systems on the Lock Opening task.

Method	TacUMI	UMI (SLAM-based)	VR Teleop
$\pi_{0.5}$	20%	0%	5%
TouchGuide	30%	5%	15%

teleoperation provides highly precise action (since it is read directly from the robot arm), teleoperating tasks, such as Lock Opening, are challenging for human demonstrators. Hesitation around the keyhole often degrades the data quality, making it unsuitable for robot learning. This issue is particularly pronounced in the low-data regime (20 demos), where such low-quality demonstrations can even hinder learning. In addition, SLAM-based UMI [14] struggles to meet the requirements of fine-grained manipulation tasks due to the limited accuracy of its SLAM module. In order to evaluate the effectiveness of our TacUMI, we trained policies using the data collected by all involved data collection systems and evaluated them on the Lock Opening task. Table V indicates that our TacUMI outperforms the other two data collection systems.

To evaluate the usability of our TacUMI, we recruited five volunteers working in robotics (two had prior experience collecting demonstrations using a handheld gripper, while three had no such experience). Each volunteer collected 20 valid demonstrations using VR Teleop and UMI with three different tracking methods (*e.g.*, SLAM-based, VR Motion Tracker, and **TacUMI**) on Lock Opening task. We used the following metrics: **Attempts** (the number of trials required to obtain the target number of demonstrations), **Length** (the dataset duration excluding preparation and post-processing time, which helps

assess system usability), **Valid Rate** (Valid Data / Attempts), and **Satisfaction** (each volunteer’s subjective usability rating of the system). As shown in Table VI, although Teleop achieves a high valid rate, the tactile feedback is not directly available, which leads to excessively long demonstrations and may degrade policy performance. SLAM-based tracking yields a very low valid rate. The VR Motion Tracker requires wearing a VR headset, making it unsuitable for prolonged data collection, and occlusions can easily cause tracking loss if the user is not constantly attentive during collection. **TacUMI** demonstrates excellent usability and receives high satisfaction ratings from the volunteers. More comparisons with existing data collection systems are provided in Appendix E1.

TABLE VI: User study on data collection systems.

Method	Attempts ↓	Length ↓	Valid Rate ↑	Satisfaction ↑
VR Teleop	100	23.0 min	100.00%	7.8
SLAM-based	143	15.3 min	69.93%	7.2
VR Motion Tracker	104	14.7 min	96.15%	9.0
TacUMI (Ours)	100	13.5 min	100.00%	9.6

VI. LIMITATIONS AND FUTURE WORK

Although our TouchGuide paradigm enables tactile fusion without retraining the base policy, the CPM still requires task-specific training, which inevitably increases training time and cost. An important direction for future work is to learn more effective tactile representations and develop a non-task-specific CPM that can generalize across tasks without retraining.

VII. CONCLUSION

In this paper, we propose a novel two-stage cross-policy multi-modal fusion paradigm (*i.e.*, **TouchGuide**), and an efficient data collection system (*i.e.*, **TacUMI**). TouchGuide steers the base policy via touch guidance to generate actions that better adhere to the contact physics, demonstrating strong performance on five challenging fine-grained manipulation tasks. TacUMI strikes a trade-off between cost and precision, minimizing overhead while maintaining high accuracy; moreover, its direct tactile feedback improves usability and helps operators collect high-quality tactile data more efficiently.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grants No. 12522202, No. 12272220 and No. 124B1038, and in part by Xense Robotics.

REFERENCES

- [1] Jorge Aldaco, Travis Armstrong, Robert Baruch, Jeff Bingham, Sanky Chan, Kenneth Draper, Debiddatta Dwibedi, Chelsea Finn, Pete Florence, Spencer Goodrich, et al. Aloha 2: An enhanced low-cost hardware for bimanual teleoperation. *arXiv preprint arXiv:2405.02292*, 2024. 2
- [2] ARX5. <https://www.arx-x.com/?product/>, 2025. 6
- [3] Oriol Barbany, Adrià Colomé, and Carme Torras. Bifold: Bimanual cloth folding with language guidance. *arXiv preprint arXiv:2501.16458*, 2025. 1
- [4] Jianxin Bi, Kevin Yuchen Ma, Ce Hao, Mike Zheng Shou, and Harold Soh. Vla-touch: Enhancing vision-language-action models with dual-level tactile feedback. *arXiv preprint arXiv:2507.17294*, 2025. 1, 2, 6
- [5] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0 : A vision-language-action flow model for general robot control, 2024. URL <https://arxiv.org/abs/2410.24164>. 1, 2
- [6] Thanpimon Buamanee, Masato Kobayashi, Yuki Urinishi, and Haruo Takemura. Bi-act: Bilateral control-based imitation learning via action chunking with transformer. In *2024 IEEE International Conference on Advanced Intelligent Mechatronics (AIM)*, pages 410–415. IEEE, 2024. 2, 19, 20
- [7] Remi Cadene, Simon Alibert, Alexander Soare, Quentin Gallouedec, Adil Zouitine, Steven Palma, Pepijn Kooijmans, Michel Aractingi, Mustafa Shukor, Dana Aubakirova, Martino Russi, Francesco Capuano, Caroline Pascal, Jade Choghari, Jess Moss, and Thomas Wolf. Lerobot: State-of-the-art machine learning for real-world robotics in pytorch. <https://github.com/huggingface/lerobot>, 2024. 3
- [8] Jiahang Cao, Yize Huang, Hanzhong Guo, Rui Zhang, Mu Nan, Weijian Mai, Jiaxu Wang, Hao Cheng, Jingkai Sun, Gang Han, et al. Compose your policies! improving diffusion-based or flow-based robot policies via test-time distribution-level composition. *arXiv preprint arXiv:2510.01068*, 2025. 1, 6
- [9] Haonan Chen, Jiaming Xu, Hongyu Chen, Kaiwen Hong, Binghao Huang, Chaoqi Liu, Jiayuan Mao, Yunzhu Li, Yilun Du, and Katherine Driggs-Campbell. Multi-modal manipulation via multi-modal policy consensus. *arXiv preprint arXiv:2509.23468*, 2025. 1, 2, 6, 7, 18, 27
- [10] Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, et al. Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088*, 2025. 2
- [11] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020. 4
- [12] Yizhou Chen, Andrea Sipos, Mark Van der Merwe, and Nima Fazeli. Visuo-tactile transformers for manipulation. *arXiv preprint arXiv:2210.00121*, 2022. 2
- [13] Zhengxue Cheng, Yiqian Zhang, Wenkang Zhang, Haoyu Li, Keyu Wang, Li Song, and Hengdi Zhang. Omnivtla: Vision-tactile-language-action model with semantic-aligned tactile sensing. *arXiv preprint arXiv:2508.08706*, 2025. 1, 2, 6
- [14] 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. 2, 3, 4, 9, 19, 21
- [15] 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, 3, 4, 6, 19, 27
- [16] Hojung Choi, Yifan Hou, Chuer Pan, Seongheon Hong, Austin Patel, Xiaomeng Xu, Mark R Cutkosky, and Shuran Song. In-the-wild compliant manipulation with umi-ft. *arXiv preprint arXiv:2601.09988*, 2026. 2, 19, 21
- [17] 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. IEEE, 2024. 2
- [18] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2, 4, 5, 15
- [19] Varad Dhat, Nick Walker, and Maya Cakmak. Using 3d mice to control robot manipulators. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, pages 896–900, 2024. 2
- [20] Runyu Ding, Yuzhe Qin, Jiyue Zhu, Chengzhe Jia, Shiqi Yang, Ruihan Yang, Xiaojuan Qi, and Xiaolong Wang. Bunny-visionpro: Real-time bimanual dexterous teleoperation for imitation learning. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 12248–12255. IEEE, 2025. 2, 19, 20
- [21] Zihao Ding, Guodong Chen, Zhenhua Wang, and Lining Sun. Adaptive visual-tactile fusion recognition for robotic operation of multi-material system. *Frontiers in Neurobotics*, 17:1181383, 2023. 2
- [22] Maximilian Du and Shuran Song. Dynaguide: Steering diffusion policies with active dynamic guidance. *arXiv preprint arXiv:2506.13922*, 2025. 2, 3, 4, 5, 23, 27

- [23] Ruiqi Feng, Chenglei Yu, Wenhao Deng, Peiyan Hu, and Tailin Wu. On the guidance of flow matching. *arXiv preprint arXiv:2502.02150*, 2025. 4, 5, 15
- [24] Flexiv Rizon4. <https://www.flexiv.com/products/rizon>, 2025. 6
- [25] 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. 2
- [26] Abraham George, Selam Gano, Pranav Katragadda, and Amir Barati Farimani. Vital pretraining: Visuo-tactile pretraining for tactile and non-tactile manipulation policies. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 258–264. IEEE, 2025. 2
- [27] Peng Hao, Chaofan Zhang, Dingzhe Li, Xiaoge Cao, Xiaoshuai Hao, Shaowei Cui, and Shuo Wang. Tla: Tactile-language-action model for contact-rich manipulation. *arXiv preprint arXiv:2503.08548*, 2025. 2
- [28] Erik Helmut, Niklas Funk, Tim Schneider, Cristiana de Farias, and Jan Peters. Tactile-conditioned diffusion policy for force-aware robotic manipulation. *arXiv preprint arXiv:2510.13324*, 2025. 2, 19, 21
- [29] 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. 2
- [30] Jialei Huang, Shuo Wang, Fanqi Lin, Yihang Hu, Chuan Wen, and Yang Gao. Tactile-vla: unlocking vision-language-action model’s physical knowledge for tactile generalization. *arXiv preprint arXiv:2507.09160*, 2025. 2
- [31] Physical Intelligence, Ali Amin, Raichelle Aniceto, Ashwin Balakrishna, Kevin Black, Ken Conley, Grace Connors, James Darpinian, Karan Dhabalia, Jared DiCarlo, et al. $\pi_{0.6}^*$: a vla that learns from experience. *arXiv preprint arXiv:2511.14759*, 2025. 1
- [32] Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. $\pi_{0.5}$: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025. 2, 3, 4, 6, 16, 19, 27
- [33] Aadithya Iyer, Zhuoran Peng, Yinlong Dai, Irmak Guzey, Siddhant Haldar, Soumith Chintala, and Lerrel Pinto. Open teach: A versatile teleoperation system for robotic manipulation. *arXiv preprint arXiv:2403.07870*, 2024. 2
- [34] Sunshine Jiang, Xiaolin Fang, Nicholas Roy, Tomás Lozano-Pérez, Leslie Pack Kaelbling, and Siddharth Ancha. Streaming flow policy: Simplifying diffusion / flow-matching policies by treating action trajectories as flow trajectories. *arXiv preprint arXiv:2505.21851*, 2025. 2
- [35] Tatsuya Kamijo, Cristian C Beltran-Hernandez, and Masashi Hamaya. Learning variable compliance control from a few demonstrations for bimanual robot with haptic feedback teleoperation system. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 12663–12670. IEEE, 2024. 2
- [36] Naveen Kuppaswamy, 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. 2
- [37] Mike Lambeta, Po-Wei Chou, Stephen Tian, Brian Yang, Benjamin Maloon, Victoria Rose Most, Dave Stroud, Raymond Santos, Ahmad Byagowi, Gregg Kammerer, et al. Digit: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation. *IEEE Robotics and Automation Letters*, 5(3):3838–3845, 2020. 2
- [38] 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. 2
- [39] Chuanyu Li, Chaoyi Liu, Daotan Wang, Shuyu Zhang, Lusong Li, Zecui Zeng, Fangchen Liu, Jing Xu, and Rui Chen. Vitamin-b: A reliable and efficient visuo-tactile bimanual manipulation interface. *arXiv preprint arXiv:2511.05858*, 2025. 2, 19, 21
- [40] Changyi Lin, Ziqi Lin, Shaoxiong Wang, and Huazhe Xu. Dtact: A vision-based tactile sensor that measures high-resolution 3d geometry directly from darkness. *arXiv preprint arXiv:2209.13916*, 2022. 2
- [41] 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*, 9(2):923–930, 2023. 2
- [42] Fangchen Liu, Chuanyu Li, Yihua Qin, 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. 2, 19, 21
- [43] Kehui Liu, Zhongjie Jia, Yang Li, Pengan Chen, Song Liu, Xin Liu, Pingrui Zhang, Haoming Song, Xinyi Ye, Nieqing Cao, et al. Fastumi-100k: Advancing data-driven robotic manipulation with a large-scale umi-style dataset. *arXiv preprint arXiv:2510.08022*, 2025. 2
- [44] Jiahua Ma, Yiran Qin, Yixiong Li, Xuanqi Liao, Yulan Guo, and Ruimao Zhang. Cdp: Towards robust autoregressive visuomotor policy learning via causal diffusion. *arXiv preprint arXiv:2506.14769*, 2025. 1, 2
- [45] HKU MMLab. A live-stream robotic teamwork for clothing manipulation from zero to hero. *HKU MMLab Research Blog*, 2025. <https://mmlab.hk/research/kai0>. 1, 27
- [46] Yao Mu, Tianxing Chen, Shijia Peng, Zhanxin Chen, Zeyu Gao, Yude Zou, Lunkai Lin, Zhiqiang Xie, and Ping Luo.

- Robotwin: Dual-arm robot benchmark with generative digital twins (early version). In *European Conference on Computer Vision*, pages 264–273. Springer, 2024. 2
- [47] Yiran Qin, Li Kang, Xiufeng Song, Zhenfei Yin, Xiaohong Liu, Xihui Liu, Ruimao Zhang, and Lei Bai. Robofactory: Exploring embodied agent collaboration with compositional constraints. *arXiv preprint arXiv:2503.16408*, 2025. 1, 27
- [48] Yuzhe Qin, Wei Yang, Binghao Huang, Karl Van Wyk, Hao Su, Xiaolong Wang, Yu-Wei Chao, and Dieter Fox. Anyteleop: A general vision-based dexterous robot arm-hand teleoperation system. *arXiv preprint arXiv:2307.04577*, 2023. 2
- [49] Jieji Ren, Jiang Zou, and Guoying Gu. Mc-tac: Modular camera-based tactile sensor for robot gripper. In *International Conference on Intelligent Robotics and Applications*, pages 169–179. Springer, 2023. 2
- [50] Mustafa Shukor, Dana Aubakirova, Francesco Capuano, Pepijn Kooijmans, Steven Palma, Adil Zouitine, Michel Aractingi, Caroline Pascal, Martino Russi, Andres Marafioti, et al. Smolvla: A vision-language-action model for affordable and efficient robotics. *arXiv preprint arXiv:2506.01844*, 2025. 2
- [51] Zhanyi Sun and Shuran Song. Latent policy barrier: Learning robust visuomotor policies by staying in-distribution. *arXiv preprint arXiv:2508.05941*, 2025. 2, 3, 4, 5, 6, 27
- [52] 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. 2
- [53] 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 *2022 International Conference on Robotics and Automation (ICRA)*, pages 10781–10787. IEEE, 2022. 2
- [54] Vive Tracker. <https://www.vive.com/hk/accessory/tracker3>, 2025. 2
- [55] Yanwei Wang, Lirui Wang, Yilun Du, Balakumar Sundaralingam, Xuning Yang, Yu-Wei Chao, Claudia Pérez-D’Arpino, Dieter Fox, and Julie Shah. Inference-time policy steering through human interactions. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 15626–15633. IEEE, 2025. 2, 4
- [56] Ziyue Wang, Li Kang, Yiran Qin, Jiahua Ma, Zhanglin Peng, Lei Bai, and Ruimao Zhang. Gaudp: Reinventing multi-agent collaboration through gaussian-image synergy in diffusion policies. *arXiv preprint arXiv:2511.00998*, 2025. 1, 2
- [57] Lai Wei, Jiahua Ma, Yibo Hu, and Ruimao Zhang. Ensuring force safety in vision-guided robotic manipulation via implicit tactile calibration. *arXiv preprint arXiv:2412.10349*, 2024. 1, 6, 18
- [58] Junjie Wen, Minjie Zhu, Yichen Zhu, Zhibin Tang, Jinming Li, Zhongyi Zhou, Chengmeng Li, Xiaoyu Liu, Yaxin Peng, Chaomin Shen, et al. Diffusion-vla: Generalizable and interpretable robot foundation model via self-generated reasoning. *arXiv preprint arXiv:2412.03293*, 2024. 2
- [59] Longyan Wu, Checheng Yu, Jieji Ren, Li Chen, Yufei Jiang, Ran Huang, Guoying Gu, and Hongyang Li. Freetacman: Robot-free visuo-tactile data collection system for contact-rich manipulation. *arXiv preprint arXiv:2506.01941*, 2025. 2, 19, 21
- [60] Philipp Wu, Yide Shentu, Zhongke Yi, Xingyu Lin, and Pieter Abbeel. Gello: A general, low-cost, and intuitive teleoperation framework for robot manipulators. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 12156–12163. IEEE, 2024. 2, 19, 20
- [61] Xense. <https://www.xenserobotics.com/product/367/detail/9>, 2025. 2
- [62] Xense SDK. <https://xensedoc.readthedocs.io/en/latest/XenseSDK/XenseSDK.html>, 2025. 2, 3, 27
- [63] Mengda Xu, Han Zhang, Yifan Hou, Zhenjia Xu, Linxi Fan, Manuela Veloso, and Shuran Song. Dexumi: Using human hand as the universal manipulation interface for dexterous manipulation. *arXiv preprint arXiv:2505.21864*, 2025. 2
- [64] Yue Xu, Litao Wei, Pengyu An, Qingyu Zhang, and Yong-Lu Li. exumi: Extensible robot teaching system with action-aware task-agnostic tactile representation. *arXiv preprint arXiv:2509.14688*, 2025. 2, 19, 21
- [65] 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, 2, 3, 6, 7, 18, 19, 20, 27
- [66] 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. 2
- [67] Samson Yu, Kelvin Lin, and Harold Soh. Demonstrating the octopi-1.5 visual-tactile-language model. *arXiv preprint arXiv:2507.09985*, 2025. 2
- [68] Wenzhen Yuan, Siyuan Dong, and Edward H Adelson. Gelsight: High-resolution robot tactile sensors for estimating geometry and force. *Sensors*, 17(12):2762, 2017. 2
- [69] Yanjie Ze, Gu Zhang, Kangning Zhang, Chenyuan Hu, Muhan Wang, and Huazhe Xu. 3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations. *arXiv preprint arXiv:2403.03954*, 2024. 1, 2
- [70] Chaofan Zhang, Peng Hao, Xiaoge Cao, Xiaoshuai Hao, Shaowei Cui, and Shuo Wang. Vtla: Vision-tactile-language-action model with preference learning for insertion manipulation. *arXiv preprint arXiv:2505.09577*, 2025. 2
- [71] 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. 2
- [72] Xiangcheng Zhang, Haowei Lin, Haotian Ye, James Zou, Jianzhu Ma, Yitao Liang, and Yilun Du. Inference-time scaling of diffusion models through classical search. *arXiv preprint arXiv:2505.23614*, 2025. 2
- [73] Yazhan Zhang, Zicheng Kan, Yang Yang, Yu Alexander Tse, and Michael Yu Wang. Effective estimation of contact force and torque for vision-based tactile sensors with helmholtz–hodge decomposition. *IEEE Robotics and Automation Letters*, 4(4):4094–4101, 2019. 2
- [74] Can Zhao, Jin Liu, and Daolin Ma. ifem2.0: Dense 3d contact force field reconstruction and assessment for vision-based tactile sensors. *IEEE Transactions on Robotics*, 2024. 2
- [75] Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023. 1, 2, 19, 20
- [76] Zhaxizhuom Zhaxizhuoma, Kehui Liu, Chuyue Guan, Zhongjie Jia, Ziniu Wu, Xin Liu, Tianyu Wang, Shuai Liang, Pengan CHEN, Pingrui Zhang, et al. Fastumi: A scalable and hardware-independent universal manipulation interface with dataset. In *Conference on Robot Learning*, pages 3069–3093. PMLR, 2025. 2, 3, 19, 21
- [77] Qinqing Zheng, Matt Le, Neta Shaul, Yaron Lipman, Aditya Grover, and Ricky TQ Chen. Guided flows for generative modeling and decision making. *arXiv preprint arXiv:2311.13443*, 2023. 4, 5, 15
- [78] Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. Dino-wm: World models on pre-trained visual features enable zero-shot planning. *arXiv preprint arXiv:2411.04983*, 2024. 4
- [79] Xinyue Zhu, Binghao Huang, and Yunzhu Li. Touch in the wild: Learning fine-grained manipulation with a portable visuo-tactile gripper. *arXiv preprint arXiv:2507.15062*, 2025. 2, 19, 21