

Semantic-Contact Fields for Category-Level Generalizable Tactile Tool Manipulation

Kevin Yuchen Ma^{1,2}, Heng Zhang^{1,3}, Weisi Lin³, Mike Zheng Shou^{*2}, and Yan Wu^{*1}

¹Institute for Infocomm Research, A*STAR ²Show Lab, National University of Singapore

³College of Computing and Data Science, Nanyang Technological University

*Corresponding authors

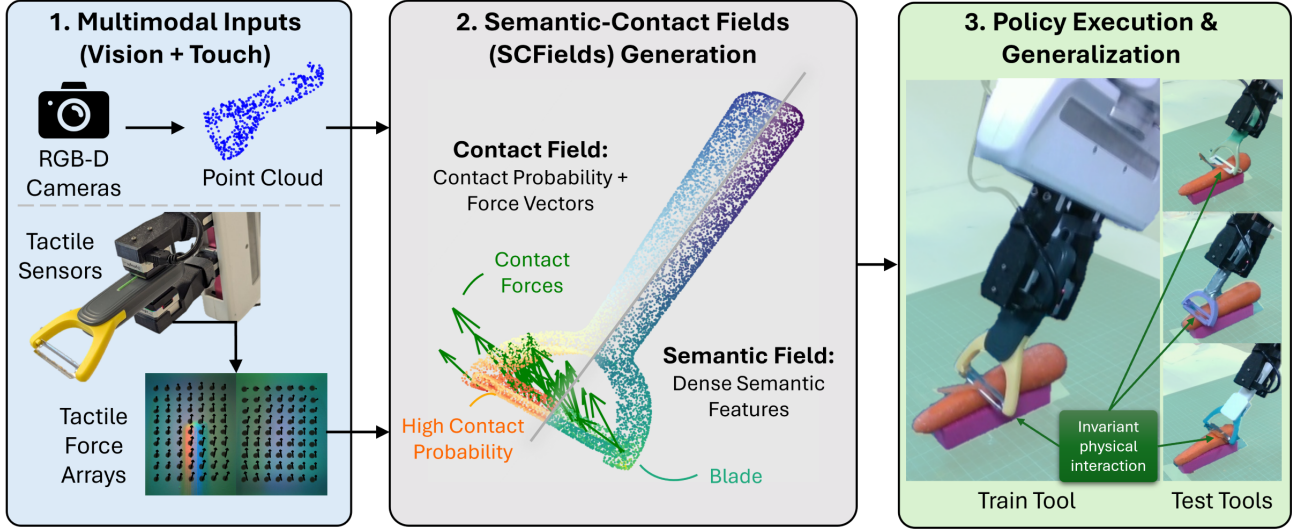


Fig. 1: **Semantic-Contact Fields (SCFields) Overview.** **1. Multimodal Inputs:** The system takes RGB-D observations and tactile readings from GelSight sensors. **2. SCFields Generation:** Our unified perception module fuses these inputs into a dense point cloud representation containing both category-level semantics (blue/green heatmap) and extrinsic contact force vectors (green arrows). **3. Policy Execution:** A diffusion policy conditioned on the SCFields enables zero-shot generalization to novel tools variants (e.g., peelers of different shapes) in contact-rich tasks by reasoning about functional affordance and contact forces simultaneously.

Abstract—Generalizing tool manipulation requires both semantic planning and precise physical control. Modern generalist robot policies, such as Vision-Language-Action (VLA) models, often lack the physical grounding required for contact-rich tool manipulation. Conversely, existing contact-aware policies that leverage tactile or haptic sensing are typically instance-specific and fail to generalize across diverse tool geometries. Bridging this gap requires learning representations that are both semantically transferable and physically grounded, yet a fundamental barrier remains: diverse real-world tactile data are prohibitive to collect at scale, while direct zero-shot sim-to-real transfer is challenging due to the complex nonlinear deformation of soft tactile sensors.

To address this, we propose Semantic-Contact Fields (SCFields), a unified 3D representation that fuses visual semantics with dense extrinsic contact estimates, including contact probability and force. SCFields is learned through a two-stage Sim-to-Real Contact Learning Pipeline: we first pre-train on large-scale simulation to learn geometry-aware contact priors, then fine-tune on a small set of real data pseudo-labeled via geometric heuristics and force optimization to align real tactile signals. The resulting force-aware representation serves as the dense observation input to a diffusion policy, enabling physical generalization to

unseen tool instances. Experiments on scraping, crayon drawing, and peeling demonstrate robust category-level generalization, significantly outperforming vision-only and raw-tactile baselines. Project page: <https://kevinskww.github.io/SCFields>.

I. INTRODUCTION

Tool use represents a hallmark of intelligence, extending a robot’s physical capabilities beyond its own embodiment. However, achieving robust category-level generalization in tool manipulation remains challenging. Effectively manipulating diverse tool variants requires a dual understanding: a semantic grasp of *where* to hold and apply the tool (functional affordance) and a physical mastery of *how* to regulate interaction forces (contact dynamics).

While recent advances in large-scale robotic learning have produced generalist policies capable of interpreting high-level semantic commands, these vision-centric models remain physically naive. Methods relying solely on visual or 3D semantic representations, such as GenDP [29], can generalize geometrically across tool variants, but fail in contact-rich tasks that are visually ambiguous and require precise force regulation.

Conversely, policies that leverage tactile or haptic sensing [2, 18] are adept at managing local contact, but are typically instance-specific. Because they map tactile signals directly to actions without an intermediate generalized representation, they fail to adapt when the tool’s geometry changes.

Our approach is grounded in a key physical insight: while the global geometry of tools within a category varies significantly, the physical interaction at the “effective part”—such as the blade of a peeler—remains invariant. Therefore, an explicit representation of *extrinsic contact*—the contact between the tool and the environment—can provide a geometry-aware physical abstraction for transferring manipulation skills across diverse tool instances. A promising candidate for this is the contact field [7], which maps tactile feedback onto the tool’s surface. However, for contact-rich tool use, contact localization alone is insufficient: the policy must also infer the magnitude and direction of interaction forces needed for regulation. Learning such representations also presents a dilemma: collecting diverse real-world tactile data to cover all geometries is prohibitively expensive, yet training entirely in simulation introduces a severe Sim-to-Real gap. Simulating the non-linear deformation of soft tactile sensors is notoriously difficult, leading models trained solely in simulation to hallucinate phantom forces or miss contacts entirely when deployed.

To bridge this gap, we propose **Semantic-Contact Fields (SCFields)** (Figure 1), a unified force-aware 3D representation trained via a two-stage Sim-to-Real framework. We decompose the contact estimation problem into learning *general geometry and interaction physics* (invariant contact distributions) and *real-world sensory alignment* (sensor-specific signal interpretation). Accordingly, our pipeline operates in two distinct stages. First, we leverage large-scale simulation across diverse tool geometries to learn geometry-aware contact priors for the invariant contact physics. Second, to address the reality gap, we introduce a **Real-World Alignment** stage. We generate pseudo-labels using geometric heuristics and analytical methods from a small set of simple real-world interactions. This alignment step adapts the simulation-trained model to interpret real sensor responses while preserving the generalizable physics learned in simulation. Importantly, the data used for this alignment stage is collected concurrently with the demonstrations used for imitation learning policy training, meaning no additional, separate data collection effort is required.

Our specific contributions are as follows:

- We propose **Semantic-Contact Fields (SCFields)**, an invariant 3D representation that fuses semantic features from pre-trained vision models with dense extrinsic contact probability and force estimates.
- We introduce a two-stage **Sim-to-Real training pipeline** that combines the scalability of simulation with the fidelity of real-world data. By pre-training on diverse simulated tools and fine-tuning with heuristic-labeled real data, we achieve robust dense contact estimates generalizable to unseen tool variants.
- We evaluate our approach on a Franka Panda robot with

Gelsight Mini tactile sensors across scraping, crayon drawing, and peeling, demonstrating category-level generalization to unseen tool instances and environmental variations.

II. RELATED WORK

A. Generalizable 3D Manipulation and Tool Use

A key limitation of 2D image-based policies is their sensitivity to viewpoint changes. This has led to a surge in 3D-centric policies that operate directly on point clouds or neural fields [33, 5, 11, 35], providing improved spatial reasoning. To enable category-level generalization, recent works have adopted semantic-centric representations. Methods such as D3Field [28], GenDP [29], and S²-Diffusion [31] use spatial or 3D semantic representations to improve category-level generalization. While these methods achieve strong geometric generalization, they remain purely vision-based and physically naive, struggling in contact-rich tasks where visual semantics alone are insufficient.

In tool manipulation, generalization requires understanding both tool geometry and functional interaction. Prior works address this by learning structured correspondences between tool instances. Some approaches focus on dense alignment, predicting motion fields or 6D functional poses to transfer manipulation motions across tools [22, 30]. Other methods utilize sparser representations to enable one-shot skill transfer, modeling tools via functional keypoints [25, 26] or leveraging vision-language models to identify affordance regions [24, 16]. While effective at identifying *where* the functional part of a tool is, these methods often do not model *how* contact should be regulated during execution, such as the force direction and magnitude required to peel or scrape.

Attempts to bridge this gap by combining vision and touch have yielded promising but limited results. 3D-ViTac [10] integrates tactile readings as occupancy points into a scene point cloud, providing a useful geometric fusion mechanism, but not an explicit estimate of dynamic extrinsic force on the tool surface. Other works rely on wrist force-torque sensors for compliance [18, 6, 17]. However, for tool manipulation, wrist-mounted sensing provides only a net wrench at the robot wrist; inferring the critical distributed contact location and force at the tool tip remains ambiguous because of tool leverage, grasp variation, and external contacts. Our work addresses this gap by representing extrinsic contact directly on the tool point cloud, combining semantic invariance with per-point contact probability and force estimates.

B. Tactile Perception and Sim-to-Real Transfer

The primary challenge in leveraging the tactile modality is converting high-dimensional, noisy sensor data into a useful representation. Vision-based sensors like GelSight [32] provide rich, high-resolution topography, while taxel-based sensors offer direct force maps. Recent advances in tactile representation learning, such as T3 [34] and Sparsh [9], utilize self-supervision to learn robust encoders. However, these representations typically encode intrinsic tactile observations

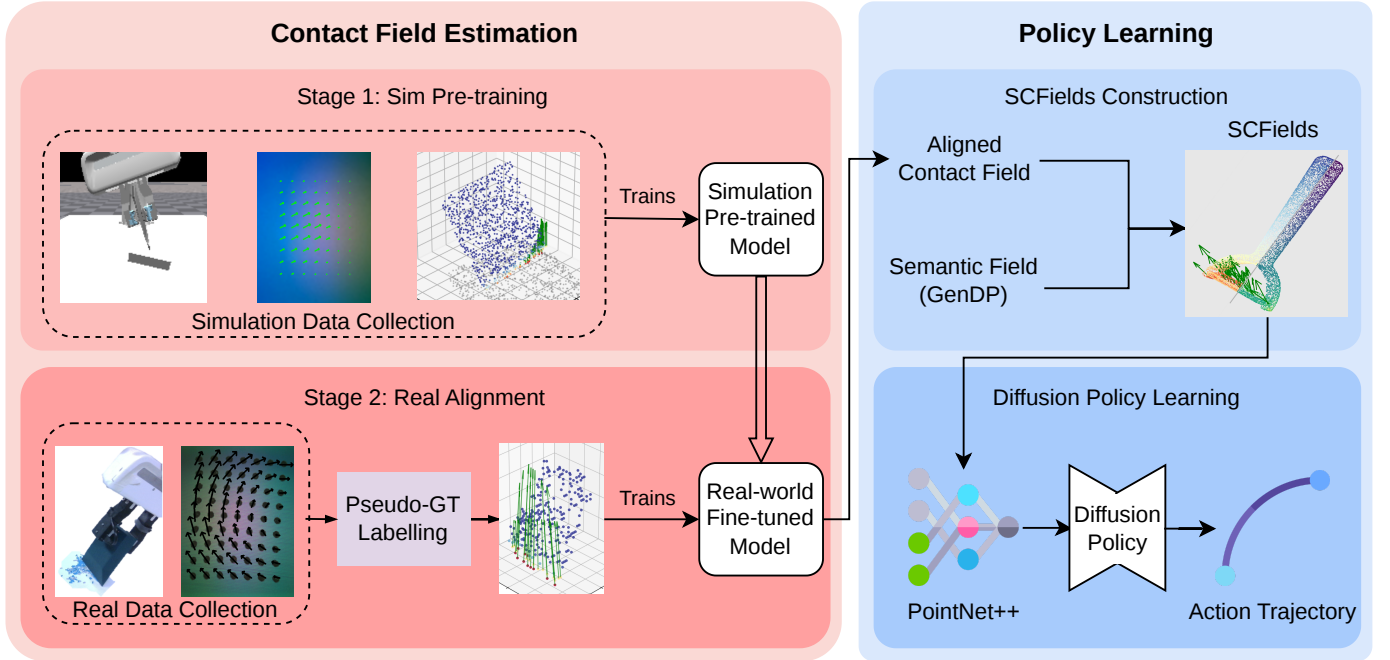


Fig. 2: **Method Overview.** **Left:** Contact Field Learning (III-B) Stage 1 learns general geometry-aware contact priors from simulated data; Stage 2 aligns the model to real tactile sensor responses using pseudo-labeled real data. **Right:** Policy Learning (III-C) A Diffusion Policy is conditioned on the resulting SCFields to achieve robust tool manipulation.

at the sensor surface, rather than grounding contact on the external tool geometry. This limits their direct use for tool manipulation, where the policy must reason about where and how the tool contacts the environment.

Tactile manipulation is further hindered by the significant gap between simulated and real tactile physics. General tactile simulators [27, 23, 1] have made strides in modeling sensor deformation and optical properties. However, accurately capturing fine-grained contact mechanics—such as friction coefficients, hysteresis, and soft-body dynamics—remains computationally intensive and difficult to calibrate. As a result, policies trained purely in simulation [3, 15, 1] often struggle to generalize to the unmodeled physical variations of the real world. To address this, we use simulation to learn geometry-aware contact priors, then fine-tune the contact estimator on a small set of pseudo-labeled real data to align real tactile sensor responses without requiring high-fidelity tactile simulation.

C. Extrinsic Contact Estimation

To manipulate tools effectively, a robot must understand *extrinsic contact*—the interaction between the tool and the environment. Prior work has attempted to estimate this property through various means. Vision-only approaches [12] can infer likely contact regions but remain ambiguous without tactile feedback, while analytical methods [19] often rely on known object geometries and specific exploratory motions.

The most promising recent direction involves learning implicit contact representations. Neural Contact Fields (NCF) [7, 8] learn a continuous function mapping surface coordinates to contact probabilities. However, NCF assumes a fixed in-hand pose of the tool relative to the sensor, which breaks down during dynamic manipulation where grasp adaptation

occurs. More recent work like VitaScope [13] relaxes this assumption by jointly estimating the tool in-hand pose and the extrinsic contact. However, VitaScope requires a known tool mesh, making category-level generalization to previously unseen tool geometries difficult. Furthermore, these methods typically model contact probability via geometric proximity, ignoring the magnitude of contact forces. In contrast, SCFields represents extrinsic contact directly on the observed tool point cloud and predicts both contact probability and force, enabling force-aware manipulation of novel tool instances without assuming a fixed in-hand pose or known mesh.

III. METHODS

Our goal is to learn a manipulation policy that generalizes to unseen tool variants by leveraging contact-rich dynamics. To achieve this, we propose a two-stage method (Figure 2):

- 1) **Contact Field Learning:** We train a multimodal perception model f_ϕ to estimate a dense Extrinsic Contact Field on the tool surface. This model is pre-trained in simulation to learn geometry-aware contact priors and fine-tuned on real-world data for domain alignment.
- 2) **Policy Learning:** We construct a unified state representation by fusing these estimated contact probabilities and forces with 3D Semantic Fields [29]. This fused representation conditions a diffusion policy π_θ capable of zero-shot transfer to novel tool instances.

A. Problem Formulation

We formulate the system as two distinct learning problems:

1. Extrinsic Contact Field Estimation: We learn a perception mapping f_ϕ that transforms raw observations—tool and environment point clouds ($P_{\text{tool}}, P_{\text{env}}$), tactile sensor readings

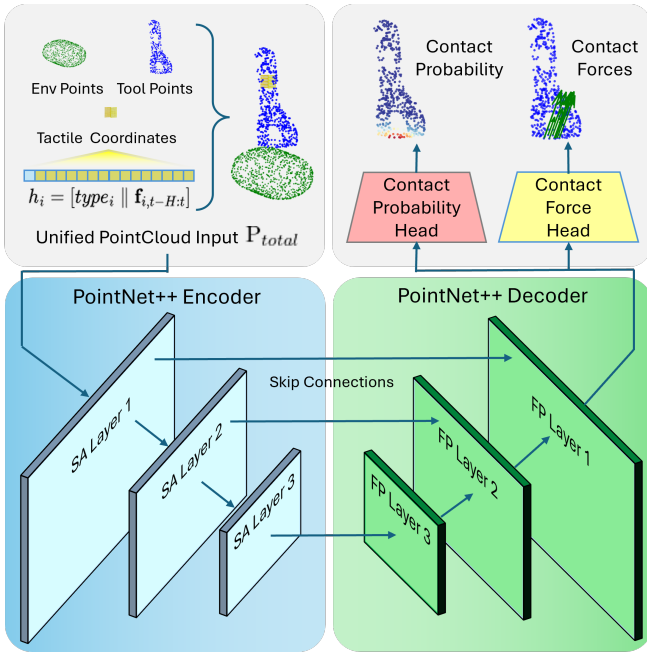


Fig. 3: Contact field model architecture. The network fuses tactile markers and force arrays with dense object geometry in a unified point cloud input to predict contact fields.

(T), and proprioceptive state (q)—into a dense extrinsic contact field F_c over the tool surface. This field assigns a contact probability c_i and a 3D force vector $\mathbf{f}_i$ to every point p_i in the tool point cloud P_{tool} . We predict both quantities because they serve different roles: c_i localizes the plausible support of tool-environment contact, while $\mathbf{f}_i$ encodes the local interaction direction and magnitude needed for force regulation.

2. Generalizable Policy Learning: We treat manipulation as conditional generation. We learn a policy $\pi_\theta(a_{t:t+H}|O_t)$ that predicts a sequence of actions a given observation O_t . The core challenge is designing an O_t that is invariant to instance-specific geometry while retaining high-fidelity physical feedback.

B. Contact Field Estimation

1) Model Architecture: Unlike prior works that process vision and touch with separate encoders [7], we employ a unified **Tactile-as-PointCloud** architecture similar to 3D-ViTac [10]. We treat tactile signals as 3D geometric entities, fusing them directly into the scene geometry using a PointNet++ [21] framework (Figure 3).

Unified Input Representation. The input is a composite point cloud $P_{\text{total}} = P_{\text{obj}} \cup P_{\text{env}} \cup P_{\text{tactile}}$, where P_{obj} and P_{env} represent the sampled tool and environment surfaces, and P_{tactile} represents the 3D coordinates of the tactile sensor markers projected into the world frame.

Each point $p_i \in P_{\text{total}}$ is augmented with a feature vector $h_i = [\text{type}_i \parallel \mathbf{f}_{i,t-H:t}]$. Here, type_i encodes the source (Object, Env, Tactile), and $\mathbf{f}_{i,t-H:t}$ encodes a H -step history of marker displacements. This representation allows the network to implicitly learn the relationship between tool geometry,

sensor deformation, and contact location without a dedicated pose encoder.

Network Architecture. We process P_{total} using a standard PointNet++ encoder-decoder. The encoder fuses sparse tactile signals with dense tool geometry based on spatial proximity. The decoder upsamples the features back to the resolution of P_{obj} , effectively propagating localized sensor information to the entire tool surface. Two parallel heads then predict the scalar contact probability $c_i \in [0, 1]$ and regress the extrinsic force vector $\mathbf{f}_i^{\text{ext}} \in \mathbb{R}^3$ for each point. Detailed layer configurations are provided in Appendix A.

2) Simulation Data Generation: Training requires dense ground-truth (GT) labels for contact locations and forces, which are physically inaccessible in the real world. We address this by generating a large-scale synthetic dataset (300 tool instances, 320,000 frames) using a multi-simulator pipeline, inspired by [13].

Simulation Pipeline. We construct a simulation environment comprising a Franka Emika Panda robot and diverse procedurally generated tools (scrapers, crayons, peelers). Data collection proceeds in two steps:

- 1) **Interaction (IsaacGym + TacSL):** We use IsaacGym [20] for high-throughput rigid-body dynamics and TacSL [1] to simulate the specific force field and depth data of GelSight sensors.
- 2) **Labeling (Open3D + PyBullet):** Since accurate contact locations and forces are not directly accessible in IsaacGym, we employ a replay strategy. We replicate the scene in Open3D to compute signed distance functions (SDF) for soft contact probability, and replay interactions in PyBullet to extract discrete contact forces, which are then extrapolated to the dense point cloud.

This pipeline produces dense contact labels for 300 unique tool geometries, providing broad geometric and contact variation for pre-training. Further details on the replay and extrapolation logic are in Appendix B.

3) Real-World Pseudo-GT Generation: Despite filtering and calibration, a significant gap remains between simulated tactile fields and real marker displacements. Furthermore, obtaining dense ground-truth extrinsic contact fields directly in the real world is extremely difficult, as it would require instrumenting the entire surface of arbitrary tools with high-resolution force sensors. To bridge these challenges, we introduce a **Real-World Alignment** stage. We collect a small real-world alignment dataset and generate pseudo-ground-truth labels using geometric heuristics and analytical force optimization.

Heuristic Contact Probability. We constrain data collection to a structured task: scraping a flat surface with known table height z_{table} . We first identify geometrically plausible contact candidates $C_{\text{candidate}} \subset P_{\text{obj}}$ using a height threshold ($p_z < z_{\text{table}} + \epsilon$). To eliminate false positives (e.g., the tool hovering near the surface without touching), we apply a signal-based gating filter. We compute the mean magnitude of tactile marker displacements relative to the initial undeformed frame.

A frame is labeled as "in contact" only if this mean delta signal exceeds a calibrated noise threshold. For these valid frames, points in $C_{candidate}$ are assigned soft contact probability labels $c_i \in [0, 1]$ inversely proportional to their distance from the table surface.

Analytical Contact Force Optimization. To estimate dense force vectors $\mathbf{f}_i^{ext}$ without ground-truth from external force sensors, we solve a convex optimization problem that explains the observed tactile net wrench $\mathbf{W}_{tac}$ using a distribution of point forces $\mathbf{f}$ at candidate contact points. We formulate this as a Second-Order Cone Program (SOCP):

$$\min_{\mathbf{f}} \quad \|\mathbf{G}\mathbf{f} - \mathbf{W}_{tac}\|_2^2 + \lambda \sum_{i \in C_{candidate}} \frac{\|\mathbf{f}_i\|_2^2}{c_i + \epsilon} \quad (1)$$

$$\text{s.t.} \quad \|\mathbf{f}_i\|_2 \leq 2(\mathbf{f}_i \cdot \mathbf{n}_i) \quad \forall i \quad (2)$$

where $\mathbf{G}$ is the grasp matrix mapping point forces to the gripper frame. The objective minimizes wrench discrepancy while regularizing force magnitudes inversely to their heuristic contact probability c_i , favoring geometrically likely contact points. The constraint enforces physical plausibility by bounding force magnitude by twice its projection onto the inward normal $\mathbf{n}_i$. This ensures forces are compressive and lie within a $\sim 60^\circ$ friction cone, preventing unrealistic "pulling" or shear. The problem is solved efficiently using the ECOS solver [4].

4) *Two-Stage Model Training:* We employ a two-stage strategy to transfer physical priors from simulation to the real world. Both stages minimize a composite loss $\mathcal{L}_{total} = \lambda_{prob}\mathcal{L}_{prob} + \lambda_{force}\mathcal{L}_{force}$. We use Focal Loss [14] for $\mathcal{L}_{prob}$ to handle class imbalance, and a combination of Adaptive Weighted MSE (magnitude) and Cosine Similarity (direction) for $\mathcal{L}_{force}$. Detailed training and loss function hyperparameters are presented in Appendix C.

Stage 1: Sim Pre-training. The model is first trained on the large-scale simulation dataset with extensive domain randomization. This establishes the fundamental mapping between tool geometry and force distribution across a wide range of tool variants and interaction poses.

Stage 2: Real-World Alignment. We fine-tune the model on the pseudo-labeled real-world dataset. Since this real-world set is small and collected under constrained conditions, we apply random translation and rotation augmentations to the input point clouds during training to enhance data diversity, preventing overfitting to the specific collection poses and ensuring that the learned sensor alignment generalizes robustly to varied spatial configurations. We use a reduced learning rate to adapt to the real sensor characteristics while preserving the general geometry-aware contact priors learned in simulation.

C. Semantic-Contact Fields for Policy Learning

We construct **Semantic-Contact Fields (SCFields)** as a unified 3D observation that fuses semantic information with aligned extrinsic contact estimates. The policy observation s_t is a dense feature field over the object point cloud P_{obj} . Each point p_i carries a feature vector $x_i = [\mathbf{f}_i^{ext} \parallel c_i \parallel S_i]$, where:

- $\mathbf{f}_i^{ext}, c_i$: **Contact Field** with contact force and probability estimates from our fine-tuned contact estimator.

- S_i : **3D Semantic Fields** adapted from [29] extracted from a pre-trained vision backbone, capturing functional semantics such as "blade" and "handle", providing geometric invariance.

We implement the manipulation policy π_θ using a 3D Diffusion Policy framework based on [29]. The dense SCFields point cloud $\{p_i, x_i\}$ is first processed by a PointNet++ backbone to extract a global feature vector that aggregates both semantic and physical information. This feature vector is then fed into the diffusion policy's denoising network as the conditioning input. During inference, the policy iteratively denoises Gaussian noise into a sequence of end-effector actions $a_{t:t+H}$, conditioned on both functional semantics and force-aware contact estimates from SCFields.

IV. EXPERIMENTS

We design our experiments to evaluate two main hypotheses: (1) Does our unified *Tactile-as-PointCloud* architecture, combined with real-world alignment, produce accurate contact estimates that generalize to unseen tools? (2) Does the resulting **Semantic-Contact Fields (SCFields)** enable a diffusion policy to perform contact-rich manipulation tasks that are robust to environmental variations and novel tool instances?

A. Contact Field Model Evaluation

We first validate the contact field perception module (f_ϕ) in isolation to ensure it provides reliable contact probability and force vector estimates before integrating it into the policy loop. For contact-field metrics, we report aggregate performance in the main text and provide confidence intervals and significance tests in Appendix F.

Models Compared. To isolate the contributions of our architecture and training pipeline, we categorize comparisons as follows:

1. Sim-to-Real Training Strategies (Validating the Pipeline):

- **Ours (Aligned):** Pre-trained on large-scale simulation, then fine-tuned on the small real-world scraper dataset.
- **Sim-Only:** Trained exclusively on simulation data. Evaluates zero-shot transfer and the magnitude of the sim-to-real gap.
- **Real-Only:** Trained from scratch on the small real-world scraper dataset. Tests if physical priors from simulation are necessary given limited real data.

2. Baselines & Ablations (Validating the Architecture):

- **No-Tactile:** Identical architecture with tactile marker features masked out. This baseline relies only on observed tool/environment geometric (P_{obj}, P_{env}), testing the necessity of dynamic tactile feedback.
- **NCF (Neural Contact Fields) [7]:** A baseline implicit representation that predicts contact probabilities but lacks explicit force vector regression.
- **Ablation - 2D Tactile Encoder:** Replaces our point-cloud fusion with a standard CNN encoder for the tactile force arrays, concatenated with the global point cloud feature. This tests the benefit of our *Tactile-as-PointCloud* fusion strategy.

TABLE I: Sim Evaluation: Architecture Capacity

Model	F1 Score $\uparrow$	Force MSE $\downarrow$
NCF [7]	0.043	N/A
No-Tactile	0.539	0.0146
Ablation - 2D Tactile Enc.	0.531	0.0147
Ablation - BCE Loss	0.123	0.0146
Ablation - No Cont. Prob.	N/A	0.0158
Ours (Sim-Only)	0.587	0.0147

TABLE II: Real-World Evaluation: Alignment & Generalization

Model	Scrapers (Seen Class)		Crayons (Unseen in Real)	
	F1 Score $\uparrow$	Force MSE $\downarrow$	F1 Score $\uparrow$	Force MSE $\downarrow$
Sim-Only	0.002	0.0435	0.008	0.0284
Real-Only	0.458	0.0221	0.614	0.0106
No-Tactile	0.411	0.0432	0.552	0.0115
No Cont. Prob.	N/A	0.0377	N/A	0.0089
Ours (Aligned)	0.534	0.0254	0.657	0.0085



Fig. 4: **Left:** Real robot experiment setup: We use a Franka Emika Panda robot with 2 GelSight Mini tactile sensors mounted on the gripper fingers, and 3 RealSense D435 cameras to capture RGB-D observations. **Right:** Training and Testing Tools

- **Ablation - Loss Function:** Replaces Focal Loss with standard BCE Loss to evaluate robustness to class imbalance.

1) *Evaluation 1: Architecture Validation (Sim-to-Sim):* We first verify the architecture’s capacity to learn complex contact physics using a held-out test set from our simulation dataset. This controlled setting allows us to compare architectural choices without domain shift noise. We report the **F1 Score** for binary contact detection and the **Mean Squared Error (MSE)** for force vector regression.

As shown in Table I, our **Tactile-as-PointCloud** architecture trained with Focal Loss achieves the best contact F1, supporting the importance of Focal Loss and preserving the 3D spatial structure of tactile markers under severe contact-class imbalance. NCF performs poorly, likely because its fixed-pose formulation is sensitive to varying in-hand tool poses. The Force MSE values are similar for several models in simulation, suggesting that force regression in the synthetic setting is less discriminative due to limited simulated tactile/contact fidelity. However, removing the contact probability head increases Force MSE, indicating that contact probability provides a useful spatial support signal for force regression.

2) *Evaluation 2: Real-World Alignment Accuracy:* We assess the efficacy of our pipeline on real-world data. We evaluate on a held-out set of **Scrapers** (Seen in Alignment) and a set of **Crayons** (Unseen in Alignment, but Seen in Sim). Ground truth is generated via our heuristic labeling pipeline.

Table II highlights the severity of the sim-to-real gap: the **Sim-Only** model fails almost completely on real tactile

inputs. The **Real-Only** model fits the seen scraper alignment data but generalizes less reliably to crayons, indicating that limited real data alone does not provide sufficient geometric contact priors. The **No-Tactile** baseline detects some contact from geometry but has substantially worse force prediction, showing that geometric proximity is insufficient for loaded contact estimation. In contrast, **Ours (Aligned)** transfers the sensor alignment learned from scraper interactions to crayons, supporting the separation between simulation-learned contact priors and real tactile-signal alignment. The **No Contact Probability** ablation further clarifies the role of the probability head. Without contact probability, force estimation degrades substantially on real scrapers, where contact is distributed along an edge, but changes little on crayons, where contact is closer to point-like. This suggests that contact probability is most useful as a spatial support estimator for extended or ambiguous contact regions, rather than as a replacement for force prediction.

3) *Evaluation 3: Qualitative Generalization:* Finally, we qualitatively evaluate generalization to the **Peeler** task, where complex interactions with irregular carrots make heuristic ground-truth labeling unreliable. For this task, the model is pre-trained on simulated peeler data but fine-tuned using **only** the real-world scraper dataset. As shown in Figure 5, Sim-Only misses real contact due to the tactile domain gap, Real-Only produces noisy forces without sufficient geometric priors, and No-Tactile fails to infer loaded blade contact from geometry alone. In contrast, Ours produces localized force estimates at the blade-carrot interface, suggesting that scraper-based real

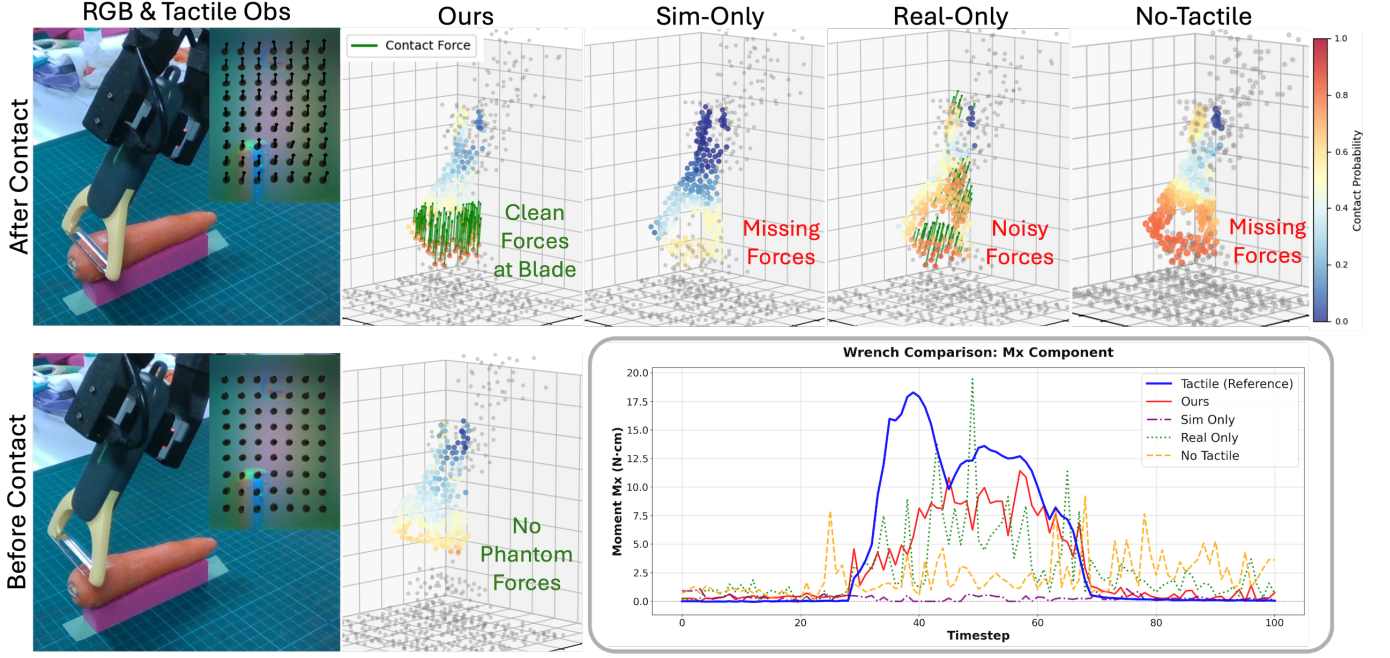


Fig. 5: Qualitative comparison of contact-field estimation on the Peeler. **Top Row:** Ours produces clean contact forces localized on the blade-carrot interface, while Sim-Only and No-Tactile miss forces, Real-Only predicts noisy forces. **Bottom Right:** Correlation between torque induced by predicted contact forces and the reference wrench from tactile signals. Ours aligns best with the reference wrench, while Real-Only and No-Tactile remain noisy.

alignment can transfer to more complex curved tools when simulation provides the relevant contact prior.

B. Policy Evaluation

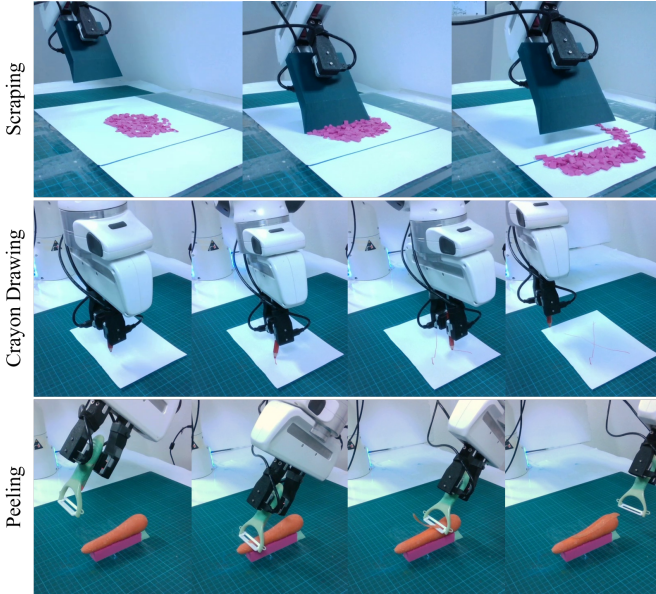


Fig. 6: Rollouts of contact-rich tasks with unseen tools. **Top:** Scraping debris past a target line. **Middle:** Drawing a cross with consistent force. **Bottom:** Peeling a carrot. Additional experiment visualizations are available in Appendix E. **Experimental Setup.** We evaluate policy performance on a real-world Franka Emika Panda robot. The robot is equipped

with a parallel gripper modified to house two GelSight Mini tactile sensors. Visual observations are captured via three calibrated RealSense D435 cameras (front, left, right). We evaluate the Diffusion Policy conditioned on SCFields across three contact-rich tasks (Figure 6).

1) *Baselines and Ablations:* We compare SCFields against the **Vision-Only** baseline based on GenDP [29], and a **Raw Tactile (End-to-End)** baseline, which concatenates raw tactile data directly into the policy observation without explicit physics supervision. Additionally, we evaluate the **Sim-Only Contact Field** and **Real-Only Contact Field** baselines defined in Section IV-A. Finally, a **No Contact Force** ablation isolates explicit contact force vectors by training the policy using only contact probability.

2) *Tasks and Metrics:* **Task 1: Scraper (Contact-Rich Cleaning).** The robot must maintain surface contact to clean debris. We trained the policy on 150 demonstration episodes collected across 3 table heights using 4 training tools. We evaluate on both seen and 4 unseen tools with 2 unseen table heights and 2 trials each (16 trials in total per split); dynamic stopping criteria yield 46–72 individual scrape attempts per split. Metrics include *Success Rate (SR)* (% of trials maintaining contact) and *Cleaning Efficiency (Eff)* (% of debris removed). An additional *Normalized Efficiency (Eff Norm)* is included to offset the effect of different scraper blade lengths. **Task 2: Crayon Drawing.** Picking up an asymmetric crayon (or pencil) to draw a cross. The policy was trained on 120 episodes across 3 heights using 3 training crayons and was evaluated on 2 unseen heights and 3 trials each (18 trials

TABLE III: Task 1: Scraper Performance. Ours outperforms baselines on unseen tools, demonstrating robust generalization.

Method	Seen Tools			Unseen Tools		
	SR (%)	Eff (%)	Eff Norm (%)	SR (%)	Eff (%)	Eff Norm (%)
Vision-Only (GenDP)	39.1	10.1	30.5	35.1	25.4	35.1
Raw Tactile	35.1	26.9	35.7	50.0	23.3	27.3
Ours (SCFields)	73.5	61.8	85.2	79.6	73.5	84.7
Ablation: Sim-Only CF	34.5	20.6	33.1	55.6	37.0	45.2
Ablation: Real-Only CF	45.8	32.3	44.5	50.0	47.5	54.2
Ablation: No Force	31.3	22.2	29.9	26.0	24.3	27.6

TABLE IV: Task 2: Crayon Drawing Consistency (Score 0-1).

Method	Seen Crayons	Unseen Crayons
Vision-Only (GenDP)	0.81	0.60
Raw Tactile	0.76	0.61
Ours (SCFields)	0.86	0.78
Ablation: Sim-Only CF	0.81	0.60
Ablation: Real-Only CF	0.68	0.74
Ablation: No Force	0.80	0.76

in total per split). Success requires precise force modulation to leave a visible trace without snapping the crayon. Metric: *Drawing Consistency* (0-1 score reflecting completeness/visibility). The preliminary stage of picking up the crayon, which requires visual semantics generalization to unseen crayons/pencils, is reported in Appendix D.

Task 3: Peeler. Peeling a carrot with a handheld peeler. We trained the policy on 60 demonstration episodes using 2 training peelers and evaluated on 20 seen-peeler trials and 30 unseen-peeler trials. This task tests the cross-category generalization of our perception module, as the contact field model was aligned using only scraper data. Metrics: Percentage of Successful *Contact* and *Cut-in*, and *Peel Quality* (avg. peel length).

predictions unreliable. Conversely, while the *Real-Only CF* ablation achieves better performance than other baselines on the seen scraper task, it fails to generalize to novel tools, confirming that simulation pre-training is requisite for learning contact representations that transfer across object categories.

Role of Explicit Force & Generalization. The **No Force** ablation isolates the value of continuous force prediction. Contact probability alone can indicate likely interaction regions, but it cannot distinguish insufficient loading from excessive pressure. As a result, the policy exhibits failure modes such as hovering above the surface or pressing too hard and slipping, showing that force vectors are necessary for regulating interaction dynamics. Similar failure modes are also present in other baseline methods that do not explicitly model contact force, as illustrated in Figure 7. Notably, the Peeler task highlights the robustness of our pipeline: although the policy used imitation learning on peeling demonstrations, the underlying perception model was aligned using only scraper data. Despite this, our model achieves an average peel length of **4.52cm**, quadrupling the Vision-Only baseline (**1.12cm**). This confirms that SCFields successfully transfers the invariant concept of "functional contact" from simulation to novel real-world tools.

V. CONCLUSION

In this work, we presented **Semantic-Contact Fields (SCFields)**, a novel 3D representation that fuses visual semantics with dense, physically-grounded contact estimates to enable category-level generalization of contact-rich tool manipulation. We addressed the fundamental challenge of tactile sim-to-real transfer through a two-stage learning pipeline: pre-training on large-scale physics simulations to learn geometric-aware contact priors, followed by a data-efficient real-world alignment stage. This approach effectively bridges the reality gap without expensive instrumentation or high-fidelity simulation. Experiments on scraping, drawing, and peeling demonstrate that SCFields significantly outperforms baselines; by grounding sparse tactile readings into dense physical estimates, our system achieves zero-shot generalization to novel tool variants and robustness to dynamic environments where traditional methods fail.

A key limitation of the current framework is its reliance on imitation learning, which restricts the robot to tool usage patterns present in the demonstrations. While SCFields enable robust generalization across geometric variants within a category, the system cannot currently discover novel functional

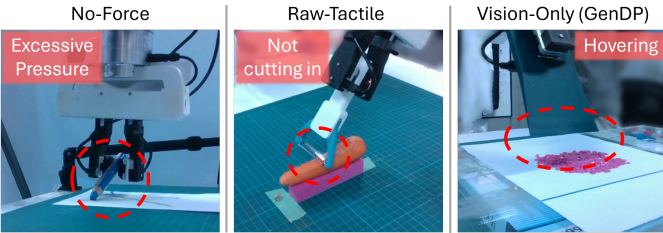


Fig. 7: Example failure modes of baseline/ablation methods.

3) *Analysis:* We present the quantitative results for all three tasks in Tables III, IV, and V. Full experiment statistics including confidence intervals and p -values are reported in Appendix F.

Comparison with Baselines. Across all tasks (Tables III, IV, V), SCFields significantly outperforms baselines, particularly on unseen tools. The *Raw Tactile* baseline struggles (e.g., 23.3% Eff on Unseen Scraper), confirming that without explicit physical grounding, end-to-end policies fail to leverage high-dimensional tactile data effectively, often overfitting to visual inputs. Similarly, the *Sim-Only CF* baseline generally matches Vision-Only performance, underscoring that without our real-world alignment stage, the domain gap renders tactile

TABLE V: Task 3: Peeler Results. Aligned model performance validates the pipeline.

Method	Seen Peelers			Unseen Peelers		
	Contact (%)	Cut-in (%)	Avg Peel Length (cm)	Contact (%)	Cut-in (%)	Avg Peel Length (cm)
Vision-Only	45.0	30.0	1.50	50.0	33.3	1.12
Raw Tactile	45.0	20.0	1.05	40.0	30.0	0.85
Ours (SCField)	80.0	70.0	4.73	90.0	73.3	4.52
Ablation: Sim-Only CF	60.0	30.0	2.00	50.0	46.7	3.05
Ablation: Real-Only CF	60.0	15.0	0.93	57.5	50.0	1.80
Ablation: No Force	65.0	30.0	1.95	40.0	13.3	1.08

affordances or alternative ways of using tools—such as repurposing a knife to peel a carrot. Exploring Reinforcement Learning or World Models to enable the autonomous discovery of such creative tool manipulation strategies would be a valuable future direction. Beyond these policy-level limitations, SCFields also inherits two system-level limitations. First, it assumes that the functional contact region is at least partially observable in the tool point cloud. Second, our current setup uses multiple RGB-D cameras and tactile sensors to obtain reliable tool geometry and contact estimates under gripper self-occlusion. Reducing sensing complexity and temporal tracking for fully occluded contact states remain important future work.

REFERENCES

- [1] Ireaiyo Akinola, Jie Xu, Jan Carius, Dieter Fox, and Yashraj Narang. Tacsl: A library for visuotactile sensor simulation and learning. *IEEE Transactions on Robotics*, 2025.
- [2] 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.
- [3] Zihan Ding, Ya-Yen Tsai, Wang Wei Lee, and Bidan Huang. Sim-to-real transfer for robotic manipulation with tactile sensory. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6778–6785. IEEE, 2021.
- [4] A. Domahidi, E. Chu, and S. Boyd. ECOS: An SOCP solver for embedded systems. In *European Control Conference (ECC)*, pages 3071–3076, 2013.
- [5] Theophile Gervet, Zhou Xian, Nikolaos Gkanatsios, and Katerina Fragkiadaki. Act3d: 3d feature field transformers for multi-task robotic manipulation. In *Conference on Robot Learning*, pages 3949–3965. PMLR, 2023.
- [6] Zihao He, Hongjie Fang, Jingjing Chen, Hao-Shu Fang, and Cewu Lu. FoAR: Force-aware reactive policy for contact-rich robotic manipulation. *IEEE Robotics and Automation Letters*, 10(6):5625–5632, 2025. doi: 10.1109/LRA.2025.3560871.
- [7] Carolina Higuera, Siyuan Dong, Byron Boots, and Mustafa Mukadam. Neural contact fields: Tracking extrinsic contact with tactile sensing. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 12576–12582. IEEE, 2023.
- [8] Carolina Higuera, Joseph Ortiz, Haozhi Qi, Luis Pineda, Byron Boots, and Mustafa Mukadam. Perceiving extrinsic contacts from touch improves learning insertion policies. *arXiv preprint arXiv:2309.16652*, 2023.
- [9] Carolina Higuera, Akash Sharma, Chaithanya Krishna Bodduluri, Taosha Fan, Patrick Lancaster, Mrinal Kalakrishnan, Michael Kaess, Byron Boots, Mike Lambeta, Tingfan Wu, and Mustafa Mukadam. Sparsh: Self-supervised touch representations for vision-based tactile sensing. In *8th Annual Conference on Robot Learning*, 2024. URL <https://openreview.net/forum?id=xYJn2e1uu8>.
- [10] Binghao Huang, Yixuan Wang, Xinyi Yang, Yiyue Luo, and Yunzhu Li. 3d-vitac: Learning fine-grained manipulation with visuo-tactile sensing. In *8th Annual Conference on Robot Learning*, 2024. URL <https://openreview.net/forum?id=bk28WlkqZn>.
- [11] Tsung-Wei Ke, Nikolaos Gkanatsios, and Katerina Fragkiadaki. 3d diffuser actor: Policy diffusion with 3d scene representations. In *8th Annual Conference on Robot Learning*, 2024.
- [12] Leon Kim, Yunshuang Li, Michael Posa, and Dinesh Jayaraman. Im2contact: Vision-based contact localization without touch or force sensing. In *Conference on Robot Learning*, pages 1533–1546. PMLR, 2023.
- [13] Jayjun Lee and Nima Fazeli. Vitascope: Visuo-tactile implicit representation for in-hand pose and extrinsic contact estimation. In *Robotics: Science and Systems (RSS)*, 2025.
- [14] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.
- [15] Yijiong Lin, Alex Church, Max Yang, Haoran Li, John Lloyd, Dandan Zhang, and Nathan F Lepora. Bi-touch: Bimanual tactile manipulation with sim-to-real deep reinforcement learning. *IEEE Robotics and Automation Letters*, 8(9):5472–5479, 2023.
- [16] Fangchen Liu, Kuan Fang, Pieter Abbeel, and Sergey Levine. Moka: Open-world robotic manipulation through mark-based visual prompting. *arXiv preprint arXiv:2403.03174*, 2024.
- [17] Jason Jingzhou Liu, Yulong Li, Kenneth Shaw, Tony Tao, Ruslan Salakhutdinov, and Deepak Pathak. Factr: Force-attending curriculum training for contact-rich policy learning. *arXiv preprint arXiv:2502.17432*, 2025.

- [18] Wenhai Liu, Junbo Wang, Yiming Wang, Weiming Wang, and Cewu Lu. Forcemimic: Force-centric imitation learning with force-motion capture system for contact-rich manipulation. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1105–1112. IEEE, 2025.
- [19] Daolin Ma, Siyuan Dong, and Alberto Rodriguez. Extrinsic contact sensing with relative-motion tracking from distributed tactile measurements. In *2021 IEEE international conference on robotics and automation (ICRA)*, pages 11262–11268. IEEE, 2021.
- [20] Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, et al. Isaac gym: High performance gpu-based physics simulation for robot learning. *arXiv preprint arXiv:2108.10470*, 2021.
- [21] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017.
- [22] Daniel Seita, Yufei Wang, Sarthak J Shetty, Edward Yao Li, Zackory Erickson, and David Held. Toolflownet: Robotic manipulation with tools via predicting tool flow from point clouds. In *Conference on Robot Learning*, pages 1038–1049. PMLR, 2023.
- [23] Zilin Si and Wenzhen Yuan. Taxim: An example-based simulation model for gelsight tactile sensors. *IEEE Robotics and Automation Letters*, 7(2):2361–2368, 2022.
- [24] Anukriti Singh, Kasra Torshizi, Khuzema Habib, Kelin Yu, Ruohan Gao, and Pratap Tokekar. Afford2act: Affordance-guided automatic keypoint selection for generalizable and lightweight robotic manipulation. *arXiv preprint arXiv:2510.01433*, 2025.
- [25] Chao Tang, Anxing Xiao, Yuhong Deng, Tianrun Hu, Wenlong Dong, Hanbo Zhang, David Hsu, and Hong Zhang. Functo: Function-centric one-shot imitation learning for tool manipulation. *arXiv preprint arXiv:2502.11744*, 2025.
- [26] Chao Tang, Anxing Xiao, Yuhong Deng, Tianrun Hu, Wenlong Dong, Hanbo Zhang, David Hsu, and Hong Zhang. Mimicfunc: Imitating tool manipulation from a single human video via functional correspondence. In *Conference on Robot Learning*, pages 4473–4492. PMLR, 2025.
- [27] Shaoxiong Wang, Mike Lambeta, Po-Wei Chou, and Roberto Calandra. Tacto: A fast, flexible, and open-source simulator for high-resolution vision-based tactile sensors. *IEEE Robotics and Automation Letters*, 7(2): 3930–3937, 2022.
- [28] Yixuan Wang, Mingtong Zhang, Zhuoran Li, Katherine Rose Driggs-Campbell, Jiajun Wu, Li Fei-Fei, and Yunzhu Li. D³ fields: Dynamic 3d descriptor fields for zero-shot generalizable robotic manipulation. In *ICRA 2024 Workshop on 3D Visual Representations for Robot Manipulation*, 2023.
- [29] Yixuan Wang, Guang Yin, Binghao Huang, Tarik Kestemur, Jiuguang Wang, and Yunzhu Li. Gendp: 3d semantic fields for category-level generalizable diffusion policy. In *8th Annual Conference on Robot Learning*, volume 2, 2024.
- [30] Yunlong Wang, Lei Zhang, Yuyang Tu, Hui Zhang, Kaixin Bai, Zhaopeng Chen, and Jianwei Zhang. Tooleenet: Tool affordance 6d pose estimation. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10519–10526. IEEE, 2024.
- [31] Quantao Yang, Michael C Welle, Danica Kragic, and Olov Andersson. S²-diffusion: Generalizing from instance-level to category-level skills in robot manipulation. *IEEE Robotics and Automation Letters*, 2025.
- [32] Wenzhen Yuan, Siyuan Dong, and Edward H Adelson. Gelsight: High-resolution robot tactile sensors for estimating geometry and force. *Sensors*, 17(12):2762, 2017.
- [33] 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.
- [34] Jialiang Zhao, Yuxiang Ma, Lirui Wang, and Edward Adelson. Transferable tactile transformers for representation learning across diverse sensors and tasks. In *8th Annual Conference on Robot Learning*, 2024. URL <https://openreview.net/forum?id=KXsroptmNI>.
- [35] Haoyu Zhen, Xiaowen Qiu, Peihao Chen, Jincheng Yang, Xin Yan, Yilun Du, Yining Hong, and Chuang Gan. 3d-vla: A 3d vision-language-action generative world model. In *International Conference on Machine Learning*, pages 61229–61245. PMLR, 2024.