IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

PoCoDP3: Pose- and Contact-Aware Visual-Tactile Policy for Contact-Rich 3D Manipulation

Zhaokun Yue, Ling Tong, *Graduate Student Member, IEEE*, and Kun Qian, *Member, IEEE*

***Abstract*—Imitation learning in contact-rich tasks requires both global spatial awareness and fine-grained in-hand interaction understanding. However, vision-only policies based on images or point clouds are often susceptible to occlusion and struggle to capture critical contact details, particularly in visually ambiguous regions or during subtle tactile interactions. In this work, we present PoCoDP3, a pose- and contact-aware visual-tactile policy that integrates 3D point clouds and tactile inputs to generate actions in contact-rich tasks. PoCoDP3 introduces a dual-branch tactile encoder that jointly models contact dynamics and estimates in-hand object pose, enabling structured tactile representations for precise contact-rich manipulation. A contact-driven cross-modal fusion mechanism adaptively prioritizes sensory modalities based on real-time interaction cues, enabling efficient visual-tactile integration. Moreover, a reference-guided diffusion policy leverages reference action offsets to reduce sampling steps, significantly accelerating inference while maintaining action quality. Experiments across simulation and real-world tasks demonstrate that PoCoDP3 consistently outperforms representative 2D and 3D policies in terms of both accuracy and inference efficiency.**

***Index Terms*—Force and tactile sensing, imitation learning, manipulation planning.**

I. Introduction

IMITATION learning provides an efficient paradigm for enabling robots to acquire complex manipulation skills by observing and replicating expert demonstrations. The effectiveness of this approach is closely tied to the completeness and quality of observational data. Benefiting from the superior capability of 3D representations in modeling spatial structures and geometric relationships, 3D vision-based imitation learning approaches [1], [2] exhibit enhanced scene understanding and better task generalization compared to image-based methods [3]. However, point clouds are sparse and susceptible to occlusion, making it challenging to capture critical in-hand interaction details, especially when interactions occur in visual blind spots or when geometric information is ambiguous. This presents a major challenge for learning fine-grained skills in contact-rich tasks.

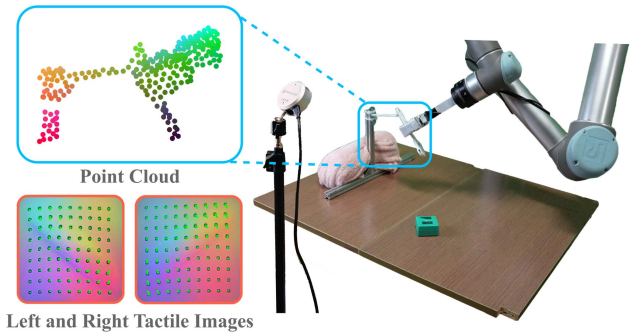

Fig. 1. **Tactile Compensation for 3D Visual Limitations:** Point clouds are sparse and prone to occlusion, whereas visuotactile sensing provides precise local feedback that complements visual perception.

As a powerful complement to visual perception, visuotactile sensors such as GelSight and GelStereo [4] provide high-resolution local feedback at contact points, illustrated in Fig. 1. Recent studies have demonstrated that integrating visual and tactile modalities into imitation learning frameworks can substantially improve robotic performance [5], [6]. However, existing methods often rely on complex pretraining procedures and primarily focus on fusing tactile signals with visual images, while paying limited attention to integrating tactile sensing with point cloud data. Moreover, common fusion strategies, such as direct feature concatenation, fail to account for the varying importance of each modality across different stages of manipulation, which may compromise generalization or lead to information loss in one of the modalities. These challenges highlight the need for more effective feature fusion strategies that not only capture cross-modal complementary information but also adaptively prioritize the most relevant sensory inputs during interaction.

In addition to comprehensive perception, efficient real-time inference is equally critical for contact-rich manipulation tasks. Recently, imitation learning based on diffusion models [1], [2], [3] has garnered significant attention due to its outstanding performance in action generation and generalization. Despite these promising results, the inherent reliance of diffusion models on multi-step denoising severely limits their inference efficiency, making them less suitable for scenarios that demand real-time responses. This limitation is particularly critical in contact-rich tasks, which require high control frequencies and rapid adaptation to external contact variations to ensure stable and safe interactions. Consequently, achieving efficient inference without

Received 10 August 2025; accepted 1 December 2025. Date of publication 12 January 2026; date of current version 20 January 2026. This article was recommended for publication by Associate Editor S. James and Editor A. Faust upon evaluation of the reviewers' comments. This work was supported in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2025A1515010397 and in part by the Project of the Joint Research and Development Center for Future Robotics and Artificial Intelligence, Southeast University. *(Corresponding author: Kun Qian.)*

Zhaokun Yue and Ling Tong are with the School of Automation, Southeast University, Nanjing 210096, China, and also with the Key Laboratory of Measurement and Control of CSE, Ministry of Education, Nanjing 210096, China (e-mail: yuezhaokun@seu.edu.cn).

Kun Qian is with Southeast University Shenzhen Research Institute, Shenzhen 518063, China, and also with the School of Automation, Southeast University, Nanjing 210096, China (e-mail: kqian@seu.edu.cn).

Project website: https://pocodp3.github.io/.

Digital Object Identifier 10.1109/LRA.2026.3653314

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

degrading policy performance remains a critical challenge for deploying diffusion models in highly dynamic environments.

In this letter, we propose Pose- and Contact-Aware DP3 (PoCoDP3), a novel policy learning framework that tightly integrates tactile sensing and point cloud observations for robust decision-making in contact-rich scenarios. PoCoDP3 takes point clouds and tactile images as inputs and generates continuous action trajectories. A dual-branch tactile encoder is designed: one branch estimates the object's pose relative to the gripper, providing a structured use of tactile information that circumvents the challenges of direct feature extraction from raw tactile images, thereby reducing training complexity and improving policy stability; the other encodes the marker motion to capture dynamic contact states. Based on this representation, we introduce a contact-driven dynamic modality selection mechanism that enables the policy to adaptively attend to the most informative sensory modality during manipulation. To improve inference efficiency, we further introduce a reference-guided diffusion mechanism inspired by [7], which decouples noise perturbation from action offset modeling. Guided by a reference trajectory, this mechanism enables the policy to generate high-quality action sequences with substantially fewer sampling steps, reducing inference overhead while maintaining control performance.

Extensive experiments in both simulation and real-world settings demonstrate that PoCoDP3 achieves superior performance and efficiency in complex, contact-rich manipulation tasks compared to state-of-the-art diffusion-based policies.

Our key contributions are summarized as follows:

- We propose a 3D visual-tactile imitation learning method that combines contact dynamics modeling and object pose estimation, enabling structured tactile representations of in-hand interaction states and improving both precision and robustness in contact-rich tasks that require pose adjustment.
- We introduce a contact-driven cross-modal fusion mechanism that adaptively allocates modality attention based on contact cues derived from shear forces, enabling efficient integration of tactile and point cloud information throughout different manipulation stages.
- We develop a reference-guided diffusion policy that significantly reduces sampling steps required for action generation, enabling faster inference while preserving both action consistency and diversity.

II. Related Work

A. Contact States and In-Hand Pose Estimation

Earlier research indicates that visuotactile sensors, such as GelSight and GelStereo [4], are capable of accurately estimating contact states. [8] proposes a constraint-optimization framework that uses tactile motion tracking to localize extrinsic contacts without requiring prior knowledge of object geometry. [9] actively estimates the extrinsic contact line via a factor graph with tactile feedback and uses it as structured input to a learned insertion policy. Object in-hand pose estimation is critical for robotic manipulation. Tactile-based methods [10] estimate the object's pose by matching observed contact shapes with those generated by a simulator. [11] estimates the 6D pose by fusing visual features with object-surface point cloud features from tactile sensors. In [12], an implicit representation is learned to enable simultaneous pose estimation and prediction of extrinsic contact locations. However, most existing methods treat contact reasoning and pose estimation as separate tasks, which limits their applicability to imitation learning. In this letter, we propose a dual-branch tactile encoder that simultaneously encodes both the in-hand object pose and implicit contact states to extract informative features for imitation learning.

B. Visual-Tactile Manipulation

Tactile sensing has been integrated into various control paradigms, including classical control, reinforcement learning, and imitation learning. [13] combines contrastive learning and Koopman-based model predictive control for efficient tactile servoing of deformable linear objects with high-dimensional inputs. Visual-tactile reinforcement learning methods have achieved notable success in contact-rich tasks such as peg insertion [14], cable-in-duct [15], and door opening [16]. However, these approaches often rely on extensive simulation training, resulting in high computational overhead. To address this, imitation learning offers an alternative paradigm that enables robots to acquire policies from human demonstrations, significantly reducing training costs. [17] enables tactile-guided control from demonstrations but omits visual input, limiting global scene understanding. Recent visual-tactile systems such as 3D-ViTac [18] demonstrate the benefits of integrating dense tactile feedback and visual perception within a unified 3D space, enabling fine-grained manipulation and robustness under occlusion. In contrast, our approach focuses on structured visual-tactile fusion that jointly captures in-hand pose and contact dynamics for robust 3D manipulation.

C. Diffusion-Based Imitation Learning

Diffusion-based imitation learning has emerged as a powerful paradigm for modeling diverse and multimodal behaviors from demonstrations. Diffusion Policy [3] and its 3D extension [1] generate action sequences through conditional denoising of visual or point cloud observations, achieving strong visuomotor performance. However, diffusion models typically require multiple iterations of denoising, which limits their suitability for real-time control. Recent studies have explored acceleration alternatives such as Flow Matching [19] and Consistency Policy [20]. Flow Matching replaces the stochastic diffusion process with a deterministic vector-field evolution, while Consistency Policy distills multi-step denoising into a single-step mapping. These approaches demonstrate impressive inference efficiency. However, their deterministic nature may limit stochastic exploration, which is often beneficial for maintaining action diversity in contact-rich manipulation. To balance efficiency and expressiveness, we propose a reference-guided diffusion mechanism that retains the stochastic denoising formulation but introduces a directional reference prior, achieving faster inference while preserving multimodal generation capability.

III. Proposed Method

Given a set of expert demonstrations for robotic skills, we aim to learn a visual-tactile policy $\pi : \mathcal{O} \mapsto \mathcal{A}$ that maps point cloud and tactile observations $o \in \mathcal{O} = \{\mathcal{O}^{pc}, \mathcal{O}^{tac}\}$ to actions $a \in \mathcal{A}$, enabling efficient multimodal fusion and real-time action inference. To this end, we propose PoCoDP3, a unified framework comprising two key components: a perception module that encodes and fuses point cloud and tactile inputs to

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

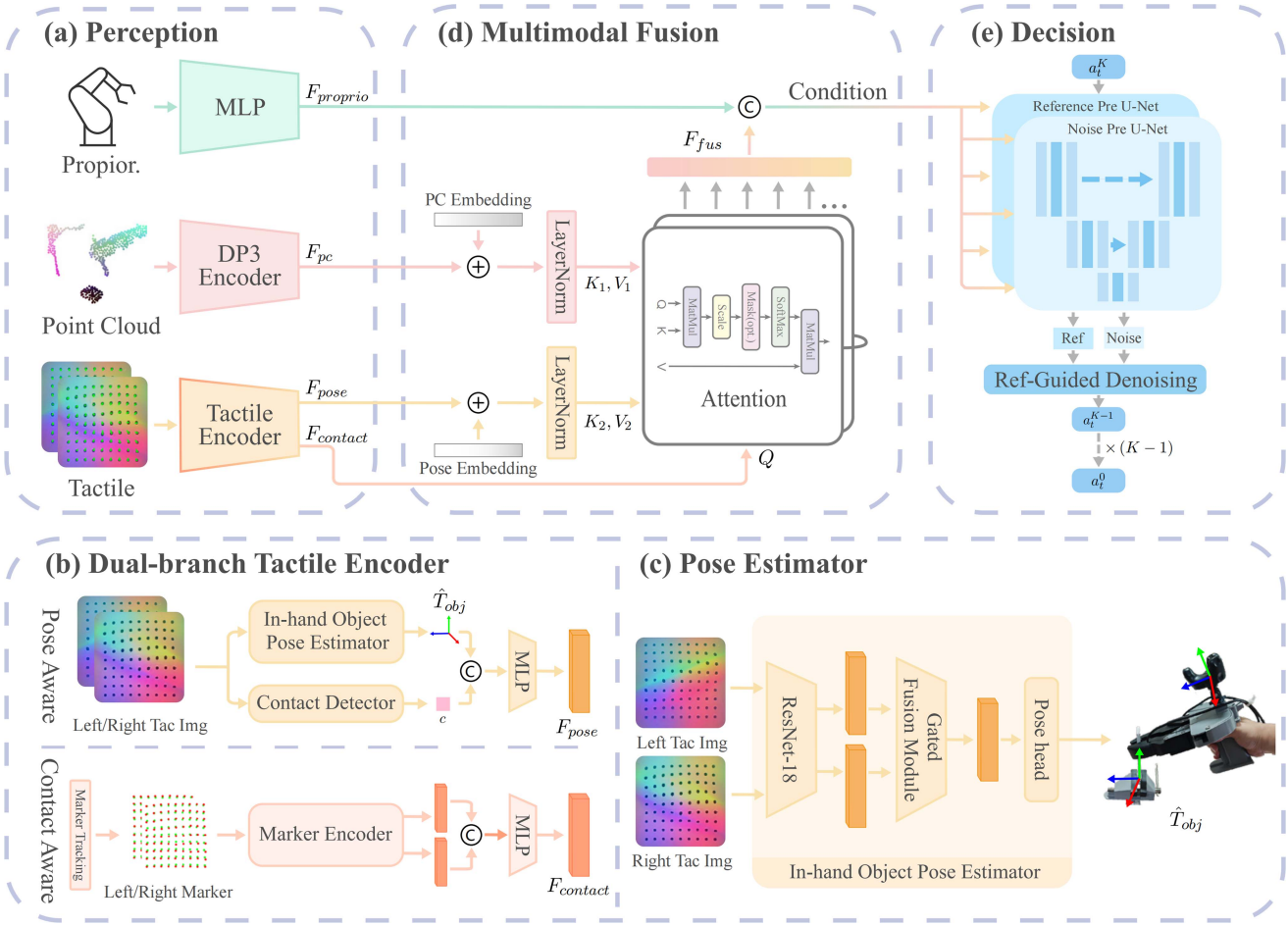

Fig. 2. **PoCoDP3 Architecture:** (a) Proprioceptive input is encoded by an MLP, point clouds by a DP3 encoder, and tactile data by a dual-branch tactile encoder. (b) The dual-branch tactile encoder uses tactile images and marker motion to encode object pose and contact state. (c) The pose estimator in the pose-aware branch predicts the 6-DoF in-hand object pose from tactile images. (d) Encoded contact information guide point cloud–tactile fusion via an attention mechanism. (e) The fused features serve as conditional input to a reference-guided diffusion policy for generating action trajectories.

construct a comprehensive representation of the environment, and a decision module that generates action sequences using a reference-guided diffusion policy conditioned on the fused visual-tactile representation. The overview of our architecture is shown in Fig. 2.

A. Learning 3D and Tactile Representation

Point Cloud Encoder: Following DP3 [1], we use a lightweight MLP-based encoder to extract a compact 3D representation $F_{pc} \in \mathbb{R}^{64}$ from the cropped and downsampled point cloud $\boldsymbol{p} \in \mathbb{R}^{N_p \times 3}$.

Tactile Encoder: We propose a dual-branch tactile encoder that processes data from visuotactile sensors mounted on both sides of the gripper to extract contact and pose features, enabling a comprehensive tactile representation for downstream decision-making.

- **Contact-Aware Branch:** is designed to capture extrinsic contact interactions by leveraging the tactile shear field represented by the motion of an $11\times$ 11 marker array. The original and displaced marker positions on both sides tactile sensors are concatenated respectively and fed into an encoder with a PointNet-like architecture [21] to extract contact features $F_{marker,l}, F_{marker,r} \in \mathbb{R}^{64}$. These features are then concatenated and passed through a single-layer MLP to generate the unified contact-aware feature $F_{contact} \in \mathbb{R}^{64}$.
- **Pose-Aware Branch:** extracts pose-related features and incorporates an in-hand pose estimator that regresses the object's SE(3) pose $\hat{T}_{obj} \in \mathbb{R}^6$ relative to the gripper based on tactile images. The regressed pose vector is concatenated with a binary contact indicator $c \in \{0, 1\}$ to form an augmented representation $\tilde{T}_{obj} = [\hat{T}_{obj}; c]$, which is subsequently passed through a three-layer MLP to produce the final pose-aware feature $F_{pose} \in \mathbb{R}^{64}$.

In-hand Pose Estimator: We train the in-hand pose estimation module using supervised learning. Specifically, tactile images from both sides are separately fed into a shared ResNet-18 backbone (trained from scratch) to extract feature maps $\mathbf{F}_l, \mathbf{F}_r \in \mathbb{R}^{512\times7\times7}$. An attention pooling operation is then applied to compress each feature map into a 1D tactile feature vector $F_{tac,l}, F_{tac,r} \in \mathbb{R}^{64}$. Subsequently, a gated fusion module aggregates the left and right features to produce a fused embedding $F_{tac,fused} \in \mathbb{R}^{64}$ as follows:

$$\mathbf{g} = \sigma\left(\mathbf{W}_g[F_{tac,l}; F_{tac,r}] + \mathbf{b}_g\right),$$

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

$$F_{tac,fused} = \mathbf{g} \odot F_{tac,l} + (1 - \mathbf{g}) \odot F_{tac,r}. \tag{1}$$

Finally, the fused tactile feature is passed through a pose regression head to predict the 6-DoF SE(3) pose of the object relative to the gripper:

$$\hat{T}_{obj} = f_{pose}(F_{tac,fused}) \in \mathbb{R}^6. \tag{2}$$

Unit quaternions are widely used to represent rotations, forming the nonlinear Riemannian manifold S^3, whereas translational displacements reside in the Euclidean space $\mathbb{R}^3$. The structural inconsistency between these two types of data leads to semantic misalignment and degraded learning performance if directly concatenated and fed into an MLP. To address this issue, we apply a logarithmic map [22] to project unit quaternions on S^3 onto the tangent space at a reference point, which locally linearizes the manifold into a Euclidean space. This allows rotational information to be embedded in the same space as translational data, thereby enabling a unified representation suitable for learning. Specifically, a unit quaternion $\boldsymbol{q} = (q_0, \vec{q}) \in S^3$ can be mapped to a rotation vector $\log_e(\boldsymbol{q}) \in \mathbb{R}^3$ via:

$$\log_e(\boldsymbol{q}) \begin{cases} \mathrm{ac}_*(q_0)\frac{\vec{q}}{\|\vec{q}\|}, & q_0 \neq 1 \\ \vec{0}, & q_0 = 1 \end{cases}, \tag{3}$$

where $\mathrm{ac}_*(\,\cdot\,)$ denotes a modified version of the arccosine, and the subscript e indicates that the logarithmic map is taken with respect to the identity element $e = (1, \vec{0}) \in S^3$.

To ensure that pose estimation is performed only under valid contact, we adopt a structural similarity (SSIM)–based method to detect contact between the tactile sensors and the object. The current left and right tactile images, I_l and I_r, are compared against reference images captured in non-contact conditions. A binary contact indicator is then computed as:

$$c = \mathrm{SSIM}_b(I_l) \quad \wedge \quad \mathrm{SSIM}_b(I_r), \tag{4}$$

where $\mathrm{SSIM}_b(\,\cdot\,) \in \{0, 1\}$ returns 1 if the similarity exceeds a predefined threshold. Pose estimation is performed only when $c = 1$, indicating valid contact; otherwise, the process is skipped, and the estimated pose is set to $\hat{T}_{obj} = \vec{0}$. This enables the policy to distinguish between contact and non-contact states, thereby improving decision-making.

B. Contact-Driven Dynamic Modality Fusion

The importance of different sensory modalities changes throughout the manipulation process. To adaptively leverage the most informative cues, we propose a contact-driven dynamic fusion mechanism that selectively integrates point cloud and tactile-derived pose features based on the current interaction state between the robot and the environment.

The shear field extracted from marker motion offers a reliable signal for estimating the interaction state. We leverage this information by using the contact-aware feature $F_{contact} \in \mathbb{R}^{64}$ as the Query in a cross-attention module, while the point cloud feature $F_{pc} \in \mathbb{R}^{64}$ and the pose-aware feature $F_{pose} \in \mathbb{R}^{64}$ serve as the Key and the Value.

To improve modality discrimination while maintaining training stability and a consistent feature distribution, we incorporate learnable modality embeddings $E_{pc}, E_{pose} \in \mathbb{R}^{64}$ directly into the Layer Normalization operation:

$$\hat{F}_{pc} = \mathrm{LayerNorm}(F_{pc} + E_{pc}),$$
$$\hat{F}_{pose} = \mathrm{LayerNorm}(F_{pose} + E_{pose}). \tag{5}$$

The query, key, and value projections are computed as follows:

$$Q = F_{contact} W_Q,$$
$$K = [\hat{F}_{pc}, \hat{F}_{pose}] W_K,$$
$$V = [\hat{F}_{pc}, \hat{F}_{pose}] W_V, \tag{6}$$

where W_Q, W_K, and W_V are learnable projection matrices, and $[\,,\,]$ denotes feature concatenation.

The attention weights over modalities are computed as:

$$\alpha_{pc}, \alpha_{pose} = \sigma\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right), \tag{7}$$

where $\sigma(\,\cdot\,)$ denotes the softmax function and d_k is the dimensionality of the Key vectors.

Finally, a unified fused feature $F_{fus} \in \mathbb{R}^{64}$ is computed as a weighted combination of the modality features:

$$F_{fus} = \alpha_{pc} V_{pc} + \alpha_{pose} V_{pose}. \tag{8}$$

The contact-aware fusion token F_{fus} is subsequently used as a conditional input to the diffusion policy module for generating context-aware action sequences.

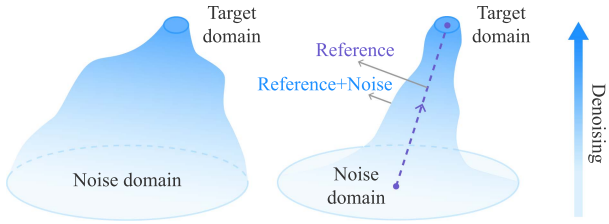

Fig. 3. **Reference-Guided Denoising Process:** Guided by the reference trajectory (purple), reference-guided DP accelerates the convergence of predicted trajectories toward the target domain during the denoising process, significantly reducing the number of required denoising steps.

C. Reference-Guided Diffusion Policy

Diffusion-based policies [3] generate actions by progressively denoising Gaussian noise conditioned on the current observation. Despite their effectiveness in imitation learning, their multi-step sampling incurs high inference latency and lacks explicit modeling of the deterministic offset between the current state and the target action, limiting their responsiveness and consistency in contact-rich tasks.

To address these limitations, we propose a reference-guided diffusion policy, which decouples the diffusion process into directional reference-guided diffusion and stochastic noise diffusion, enhancing both inference efficiency and deterministic control, as shown in Fig. 3. This mechanism enables the policy to match the performance of conventional diffusion policy using only a few sampling steps (e.g., 2–5 steps). Specifically, the forward process is defined as:

$$A_t = A_{t-1} + A^t_{ref}, \quad A^t_{ref} \sim \mathcal{N}(\alpha_t A_{ref}, \beta_t^2 \mathbf{I}). \tag{9}$$

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

Here, A^t_{ref} represents a directional mean shift with added random perturbation, transitioning the action from state A_{t-1} to A_t. A_{ref} in A^t_{ref} represents the reference action that progressively guides the generation process from noise toward the target action A_0. To provide this directional prior, we define the reference guidance as the offset from a hypothetical zero action to the target action, i.e., $A_{ref} = -A_0$. α_t and β_t are two independent coefficient schedules that respectively control the reference-guided and noise diffusion processes.

Subsequently, the forward process unfolds as:

$$\begin{aligned} A_t &= A_{t-1} + \alpha_t A_{ref} + \beta_t \epsilon_{t-1} \\ &= A_{t-2} + (\alpha_{t-1} + \alpha_t) A_{ref} + \left(\sqrt{\beta^2_{t-1} + \beta^2_t}\right) \epsilon_{t-2} \\ &= \dots \\ &= A_0 + \bar{\alpha}_t A_{ref} + \bar{\beta}_t \epsilon, \end{aligned} \tag{10}$$

where $\epsilon_{t-1}, \dots, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\bar{\alpha}_t = \sum_{i=1}^t \alpha_i$ and $\bar{\beta}_t = \sqrt{\sum_{i=1}^t \beta_i^2}$.

In the forward process, the reference guidance A_{ref} and noise ϵ are gradually injected to generate A_t from A_0. In contrast, the reverse process from A_t to A_0 requires estimating and removing the added guidance and noise components to recover the target action. The estimated target action is computed as:

$$A^\theta_0 = A_t - \bar{\alpha}_t \hat{A}^\theta_{ref} - \bar{\beta}_t \hat{\epsilon}_\theta. \tag{11}$$

Correspondingly, the reverse denoising step can be formulated as:

$$A_{t-1} = A_t - (\bar{\alpha}_t - \bar{\alpha}_{t-1}) \hat{A}^\theta_{ref} - (\bar{\beta}_t - \bar{\beta}_{t-1}) \hat{\epsilon}_\theta, \tag{12}$$

where $\hat{A}^\theta_{ref}$ and $\hat{\epsilon}_\theta$ are predicted by the reference prediction network $A^\theta_{ref}(A_t, t)$ and the noise prediction network $\epsilon_\theta(A_t, t)$, respectively.

We define the following loss function for model training:

$$\begin{aligned} \mathcal{L}_{ref}(\theta) &= \mathbb{E}\left[\lambda_{ref} \| A_{ref} - A^\theta_{ref}(A_t, t) \|^2\right], \\ \mathcal{L}_{\epsilon}(\theta) &= \mathbb{E}\left[\lambda_{\epsilon} \| \epsilon - \epsilon_\theta(A_t, t) \|^2\right], \end{aligned} \tag{13}$$

where $\lambda_{ref}, \lambda_\epsilon \in 0, 1$, and during training, the input action trajectory A_t is generated from A_{ref} and ϵ according to the following equation:

$$A_t = (\bar{\alpha}_t - 1) A_{ref} + \bar{\beta}_t \epsilon. \tag{14}$$

IV. Experiment

A. Pose Estimator Training

The tactile in-hand object pose estimation module takes tactile images as input and predicts the relative pose between the object and the gripper. For real-world experiments, we adopt a handheld device integrated with a Vive Tracker and tactile sensors, as illustrated in Fig. 4, to efficiently collect data. The data collection process involves two stages: (a) **Object Frame Setup**: The tracker is placed on a mounting platform to define the object coordinate frame origin; (b) **Data Collection**: The object is fixed at the origin, and the relative pose ${}^{eef}_{obj}\mathbf{T} = ({}^{obj}_{eef}\mathbf{T})^{-1}$ is computed from the tracker readings ${}^{obj}_{eef}\mathbf{T}$. For each object used in the experiments, randomized grasps with varying poses and forces are applied using the collection device, followed by random shaking. Grasps are reset every 10 seconds, and 100 sequences are collected per object.

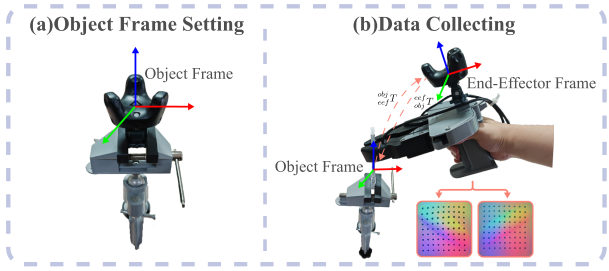

Fig. 4. **Data Collection for Pose Estimation:** The tracker is used to define the object frame and record gripper-object relative poses during randomized grasp interactions.

It is worth noting that in the object coordinate frame, certain degrees of freedom, such as rotation around the central axis of cylindrical objects, are less relevant to policy learning and difficult to infer using only tactile data. To improve training stability, we set these components to zero.

B. Experiment Setup

We conduct a comprehensive evaluation of our method across six tasks, including three simulation tasks from Robomimic [23] and MimicGen [24], and three real-world tasks. For the simulation tasks, we integrate a FOTS [25] visuotactile sensor simulator into the environment to generate tactile images and marker motion data. For the real-world tasks, our experimental setup consists of a UR5 robot equipped with an Inspire EG2-4C2 gripper, with two Vitai visuotactile sensors mounted on the fingertips. An Intel RealSense L515 RGB-D camera is positioned in front of the workspace to provide visual and point cloud observations of the scene. All devices are connected to a workstation equipped with an Intel Core i5-14600KF CPU and an NVIDIA RTX 4070 Ti Super GPU, which is used for data collection, policy training, and evaluation.

Baseline and Ablation Methods: We select two state-of-the-art diffusion-based policies as baselines for comparison: (1) Diffusion Policy (DP) [3], which generates actions conditioned on image observations; and (2) DP3 [1], which replaces image inputs with point cloud observations. In addition, we compare four ablated or variant models: (1) PoCoDP3 (w/o PoseEst): replaces the pose features with ResNet-encoded tactile image features; (2) PoCoDP3 (w/o Attn): directly concatenates the pose and point cloud features without using the dynamic modality fusion mechanism; (3) PoCoDP3 (w/o Ref): uses the standard diffusion policy instead of reference-guided one; and (4) PoCoDP3 (FM): adopts Flow-Matching [19] instead of diffusion for action generation. All methods are evaluated in terms of inference time and success rate.

Implementation Details: In simulation, the RGB images from both cameras and tactile sensors have a resolution of 84× 84. In real-world experiments, RGB images from the global camera with a resolution of 320× 240 are used, and the tactile images are resized to 224× 224. In both settings, the point clouds captured from the global camera have a shape of 512× 3. The diffusion-based policies use 100 training and sampling steps for DDPM. All policies are trained for 1000 epochs with a chunking length of 8 and a prediction horizon of 16. During inference, the DDIM scheduler is employed with 15 steps for DP and 5 steps for reference-guided DP, whereas FM uses 5 inference steps.

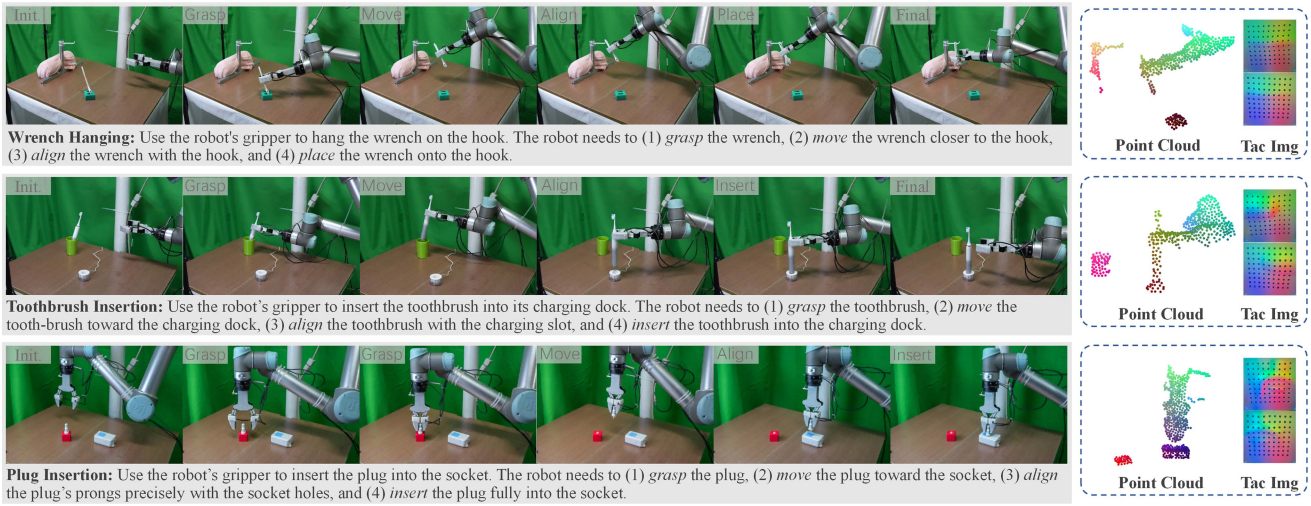

Fig. 5. **Real-World Experiment:** We carefully design three challenging contact-rich tasks that involve both non-contact and contact phases, enabling comprehensive evaluation.

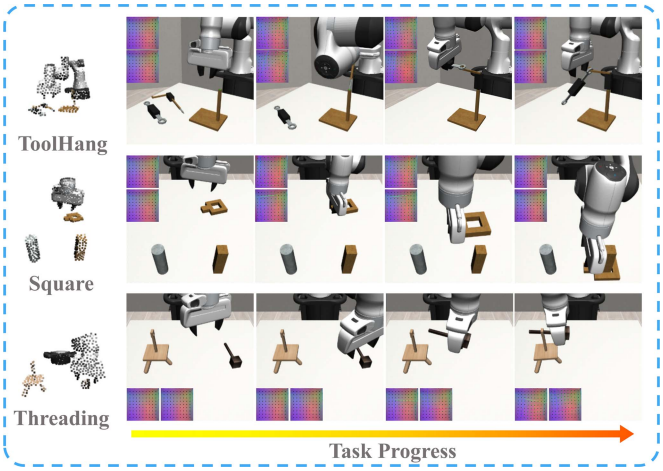

Fig. 6. **Simulation Experiment:** We evaluate our method on tasks from the Robomimic [23] and MimicGen [24] benchmarks. FOTS [25] is integrated into the simulator to generate real-time tactile images and marker motion. From top to bottom, the tasks are ToolHang, Square, and Threading.

TABLE I
SIMULATION QUANTITATIVE RESULTS

Method	Inference Time (s)	Average Success Rate		
		ToolHang	Threading	Square
Diffusion Policy	0.184	0.73	0.65	0.92
3D Diffusion Policy	0.173	0.65	0.59	0.90
PoCoDP3(w/o Attn)	0.092	0.77	0.69	0.92
PoCoDP3(w/o PoseEst)	0.094	0.18	0.19	0.22
PoCoDP3(w/o Ref)	0.178	**0.83**	0.74	0.93
PoCoDP3(FM)	**0.072**	0.78	0.72	0.92
PoCoDP3(Ours)	0.094	0.82	**0.75**	**0.93**

C. Simulation Performance

We evaluate our method on three simulation tasks: ToolHang and Square from Robomimic, and Threading D0 from MimicGen, as shown in Fig. 6. Each policy is trained on 200 demonstrations and evaluated by the average success rate over 50 environment initializations, across 10 checkpoints over 3 seeds. Results are summarized in Table I.

PoCoDP3 achieves superior or comparable performance to all baseline methods and ablated variants across the three tasks. Specifically, compared to the baseline methods that only use visual or point cloud inputs, our method leverages tactile sensing to capture fine-grained in-hand object states, effectively compensating for missing local interaction cues caused by occlusions and point cloud sparsity in contact-rich tasks. Furthermore, the contact-driven attention strengthens the dynamic fusion between tactile and point cloud inputs, jointly boosting task success rates. Additionally, the reference-guided diffusion policy significantly reduces inference time without compromising performance, and achieves comparable or even better performance than the Flow Matching–based variant. Finally, comparisons with PoCoDP3(w/o PoseEst) show that the absence of the structured tactile representations hinders effective fusion with the point cloud features, resulting in degraded performance.

This validates the importance of explicit in-hand object pose estimation for achieving effective fusion between tactile and point cloud modalities, providing structured representations that substantially stabilize policy learning.

D. Real-World Performance

To confirm that PoCoDP3 performs well in real-world settings, we design three challenging contact-rich manipulation tasks requiring precise pose control: Wrench Hanging (W-Hang), Toothbrush Insertion (T-Ins), and Plug Insertion (P-Ins), as illustrated in Fig. 5. Each task involved both contact and non-contact phases, with the aim of evaluating the policy's capability to adapt to dynamic contact conditions and to robustly fuse visual and tactile inputs throughout the manipulation process. For data collection, a teleoperation system based on the Vive Tracker was employed to collect 80 demonstrations for each task. For evaluation, we report success rates averaged over 20 trials for each task, with randomized initial configurations between trials.

The results are summarized in Table II, where PoCoDP3 demonstrates strong overall performance across all real-world

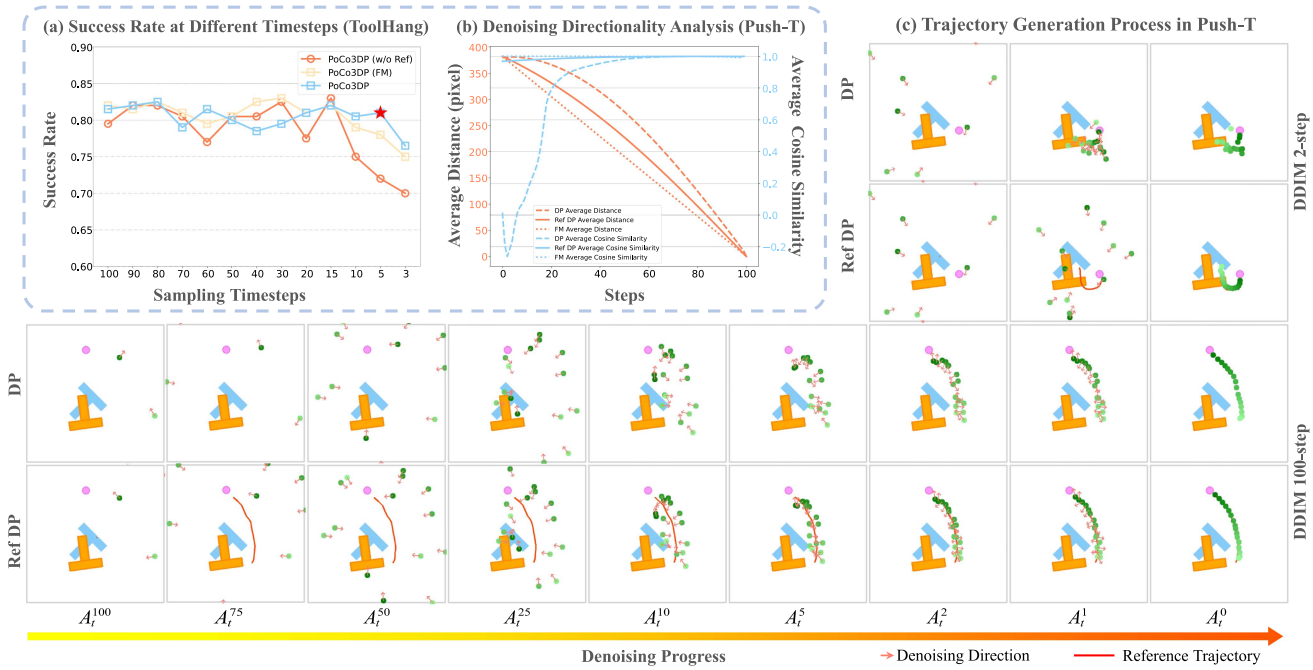

Fig. 7. **Efficiency and Effectiveness of Reference-Guided DP:** (a) On the ToolHang task, PoCoDP3 maintains high success rates and faster inference under fewer sampling steps, outperforming its unguided variant. (b) In the Push-T task, reference guidance improves denoising directionality, leading to earlier convergence and higher cosine similarity. (c) Trajectory visualizations under 2-step and 100-step DDIM further highlight the benefits of guidance.

TABLE II
REAL-WORLD QUANTITATIVE RESULTS

Method	Inference Time (s)	Success Rate		
		W-Hang	T-Ins	P-Ins
Diffusion Policy	0.190	0.65	0.75	0.45
3D Diffusion Policy	0.177	0.50	0.70	0.40
PoCoDP3(w/o Attn)	0.094	0.65	0.80	0.55
PoCoDP3(w/o PoseEst)	0.096	0.15	0.20	0.05
PoCoDP3(w/o Ref)	0.184	0.80	0.85	**0.75**
PoCoDP3(FM)	**0.075**	0.70	0.85	0.65
PoCoDP3(Ours)	0.098	**0.80**	**0.90**	0.70

tasks. For Wrench Hanging, we observe that the most common failure mode of DP and DP3 is the premature release of the gripper before the wrench hole is properly aligned with the hook. For Toothbrush and Plug Insertion, failures are often caused by inaccurate pose adjustment of the manipulated object, resulting in unsuccessful insertion. In contrast, PoCoDP3 benefits from local tactile feedback that complements visual perception, making it less prone to such failures. Removing the attention module reduces success rates, confirming its role in robust visual-tactile fusion, whereas removing the pose estimation module results in an even more pronounced degradation, underscoring the importance of structured tactile representations.

E. Sampling Efficiency and Action Diversity Evaluation

We evaluate PoCoDP3, PoCoDP3 (w/o Ref) and PoCoDP3 (FM) under varying sampling steps on the ToolHang task, as shown in Fig. 7(a). The success rates are measured using the final checkpoint over 200 environment initializations. While PoCoDP3 and PoCoDP3 (w/o Ref) perform similarly with a large number of steps, the success rate of PoCoDP3 (w/o Ref) drops significantly when the number of steps falls below 15, whereas PoCoDP3 consistently maintains high performance. With only 5 denoising steps, PoCoDP3 achieves the fastest inference without compromising performance. These results demonstrate that reference guidance can significantly reduce the number of denoising steps and accelerate inference while preserving task success.

To further investigate the impact of reference guidance on the generation process and action diversity, we conduct experiments on the Push-T task proposed in [3]. Specifically, we analyze the average distance between intermediate and target actions, as well as the average cosine similarity between action update directions and target directions during the denoising process, as shown in Fig. 7(b). In addition, Fig. 7(c) illustrates the trajectory generation processes of both DP and reference-guided DP. Experimental results indicate that reference guidance substantially enhances the directionality of the denoising process. Compared with DP, the reference-guided DP approaches the target trajectory earlier and maintains higher directional consistency throughout the generation process. By comparison, the Flow Matching model tends to converge toward the target in a fully deterministic manner. Under the 2-step DDIM setting, the unguided model fails to generate valid trajectories, whereas the reference-guided DP still produces reasonable and convergent action sequences, validating its effectiveness in improving both generation quality and inference efficiency. Furthermore, we randomly generate 100 trajectories under the same observation, as shown in Fig. 8. The results show that, due to its fully deterministic nature and the lack of intrinsic stochasticity during sampling, Flow Matching may exhibit reduced trajectory diversity. In contrast, the reference-guided DP maintains stronger deterministic prediction capability while still modeling

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

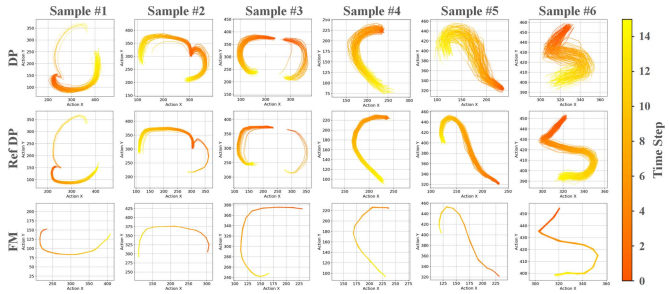

Fig. 8. **Action Diversity:** In Push-T task, at the given state, 100 trajectories are generated. Reference-guided DP achieves more deterministic trajectory predictions while preserving the capability to model multi-modal actions, whereas FM [19] exhibits reduced diversity in action generation.

multimodal action distributions, thereby achieving a favorable balance between determinism and diversity.

V. Conclusion

In this letter, we present PoCoDP3, a pose- and contact–aware visual-tactile policy for contact-rich 3D manipulation. Compared to other existing methods, PoCoDP3 more effectively captures fine-grained in-hand interactions by leveraging pose-aware structured tactile representations and contact-driven cross-modal fusion. Furthermore, the proposed reference-guided diffusion mechanism significantly improves inference efficiency, enabling fast and high-quality action generation. Extensive experiments across simulation and real-world settings demonstrate the superior performance of PoCoDP3 in contact-rich tasks.

While PoCoDP3 achieves strong performance on contact-rich manipulation tasks, it currently assumes well-defined object poses and is therefore primarily applicable to rigid-object manipulation. Extending the framework to deformable or soft objects will require implicit or deformation-aware tactile representations, which will be investigated in future work.

References

[1] Y. Ze, G. Zhang, K. Zhang, C. Hu, M. Wang, and H. Xu, "3D diffusion policy: Generalizable visuomotor policy learning via simple 3D representations," in *Proc. Robot.: Sci. Syst.*, 2024.

[2] C. Wang, H. Fang, H.-S. Fang, and C. Lu, "RISE: 3D perception makes real-world robot imitation simple and effective," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2024, pp. 2870–2877.

[3] C. Chi et al., "Diffusion policy: Visuomotor policy learning via action diffusion," in *Proc. Robot.: Sci. Syst.*, 2023.

[4] C. Zhang et al., "Gelstereo 2.0: An improved GelStereo sensor with multimedium refractive stereo calibration," *IEEE Trans. Ind. Electron.*, vol. 71, no. 7, pp. 7452–7462, Jul. 2024.

[5] H. Xue et al., "Reactive diffusion policy: Slow-fast visual-tactile policy learning for contact-rich manipulation," in *Proc. Robot.: Sci. Syst.*, 2025.

[6] Z. Xu et al., "UniT: Data efficient tactile representation with generalization to unseen objects," *IEEE Robot. Automat. Lett.*, vol. 10, no. 6, pp. 5481–5488, Jun. 2025.

[7] J. Liu, Q. Wang, H. Fan, Y. Wang, Y. Tang, and L. Qu, "Residual denoising diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 2773–2783.

[8] D. Ma, S. Dong, and A. Rodriguez, "Extrinsic contact sensing with relative-motion tracking from distributed tactile measurements," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2021, pp. 11262–11268.

[9] S. Kim and A. Rodriguez, "Active extrinsic contact sensing: Application to general peg-in-hole insertion," in *Proc. Int. Conf. Robot. Automat.*, 2022, pp. 10241–10247.

[10] M. Bauza, A. Bronars, and A. Rodriguez, "Tac2Pose: Tactile object pose estimation from the first touch," *Int. J. Robot. Res.*, vol. 42, no. 13, pp. 1185–1209, 2023.

[11] S. Dikhale et al., "VisuoTactile 6D pose estimation of an in-hand object using vision and tactile sensor data," *IEEE Robot. Automat. Lett.*, vol. 7, no. 2, pp. 2148–2155, Apr. 2022.

[12] J. Lee and N. Fazeli, "ViTaSCOPE: Visuo-tactile implicit representation for in-hand pose and extrinsic contact estimation," in *Proc. Robot.: Sci. Syst.*, 2025.

[13] A. Liu, K. Qian, B. Duan, and S. Luo, "Dynamic contrastive Koopman operator for tactile servo of deformable linear objects," *IEEE Trans. Ind. Electron.*, vol. 72, no. 12, pp. 13716–13728, Dec. 2025.

[14] C. Sferrazza, Y. Seo, H. Liu, Y. Lee, and P. Abbeel, "The power of the senses: Generalizable manipulation from vision and touch through masked multimodal learning," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2024, pp. 9698–9705.

[15] B. Duan, K. Qian, A. Liu, and S. Luo, "Visual–tactile learning of robotic cable-in-duct installation skills," *Automat. Construction*, vol. 170, 2025, Art. no. 105905.

[16] Q. Wu, H. Wang, J. Zhou, X. Xiong, and Y. Lou, "TARS: Tactile affordance in robot synesthesia for dexterous manipulation," *IEEE Robot. Automat. Lett.*, vol. 10, no. 1, pp. 327–334, Jan. 2025.

[17] K. Yu, Y. Han, Q. Wang, V. Saxena, D. Xu, and Y. Zhao, "MimicTouch: Leveraging multi-modal human tactile demonstrations for contact-rich manipulation," in *Proc. 8th Annu. Conf. Robot Learn.*, 2024, pp. 4844–4865.

[18] B. Huang, Y. Wang, X. Yang, Y. Luo, and Y. Li, "3D-ViTac: Learning fine-grained manipulation with visuo-tactile sensing," in *Proc. 8th Annu. Conf. Robot Learn.*, 2024, pp. 2557–2578.

[19] F. Zhang and M. Gienger, "Affordance-based robot manipulation with flow matching," 2024, *arXiv:2409.01083*.

[20] A. Prasad, K. Lin, J. Wu, L. Zhou, and J. Bohg, "Consistency policy: Accelerated visuomotor policies via consistency distillation," in *Proc. Robot.: Sci. Syst.*, 2024.

[21] W. Chen, J. Xu, F. Xiang, X. Yuan, H. Su, and R. Chen, "General-purpose Sim2Real protocol for learning contact-rich manipulation with marker-based visuotactile sensors," *IEEE Trans. Robot.*, vol. 40, pp. 1509–1526, 2024.

[22] M. J. Zeestraten, I. Havoutis, J. Silvério, S. Calinon, and D. G. Caldwell, "An approach for imitation learning on riemannian manifolds," *IEEE Robot. Automat. Lett.*, vol. 2, no. 3, pp. 1240–1247, Jul. 2017.

[23] A. Mandlekar et al., "What matters in learning from offline human demonstrations for robot manipulation," in *Proc. 5th Annu. Conf. Robot Learn.*, 2021, pp. 1678–1690.

[24] A. Mandlekar et al., "MimicGen: A data generation system for scalable robot learning using human demonstrations," in *Proc. 7th Annu. Conf. Robot Learn.*, 2023, pp. 1820–1864.

[25] Y. Zhao, K. Qian, B. Duan, and S. Luo, "FOTS: A fast optical tactile simulator for Sim2Real learning of tactile-motor robot manipulation skills," *IEEE Robot. Automat. Lett.*, vol. 9, no. 6, pp. 5647–5654, Jun. 2024.