UniVTAC: A Unified Simulation Platform for Visuo-Tactile Manipulation Data Generation, Learning, and Benchmarking

Author Names Omitted for Anonymous Review. Paper-ID [175]

***Abstract*—Robotic manipulation has seen rapid progress with vision–language–action (VLA) policies. However, visuo-tactile perception is critical for contact-rich manipulation, as tasks such as insertion are difficult to complete robustly using vision alone. At the same time, acquiring large-scale and reliable tactile data in the physical world remains costly and challenging, and the lack of a unified evaluation platform further limits policy learning and systematic analysis. To address these challenges, we propose UniVTAC, a simulation-based visuo-tactile data synthesis platform that supports three commonly used visuo-tactile sensors and enables scalable and controllable generation of informative contact interactions. Based on this platform, we introduce the UniVTAC Encoder, a visuo-tactile encoder trained on large-scale simulation-synthesized data with designed supervisory signals, providing tactile-centric visuo-tactile representations for downstream manipulation tasks. In addition, we present the UniVTAC Benchmark, which consists of eight representative visuo-tactile manipulation tasks for evaluating tactile-driven policies. Experimental results show that integrating the UniVTAC Encoder improves average success rates by 17.1% on the UniVTAC Benchmark, while real-world robotic experiments further demonstrate a 25% improvement in task success.**

I. Introduction

Recent advances in robotic manipulation have enabled impressive performance across a broad range of tasks. Modern policies [13, 40, 6, 42, 25, 39, 21, 15, 2, 26] demonstrate strong capabilities in flexible object manipulation and fine-grained skills, supporting complex activities such as coffee making and tool use. Nevertheless, precise and reliable manipulation remains fundamentally challenging, particularly for tasks that require accurate reasoning about object pose and contact states, including insertion and alignment. In these settings, vision-based perception alone is often insufficient due to occlusions introduced by the end-effector, limited depth accuracy at close range, and the lack of direct observability of contact interactions once physical engagement occurs.

Tactile sensing provides complementary information by directly capturing localized contact geometry, force distribution, and relative motion at the interaction interface, which enables the detection of misalignment and supports closed-loop corrective control during manipulation. Building upon these advantages, visuo-tactile perception integrates visual appearance with tactile deformation cues, yielding dense and spatially resolved observations of contact phenomena such as pressure distribution, surface texture, and material properties. This combination makes visuo-tactile perception particularly well suited for contact-rich manipulation scenarios where reliable interaction depends on accurate perception under occlusion and close-range contact.

However, the current infrastructure for visuo-tactile manipulation remains underdeveloped. On the one hand, the scarcity of large-scale tactile data severely limits the training of tactile-centric representation models, leading to suboptimal performance of tactile-based manipulation policies. This limitation largely stems from practical constraints of real-world tactile sensing hardware, including the lack of standardized sensor designs, high manufacturing and deployment costs, and the difficulty of large-scale production and data collection. As a result, both representation learning and policy optimization for tactile manipulation are significantly constrained.

On the other hand, the absence of unified and comprehensive benchmarks for visuo-tactile manipulation hinders systematic evaluation and iterative improvement of tactile-driven strategies. Addressing this gap requires a reproducible platform that supports both scalable data acquisition and standardized policy evaluation, enabling fair comparison and principled analysis across different methods.

To address these challenges, we propose UniVTAC, a simulation-based framework for synthesizing visuo-tactile manipulation data that supports three widely used visuo-tactile sensors. Leveraging high-fidelity simulation, UniVTAC enables large-scale and controllable generation of visuo-tactile interaction data, including pressure patterns, marker deformations, and tangential force cues. Based on the synthesized data, we design multiple supervision objectives, such as visuo-tactile image reconstruction and pose estimation, and train a unified visuo-tactile encoder under this multi-task supervision. Through large-scale pretraining, the encoder learns tactile-centric representations that capture fine-grained contact boundaries while remaining sensitive to object pose and interaction dynamics.

To systematically evaluate the effectiveness of the synthesized data and the learned representations, we further introduce a simulation-based visuo-tactile manipulation benchmark built on TacEx [33] and NVIDIA Isaac Sim [34]. The benchmark comprises eight representative tactile manipulation tasks and supports automated data synthesis and unified policy evaluation. Together with UniVTAC, it enables systematic and reproducible analysis of visuo-tactile manipulation strategies under diverse contact scenarios. Our contributions are summarized as follows:

(1) **Simulation-based visuo-tactile data synthesis.** We

propose UniVTAC, a scalable and controllable simulation framework for synthesizing visuo-tactile contact and manipulation data, supporting three commonly used visuo-tactile sensors. Within this framework, we design task-specific supervisory signals tailored for tactile manipulation and use them to generate large-scale annotated synthetic visuo-tactile data. Leveraging this data, we pretrain a visuo-tactile encoder designed to support downstream tactile manipulation tasks.

(2) **UniVTAC Benchmark.** We introduce UniVTAC Benchmark, a simulation-based benchmark comprising eight dexterous visuo-tactile manipulation tasks. Built upon the UniVTAC platform, the benchmark supports automated data generation and unified policy evaluation, enabling systematic and reproducible analysis of visuo-tactile manipulation strategies.

(3) **Comprehensive evaluation and real-world validation.** We conduct extensive experiments on UniVTAC Benchmark to analyze the performance of representative manipulation policies on tactile-dependent tasks. We demonstrate the effectiveness of the UniVTAC Encoder representations in simulation and further validate their applicability through real-world robotic experiments.

II. Related Work

A. Visuo-tactile Sensor Simulation

The development of high-fidelity simulators for visuo-tactile sensors is a cornerstone for data-driven tactile perception research. Recent progress has been driven by advancements in physics-based simulation and differentiable rendering. Existing frameworks can be categorized by their underlying physical formulation.

The Incremental Potential Contact (IPC) method [20] has emerged as a prominent framework for contact-rich robotic simulation, offering rigorous non-penetration guarantees and robust handling of complex frictional interactions. Building on IPC, [4] and [18] integrated it into robotic manipulation pipelines through SapienIPC, a soft-body capable extension of the original SAPIEN simulator, enabling high-fidelity sim-to-real transfer. Taccel [22] significantly accelerated this pipeline via a GPU-optimized IPC backend, achieving real-time or faster-than-real-time simulation. TacEx [33] couples the modern IPC library libuipc [11, 12] with a physically based visuo-tactile renderer in NVIDIA Isaac Sim, supporting real-time, high-fidelity simulation of GelSight-style sensors. The Finite Element Method (FEM) offers an alternative route focused on geometric and material fidelity, exemplified by TacFlex [49]. By coupling FEM-based elastic deformation models with ray-traced optical rendering. In contrast, DiffTactile [37] adopts the Material Point Method (MPM) to model highly deformable objects, leveraging its compatibility with automatic differentiation for gradient-based policy optimization. Separately, Tacchi [5] utilizes MPM for its efficiency in pluggable, modular simulation settings, where fast setup and moderate accuracy are prioritized.

Our work is built upon the TacEx framework [33], leveraging its physical accuracy and integration in Isaac Sim, introducing multiple sensors support and automatic manipulation APIs. Our marker deformation simulation is also inspired by the IPC-based contact modeling in [4] and [18].

B. Visuo-tactile Representation Learning

Learning effective representations from tactile signals typically follows three paradigms: reconstruction, explicit geometric supervision, and multi-modal alignment. Reconstruction-based methods focus on capturing visual distribution. Marker-Embedded GAN [16] utilizes GANs to generate RGB images from depth maps to bridge the cross-modal gap, while UniT [44] employs VQGAN to reconstruct tactile images for data-efficient encoding. However, as these methods optimize for pixel-level fidelity, they often overlook the underlying contact physics and force distributions.

To capture finer contact details, explicit geometric supervision has been explored. RDP [45] treats markers as a 2D grid-structured point matrix and applies PCA to capture dominant deformation modes. In contrast, the framework of [4] models markers as 3D point clouds and uses a PointNet-based architecture to learn nonlinear geometric features from full spatial coordinates. While these methods are robust in modeling physical contact, they often sacrifice rich tactile textures and object-specific shape information by discarding high-dimensional image features.

Recently, vision-tactile contrastive learning has gained traction by aligning tactile embeddings with multi-view or temporal visual signals [9, 24]. Although these approaches achieve impressive results in contact-rich manipulation, the resulting representations are often task-specific. Relying on global alignment (e.g., CLIP-like objectives) often necessitates training the encoder for each new task, thereby limiting the generalizability of the learned tactile representation.

C. Simulation for Data Generation

Many simulation-based methods have been proposed for robotic policy learning [43, 10, 46, 29, 32, 8, 7, 35, 23, 17, 3, 31], providing scalable environments for training and evaluation. Building upon these platforms, a number of simulation-based pipelines further focus on large-scale robot data generation. For example, the RoboTwin series [30] and RoboGen [41] employ expert code to synthesize high-quality, near-perfect manipulation trajectories. MimicGen [28] and DexMimicGen [14] extend data collection through human teleoperation, while other works [47, 19] obtain expert policies via reinforcement learning with carefully designed reward functions and subsequently use them for data synthesis. Collectively, these approaches enable efficient large-scale data collection and, in some cases, support natural-language task specifications for novel tasks [27].

Despite their success, most existing simulation-based data generation pipelines primarily focus on rigid or articulated object manipulation and provide limited support for modeling the complex dynamics of tactile interaction. Standard protocols typically rely on heuristic gripper commands and binary success criteria based on final object states, without explicitly modeling transient contact forces, deformation, or slippage during execution. As a result, such open-loop or weakly constrained execution may produce trajectories with excessive grasping forces or unstable contact configurations, which would be unsafe for real-world tactile sensors and insufficient for learning tactile-dependent manipulation skills.

In contrast, our work explicitly incorporates tactile awareness into the data generation process through a closed-loop control scheme for gripper aperture. By integrating runtime validity checks and contact-sensitive feedback, our pipeline avoids sensor-destructive behaviors and ensures that synthesized trajectories capture rich, physically meaningful contact responses. This design enables the generation of tactile-aware data that is better aligned with the requirements of visuo-tactile manipulation and downstream tactile-based policy learning.

III. UniVTAC

The representation of observations plays a central role in effective robotic manipulation, particularly in contact-rich settings. However, acquiring large-scale and reliable tactile data in the physical world remains difficult to scale, as real-world tactile sensors provide only raw observations and incur substantial data collection costs. Simulation environments, by contrast, offer access to ground-truth physical states and enable scalable synthesis of contact interactions.

In this section, we present UniVTAC as an end-to-end simulation-driven framework for visuo-tactile manipulation, encompassing data synthesis, representation learning, and systematic evaluation. UniVTAC enables automated generation of large-scale annotated visuo-tactile interaction data, which we use to train the UniVTAC Encoder via auxiliary supervision derived from simulation-specific physical signals. The resulting tactile-centric visuo-tactile representations are designed to support downstream policy learning in contact-rich manipulation tasks. To systematically assess the effectiveness of the learned representations and the policies built upon them, we further introduce the UniVTAC Benchmark, a simulation benchmark comprising eight representative visuo-tactile manipulation tasks that span diverse contact scenarios.

A. UniVTAC Platform

Built upon the TacEx framework [33], UniVTAC extends the soft-body simulation capabilities to support diverse sensor types and complex manipulation tasks.

1) Sensor Configuration and Heterogeneity: UniVTAC integrates three mainstream visuotactile sensors: *GelSight Mini*[48], *ViTai GF225* [36], and *Xense WS* [1]. The optical and mechanical properties of each sensor are modeled by adjusting internal camera intrinsics, gelpad meshes, and rendering methods. This modularity allows researchers to validate tactile algorithms across different hardware specifications within a unified environment.

2) Automated Manipulation APIs: To facilitate large-scale and high-fidelity data collection, we implement a library of atomic manipulation APIs: *Grasp*, *Move*, *Place*, *Probe*, and *Rotate*. While *Move* and *Place* focus on precise trajectory generation via *cuRobo* [38], the *Grasp* and *Probe* primitives incorporate a tactile-reactive adaptive control mechanism to ensure physical consistency. Specifically, during the *Grasp* sequence, the gripper's joint velocity $\dot{q}$ is governed by a state-dependent feedback law to prevent non-physical penetration:

$$\dot{q} = \begin{cases} v_{\text{fast}} & \text{if } d_{\min} = d_{\max} \\ \min\left(|d_{\min} - \delta_{\text{th}}|, v_{\text{slow}}\right) & \text{if } d_{\min} < d_{\max} \end{cases} \tag{1}$$

where $d_{\min}$ is the real-time minimum depth from the tactile sensors, $d_{\max}$ denotes the zero-contact depth, and δ_{th} is the reaching depth threshold. The v_{fast} and v_{slow} constrains the maximum velocity during the phase without or with contact. This closed-loop approach effectively mitigates non-physical clipping artifacts and ensures that the captured tactile imprints remain within a realistic deformation manifold, providing high-quality training data for the subsequent representation learning.

3) Tactile Representation Prerequisites: To learn a robust general-purpose tactile representation, we posit that the encoder must disentangle three distinct physical properties from raw tactile observations. We categorize these into three perception pathways, which guide our data generation and subsequent supervision strategies. **Shape Perception** targets global geometry, compelling the model to disentangle the object's intrinsic visual appearance from sensor-specific artifacts such as markers and lighting. **Contact Perception** focuses on local dynamics, explicitly modeling interaction mechanics through surface deformation and shear-induced marker displacement. **Pose Perception** provides spatial grounding, anchoring tactile signals in external metric space to enable consistent spatial reasoning for downstream policies. To support these pathways, our pipeline is designed to capture high-fidelity ground truth for each modality alongside the raw sensor input.

4) Contact-Rich Data Generation: We curate a geometrically diverse dataset to ensure the encoder encounters broad variations in local curvature and contact topology. Following protocols in [16], we employ 14 distinct geometric primitives as indenters, ranging from standard convex shapes (e.g., *spheres, cones*) to complex non-convex patterns (e.g., *stars, cross-shapes*). These objects are mounted on a standardized prism base to facilitate varying grasp interactions.

For each episode, the robotic gripper approaches a prism on the ground plane with shapes on its surface and initiates contact. To capture the dynamic relationship between gripper width and gel deformation, we vary the grasping tightness stochastically by setting the reaching depth threshold δ_{th}. This generates data ranging from light touches to deep indentations. Once in contact, the robot executes a series of randomized small-scale rotations through the *Move* and *Rotate* APIs. These actions induce rich shear force patterns and marker displacements, which are critical for the contact perception. During this interaction sequence, we synchronously record the raw tactile images with markers I_{marked}, along with the simulation-privileged ground truth: pure contact images I_{pure}, depth maps D, 2D projections of fiducial markers M, and the pose of the object in the gelpad center local frame p.

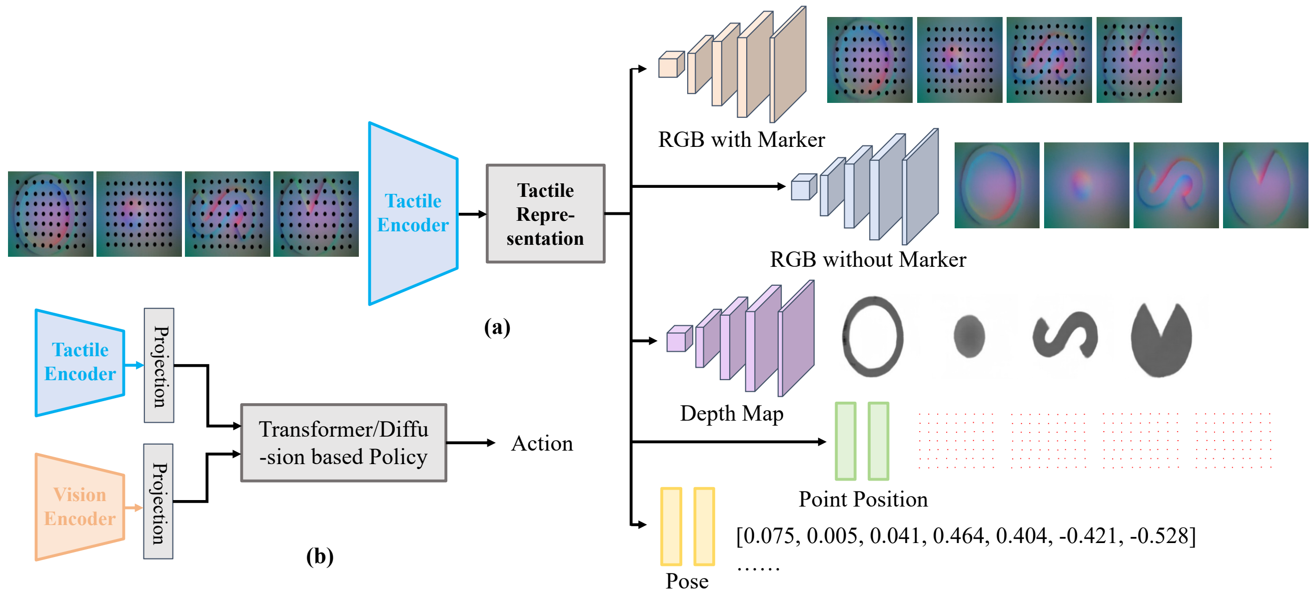


Fig. 1. **UniVTAC Encoder Framework and Integration into Policy Learning.** (a) The UniVTAC Encoder is pretrained with three self-supervised objectives, including shape reconstruction, contact deformation prediction, and object pose regression, to learn a structured tactile-centric representation from raw visuo-tactile observations. (b) At deployment time, the pretrained encoder is integrated as a perception module for downstream manipulation policies, enabling end-to-end policy learning from raw tactile images without introducing additional inference-time overhead.

This automated pipeline generates approximately 14,000 interaction frames per shape, yielding a total dataset of 205,826 samples.

B. UniVTAC Encoder

We formulate UniVTAC Encoder as a multi-pathway representation learning framework that embeds structured physical priors into tactile observations through task-driven supervision. As illustrated in Figure 1, a shared encoder maps each visuo-tactile observation into a compact latent representation, which is then decoded by multiple pathway-specific heads during training. These decoding pathways impose complementary inductive biases on the learned representation, while all decoders are discarded at deployment time, ensuring that only the encoder is retained with no additional inference-time overhead.

The design of UniVTAC is grounded in three perceptual prerequisites that are critical for contact-rich manipulation: object shape understanding, local contact dynamics, and spatial pose awareness. To support shape perception, UniVTAC Encoder employs a dual-view reconstruction strategy in which the encoder is supervised to reconstruct both the raw tactile image with marker patterns, denoted as $I_{\text{marked}} \in \mathbb{R}^{3\times H\times W}$, and a counterfactual marker-free image $I_{\text{pure}} \in \mathbb{R}^{3\times H\times W}$. Jointly optimizing these two reconstruction objectives encourages the encoder to disentangle intrinsic object geometry from sensor-specific marker artifacts, resulting in shape-aware representations that are robust to appearance variations.

Contact perception is enforced by supervising the encoder with physically grounded signals that capture local deformation dynamics. Specifically, UniVTAC Encoder predicts a dense surface depth map $D \in \mathbb{R}^{H\times W}$ representing normal indentation of the gelpad, together with the 2D projections of N fiducial markers, denoted as $M \in \mathbb{R}^{2\times N}$, which encode lateral shear and tangential deformation. By jointly modeling geometric deformation and marker displacement, the learned representation is encouraged to capture true contact mechanics rather than relying on pixel-level cues alone.

To anchor tactile observations within a global spatial context, UniVTAC Encoder further incorporates a pose perception pathway that regresses the relative object pose $p \in \mathbb{R}^7$, consisting of 3D translation and a quaternion representation of orientation. This explicit spatial supervision grounds the latent representation in a metric space, enabling downstream policies to perform spatially consistent reasoning during manipulation.

Architecturally, UniVTAC Encoder adopts a ResNet-18 backbone as the shared encoder, which projects each tactile observation into a latent feature vector. The shape reconstruction heads employ transposed convolutional decoders to generate RGB outputs, while the depth decoder follows a similar architecture with a single-channel output. Marker flow is predicted using a lightweight multilayer perceptron, and pose regression is performed by a compact MLP that outputs the 7-dimensional pose vector. Reconstruction examples are shown in Figure 2.

The shared encoder and task-specific heads are trained end-to-end using a multi-task objective. We employ Mean Squared Error (MSE) as the primary optimization criterion for all pathways. The total loss $\mathcal{L}_{\text{total}}$ is defined as a weighted sum of the following components:

Shape Reconstruction Loss. This loss encourages the model to capture global geometry by reconstructing both the marked and marker-free observations:

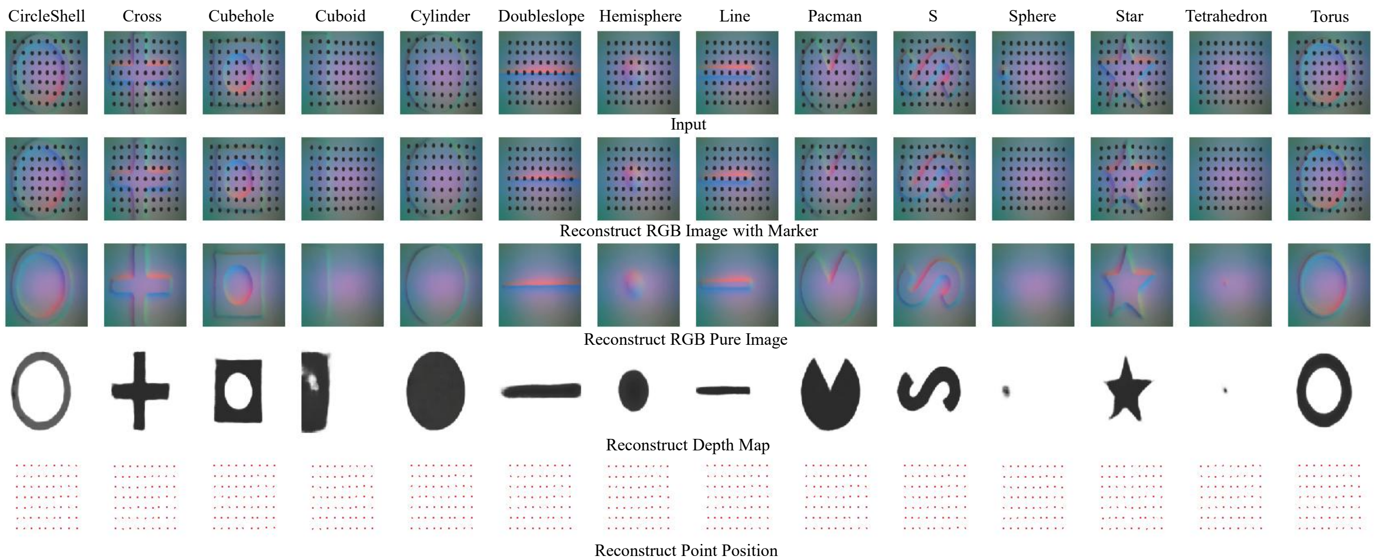


Fig. 2. **Reconstruction Result**. From a tactile image with markers, UniVTAC Encoder reconstructs complementary physical signals, including the marker-free tactile image, gelpad deformation depth map, and marker point positions, across diverse contact geometries. These results show that the learned representation captures both global shape cues and fine-grained contact deformation beyond sensor-specific visual patterns.

$$\mathcal{L}_{\text{shape}} = MSE(\hat{I}_{\text{marked}}, I_{\text{marked}}) + MSE(\hat{I}_{\text{pure}}, I_{\text{pure}}) \quad (2)$$

where $\hat{I}$ and I denote the predicted and ground-truth tactile images, respectively.

Contact Deformation Loss. To supervise local interaction dynamics, we minimize the discrepancy in surface depth and marker displacements:

$$\mathcal{L}_{\text{contact}} = MSE(\hat{D}, D) + MSE(\hat{M}, M) \quad (3)$$

where D is the depth map and M represents the marker positions porjected onto the image plane.

Pose Regression Loss. The spatial grounding is optimized via:

$$\mathcal{L}_{\text{pose}} = MSE(\hat{p}, p) \quad (4)$$

where p represents the object relevant pose (3D translation and 4D quaternion).

The final training objective is

$$\mathcal{L}_{\text{total}} = \lambda_s \mathcal{L}_{\text{shape}} + \lambda_c \mathcal{L}_{\text{contact}} + \lambda_p \mathcal{L}_{\text{pose}} \quad (5)$$

In our experiments, we empirically set the balancing hyper-parameters to $\lambda_s = 1.0$, $\lambda_c = 0.5$, and $\lambda_p = 0.5$.

C. UniVTAC Benchmark

Following the RoboTwin paradigm and the tactile perception design of UniVTAC, we leverage simulation APIs, annotated object assets, and expert programs to enable automated task-level data synthesis and evaluation. This design integrates data collection, model training, and policy evaluation into a unified pipeline, supporting scalable and reproducible experimentation.

Expert-driven data synthesis methods are capable of generating near-perfect manipulation trajectories. For example, insertion tasks can be synthesized with highly precise and collision-free executions. However, such trajectories are sub-optimal for tactile learning, as they lack meaningful variations in contact perception. In these executions, contact events are often trivial or instantaneous, preventing tactile policies from learning informative, contact-dependent behaviors.

To address this limitation, we introduce stochasticity and corrective behaviors into the synthesis process. Specifically, for contact-rich skills such as insertion, we intentionally introduce randomized failures during execution and allow the expert controller to perform corrective actions based on perceived contact feedback, which in simulation is provided by ground-truth physical states. This process more closely resembles human manipulation behavior, where coarse alignment is followed by iterative contact-based correction until task completion. As a result, the synthesized trajectories exhibit diverse contact patterns and corrective interactions, providing rich tactile supervision that enables policies to effectively leverage visuo-tactile information.

To ensure reliable evaluation, the UniVTAC Benchmark incorporates tactile-specific, physics-based success criteria that extend beyond reaching target coordinates. A trial is considered invalid if the maximum penetration depth exceeds a predefined safety threshold or if significant relative slip is detected between the gelpad and the object surface. These constraints prevent degenerate solutions that exploit simulation artifacts and ensure that evaluation outcomes reflect physically meaningful manipulation behaviors rather than numerical or physical loopholes.

We design eight visuo-tactile manipulation tasks spanning three perceptual dimensions. All tasks are visualized in Figure 3, with detailed task descriptions provided in Appendix A. Pose reasoning tasks, including *Lift Bottle*, *Lift Can*, and

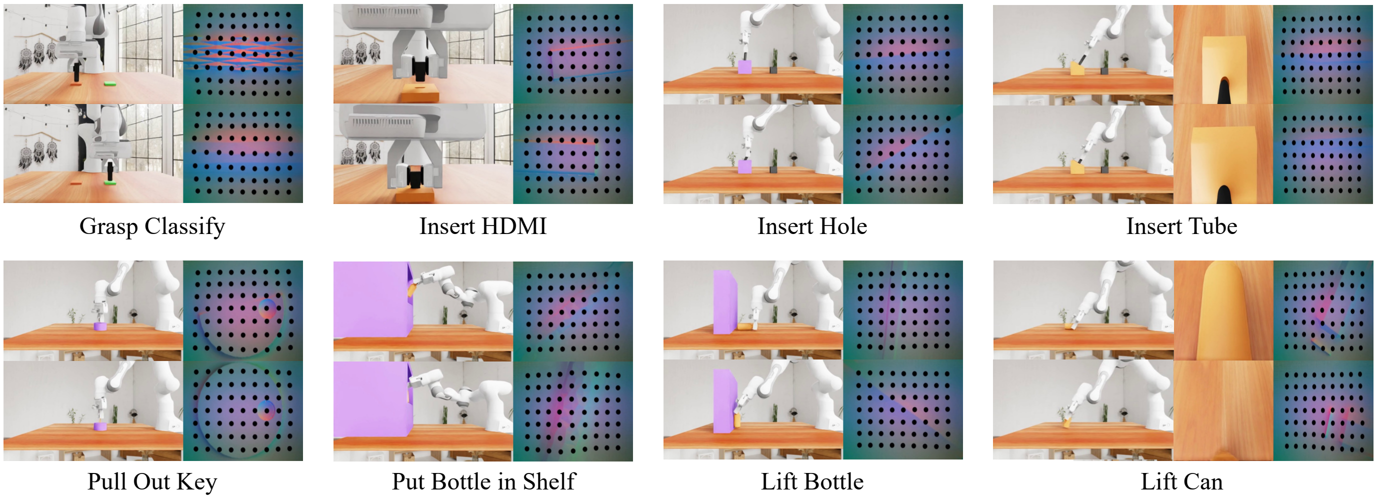


Fig. 3. **UniVTAC Benchmark Tasks**. The UniVTAC Benchmark comprises eight representative visuo-tactile manipulation tasks spanning shape recognition, pose reasoning, and contact-rich interaction, and is designed to systematically evaluate tactile-dependent manipulation policies. For each task, we visualize two representative key frames corresponding to critical stages of execution. Each key frame includes both a visuo-tactile observation and a standard visual observation. For clarity of presentation, we display the tactile observation from only one side of the gripper, although tactile sensing is available on both fingertips during execution.

Put Bottle in Shelf, require policies to estimate the relative spatial relationship between the gripper and the object during interaction, relying primarily on accurate perception of object position and orientation. Shape perception tasks, such as *Grasp Classify*, focus on distinguishing object geometries from high-dimensional tactile observations, emphasizing the ability of learned representations to capture shape-related contact cues. Contact-rich interaction tasks, including *Insert Hole*, *Insert Tube*, *Insert HDMI*, and *Pull Out Key*, involve fine-grained alignment and sequential contact transitions, requiring precise reasoning over contact dynamics throughout the manipulation process.

IV. Experiment

Our experiments are designed to investigate three key questions. (1) We evaluate the performance of representative manipulation policies on the UniVTAC Benchmark to characterize their behavior across diverse visuo-tactile tasks; (2) We examine the extent to which the UniVTAC Encoder supports different classes of policies, assessing its effectiveness as a representation module across varying policy architectures; (3) We conduct simulation-to-real experiments to assess whether the UniVTAC Encoder, trained solely on simulation-synthesized data, transfers effectively to real-world robotic manipulation.

A. Evaluation on the UniVTAC Benchmark

We evaluate representative manipulation policies on the UniVTAC Benchmark, focusing on Action Chunking Transformers (ACT) [50]. ACT models manipulation as sequence prediction using a Transformer-based conditional variational framework, enabling temporally coherent action generation and flexible fusion of sensory inputs through attention mechanisms. To assess the effectiveness of tactile-centric representations, we further evaluate ACT augmented with the UniVTAC Encoder, which incorporates visuo-tactile features learned through large-scale simulation pretraining. In addition, we include VITaL [9], a representative visuo-tactile manipulation policy that leverages visuo-tactile pretraining, serving as a strong baseline that explicitly exploits tactile representations.

All policies are trained on 50 automatically collected full trajectories per task and evaluated over 100 test rollouts.

Table I summarizes the performance of representative manipulation policies on the UniVTAC Benchmark. Overall, the results highlight both the effectiveness of the benchmark design and the importance of tactile feedback for contact-rich manipulation. The benchmark spans tasks with varying degrees of perceptual and interaction complexity, enabling a nuanced analysis of how different policies exploit visual and tactile information.

Across all tasks, ACT augmented with the UniVTAC Encoder achieves a substantial performance improvement over its vision-only counterpart, increasing the average success rate from 30.9% to 48.0%. The improvement is consistently observed across multiple contact-sensitive tasks, including insertion and pull-out scenarios. This trend indicates that the UniVTAC Encoder provides tactile-centric representations that can be effectively exploited by attention-based policy architectures. Moreover, these results suggest that the UniVTAC Benchmark captures manipulation scenarios in which tactile perception plays a decisive role, and that ACT's Transformer-based attention mechanism is well suited for integrating heterogeneous visuo-tactile features.

We also observe that VITaL achieves strong performance on several tasks, most notably attaining near-perfect accuracy on *Grasp Classify*. This behavior is expected, as VITaL benefits from explicit visuo-tactile pretraining, enabling the extraction of discriminative tactile features that are closely

TABLE I
UNIVTAC BENCHMARK. WE REPORT THE SUCCESS RATES AND AVERAGE PERFORMANCE OF ACT, VITAL, AND ACT W/ UNIVTAC ACROSS THE EIGHT TASKS IN THE UNIVTAC BENCHMARK.

Method	Lift Bottle	Pull-out Key	Lift Can	Put Bottle in Shelf	Insert Hole	Insert HDMI	Insert Tube	Grasp Classify	***Average***
ACT	42.0%	28.0%	20.0%	28.0%	19.0%	15.0%	45.0%	50.0%	30.9%
VITaL	72.0%	47.0%	8.0%	32.0%	25.0%	6.0%	34.0%	100.0%	40.5%
ACT w/ UniVTAC	71.0%	46.0%	29.0%	31.0%	24.0%	28.0%	56.0%	99.0%	48.0%

aligned with shape-related contact cues. As such, VITaL serves as a meaningful reference point, illustrating the effectiveness of representation pretraining for visuo-tactile manipulation and providing a practical upper bound for evaluating the UniVTAC Encoder under comparable experimental settings.

These observations indicate that the UniVTAC Benchmark spans a continuum of task dependencies, ranging from visually solvable manipulation to scenarios that fundamentally rely on tactile feedback. While vision-only policies can achieve competitive performance on tasks dominated by geometric visibility, tactile perception becomes essential for robust execution in contact-rich and correction-intensive settings. Consequently, the benchmark enables a more principled evaluation of manipulation policies by revealing not only aggregate task success, but also how effectively different policy architectures exploit visuo-tactile representations.

B. Efficacy of UniVTAC Encoder

We further evaluate the UniVTAC Encoder by comparing it against baselines and ablation variants.

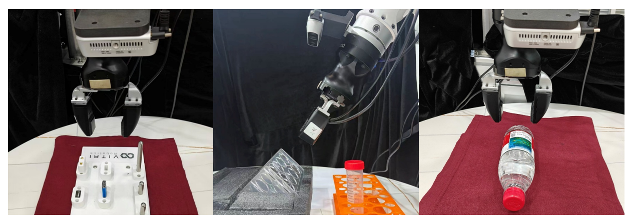
We first compare our method with a CLIP-based visuo-tactile encoder. When integrated with ACT, the UniVTAC Encoder achieves an average success rate of 48.0%, outperforming the contrastive learning-based VITaL, which attains 40.5%. This performance gap suggests that the reconstruction-based pretraining strategy adopted in UniVTAC captures more actionable and physically grounded contact information than contrastive representations alone, which primarily emphasize global feature alignment.

TABLE II
ABLATION STUDY ON ACT. WE REPORT THE PERFORMANCE OF ACT-BASED VARIANTS, INCLUDING ACT W/ UNIVTAC, MARKED_RGB, AND FROM_SCRATCH, UNDER THE SAME EVALUATION PROTOCOL.

Task	ACT w/ UniVTAC	marked_rgb	from_scratch
Lift Bottle	71.0%	56.0%	56.0%
Pull-out Key	46.0%	41.0%	40.0%
Lift Can	29.0%	11.0%	33.0%
Put Bottle in Shelf	31.0%	25.0%	34.0%
Insert Hole	24.0%	12.0%	3.0%
Insert HDMI	28.0%	24.0%	16.0%
Insert Tube	56.0%	65.0%	45.0%
Average	**40.7%**	**33.4%**	**32.4%**

We further conduct ablation studies to validate the effectiveness of individual design components in UniVTAC. Specifically, we compare against two variants implemented on ACT, as reported in Table II. The *marked_rgb* variant pretrains the encoder using only marker-based visual observations, while the *from_scratch* variant trains the encoder end-to-end from random initialization. Our full model achieves an average success rate of 40.7%, consistently outperforming both *marked_rgb* (33.4%) and *from_scratch* (32.4%). The improvement over *marked_rgb* highlights the necessity of dedicated tactile encoding beyond appearance cues, while the gain over *from_scratch* demonstrates that the proposed pretraining strategy yields more robust and transferable representations than learning from scratch, even when full encoder fine-tuning is permitted.

C. Real-World Experiments



Insert Tube **Insert USB** **Bottle Upright**

Fig. 4. **Real World Taks**. We illustrate the experimental settings for the three real-world manipulation tasks evaluated in this work.

We evaluate our approach on three physically challenging dexterous manipulation tasks using a real-world robotic platform: *Insert Tube*, *Insert USB*, and *Bottle Upright*. All experiments are conducted on the Tianji Robotics Marvin manipulator, a 7-DoF robotic arm equipped with a parallel gripper and a wrist-mounted RGB camera for visual feedback. Tactile sensing is provided by two ViTai GF225 visuo-tactile sensors embedded in the gripper fingertips, each capturing high-resolution deformation images at 30 Hz. The sensors output marker-based RGB observations that encode rich contact geometry and pressure distribution during interaction. The experimental setups are illustrated in Figure 4, and representative key frames of the tasks are shown in Figure 5.

To collect expert demonstrations, we employ VR-based teleoperation using Meta Quest headsets. Human operators control the end-effector pose via hand-held controllers while observing live video streams from the wrist-mounted camera. For each task, we collect 150 high-quality demonstration trajectories, which are used to train an Action Chunking Transformer (ACT) integrated with a diffusion-based policy head. The policy maps sequences of multimodal observations

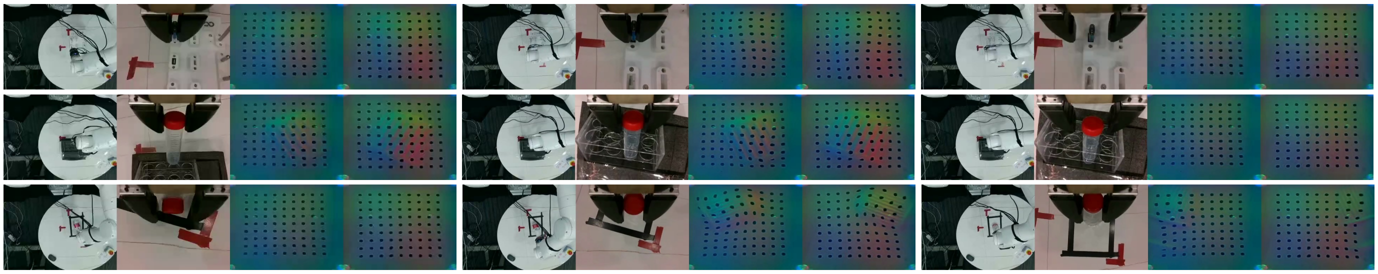


Fig. 5. **Real-world Task Key Frames**. Representative key frames from three real-world visuo-tactile manipulation tasks, showing synchronized wrist RGB images (left) and marker-based tactile observations (right) at the initial approach, contact-rich interaction, and final completion stages. The intermediate frames highlight evolving contact states and deformation cues that support fine-grained alignment and correction beyond vision-only perception.

to future actions and outputs joint positions as direct action commands.

The UniVTAC Encoder is first pretrained in simulation using large-scale synthesized visuo-tactile contact data to learn generalizable contact representations. To rigorously evaluate the generalization of the pretrained feature, the encoder is transferred to the real robot without further fine-tuning at deployment time. We conduct ablation studies comparing policies with tactile input enabled via the UniVTAC Encoder against vision-only baselines. All policies are trained separately for each task and evaluated over 20 rollouts under identical initial conditions, with task success determined by human observers.

TABLE III
REAL-WORLD EVALUATION RESULTS. WE REPORT THE REAL-WORLD PERFORMANCE ON THREE MANIPULATION TASKS, COMPARING A VISION-ONLY OBSERVATION SETTING WITH A VISION-AND-TACTILE FUSION SETTING.

Task	Vision	Vision + UniVTAC
Insert Tube	55.0%	**85.0%**
Insert USB	15.0%	**25.0%**
Bottle Upright	60.0%	**95.0%**
Average	43.3%	**68.3%**

We further analyze the real-world experimental results to understand how the UniVTAC Encoder contributes to performance gains across different manipulation tasks. As summarized in Table III, integrating the UniVTAC Encoder leads to consistent improvements across all evaluated tasks, with an average success rate increase of 25%. In fine-grained insertion tasks, including *Insert Tube* and *Insert USB*, policies augmented with the UniVTAC Encoder achieve performance gains of 30% and 10%, respectively. These improvements indicate that UniVTAC provides high-quality tactile feedback that enables policies to reason over subtle contact cues during precise alignment and insertion. Successful execution in these tasks often depends on detecting partial misalignment and responding with incremental corrective motions, which cannot be reliably inferred from vision alone.

In the *Bottle Upright* task, incorporating the UniVTAC Encoder results in a performance improvement of 35%. This outcome highlights that the benefits of UniVTAC extend beyond direct contact detection to object pose understanding. Although the encoder is pretrained purely in simulation, its representations implicitly capture object orientation and pose-related cues that are critical for maintaining stability during manipulation. The strong performance observed in real-world execution demonstrates that such pose-aware representations transfer effectively to physical robotic systems.

Overall, the average improvement of 25% in real-world success rates demonstrates that the UniVTAC Encoder not only provides informative tactile representations, but also enables effective deployment in physical robotic systems despite being pretrained exclusively on simulation-synthesized data. These results underscore the value of simulation-based visuo-tactile data synthesis and highlight UniVTAC as a practical and scalable approach to enhance real-world dexterous manipulation through tactile perception.

V. CONCLUSION

We presented UniVTAC, a unified simulation platform for scalable visuo-tactile data generation, representation learning, and benchmarking in contact-rich robotic manipulation. Built on high-fidelity tactile simulation, UniVTAC enables controllable synthesis of diverse visuo-tactile interactions and provides structured supervisory signals for learning tactile-centric representations. Leveraging this platform, we introduced UniVTAC Encoder, a visuo-tactile encoder pretrained via multi-pathway supervision that captures object shape, contact deformation, and pose information from tactile observations.

To facilitate systematic evaluation, we further proposed UniVTAC Benchmark, a benchmark comprising eight representative visuo-tactile manipulation tasks that emphasize tactile-dependent reasoning under occlusion and contact uncertainty. Extensive simulation experiments demonstrate that UniVTAC Encoder consistently improves manipulation performance when integrated with modern policy architectures, while real-world evaluations confirm effective sim-to-real transfer despite being pretrained purely in simulation.

We believe UniVTAC establishes a practical and extensible foundation for visuo-tactile manipulation research, enabling scalable data generation, robust tactile representation learning and fair benchmarking. Future work will extend this framework to more diverse sensor modalities, dynamic interactions, and open-world manipulation scenarios.

REFERENCES

[1] Xense. https://www.xenserobotics.com/product/367/detail/9, 2025. 2.

[2] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

[3] 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.

[4] Weihang Chen, Jing Xu, Fanbo Xiang, Xiaodi Yuan, Hao Su, and Rui Chen. General-Purpose Sim2Real Protocol for Learning Contact-Rich Manipulation With Marker-Based Visuotactile Sensors. *IEEE Transactions on Robotics*, 40:1509–1526, 2024. ISSN 1552-3098, 1941-0468. doi: 10.1109/TRO.2024.3352969. URL https://ieeexplore.ieee.org/document/10388459/.

[5] Zixi Chen, Shixin Zhang, Shan Luo, Fuchun Sun, and Bin Fang. Tacchi: A Pluggable and Low Computational Cost Elastomer Deformation Simulator for Optical Tactile Sensors. *IEEE Robotics and Automation Letters*, 8 (3):1239–1246, 2023-03. ISSN 2377-3766, 2377-3774. doi: 10.1109/LRA.2023.3237042. URL http://arxiv.org/abs/2301.08343.

[6] Shengliang Deng, Mi Yan, Songlin Wei, Haixin Ma, Yuxin Yang, Jiayi Chen, Zhiqi Zhang, Taoyu Yang, Xuheng Zhang, Heming Cui, et al. Graspvla: a grasping foundation model pre-trained on billion-scale synthetic action data. *arXiv preprint arXiv:2505.03233*, 2025.

[7] Ricardo Garcia, Shizhe Chen, and Cordelia Schmid. Towards generalizable vision-language robotic manipulation: A benchmark and llm-guided 3d policy. In *International Conference on Robotics and Automation (ICRA)*, 2025.

[8] Haoran Geng, Feishi Wang, Songlin Wei, Yuyang Li, Bangjun Wang, Boshi An, Charlie Tianyue Cheng, Haozhe Lou, Peihao Li, Yen-Jen Wang, et al. Roboverse: Towards a unified platform, dataset and benchmark for scalable and generalizable robot learning. *arXiv preprint arXiv:2504.18904*, 2025.

[9] Abraham George, Selam Gano, Pranav Katragadda, and Amir Barati Farimani. VITaL Pretraining: Visuo-Tactile Pretraining for Tactile and Non-Tactile Manipulation Policies, 2024-09-26. URL http://arxiv.org/abs/2403.11898.

[10] Jiayuan Gu, Fanbo Xiang, Xuanlin Li, Zhan Ling, Xiqiang Liu, Tongzhou Mu, Yihe Tang, Stone Tao, Xinyue Wei, Yunchao Yao, et al. Maniskill2: A unified benchmark for generalizable manipulation skills. In *The Eleventh International Conference on Learning Representations*, 2023.

[11] Kemeng Huang, Floyd M. Chitalu, Huancheng Lin, and Taku Komura. Gipc: Fast and stable gauss-newton optimization of ipc barrier energy. *ACM Trans. Graph.*, 43 (2), mar 2024. ISSN 0730-0301. doi: 10.1145/3643028.

[12] Kemeng Huang, Xinyu Lu, Huancheng Lin, Taku Komura, and Minchen Li. Stiffgipc: Advancing gpu ipc for stiff affine-deformable simulation. *ACM Trans. Graph.*, 44(3), May 2025. ISSN 0730-0301. doi: 10.1145/3735126.

[13] Physical Intelligence, Ali Amin, Raichelle Aniceto, Ashwin Balakrishna, Kevin Black, Ken Conley, Grace Connors, James Darpinian, Karan Dhabalia, Jared DiCarlo, Danny Driess, Michael Equi, Adnan Esmail, Yunhao Fang, Chelsea Finn, Catherine Glossop, Thomas Godden, Ivan Goryachev, Lachy Groom, Hunter Hancock, Karol Hausman, Gashon Hussein, Brian Ichter, Szymon Jakubczak, Rowan Jen, Tim Jones, Ben Katz, Liyiming Ke, Chandra Kuchi, Marinda Lamb, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Yao Lu, Vishnu Mano, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Charvi Sharma, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, Will Stoeckle, Alex Swerdlow, James Tanner, Marcel Torne, Quan Vuong, Anna Walling, Haohuan Wang, Blake Williams, Sukwon Yoo, Lili Yu, Ury Zhilinsky, and Zhiyuan Zhou. $\pi^*_{0.6}$: a vla that learns from experience, 2025. URL https://arxiv.org/abs/2511.14759.

[14] Zhenyu Jiang, Yuqi Xie, Kevin Lin, Zhenjia Xu, Weikang Wan, Ajay Mandlekar, Linxi Fan, and Yuke Zhu. Dexmimicgen: Automated data generation for bimanual dexterous manipulation via imitation learning. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 2025.

[15] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R Sanketi, Quan Vuong, et al. Openvla: An open-source vision-language-action model. In *8th Annual Conference on Robot Learning*.

[16] Won Dong Kim, Sanghoon Yang, Woojong Kim, Jeong-Jung Kim, Chang-Hyun Kim, and Jung Kim. Marker-Embedded Tactile Image Generation via Generative Adversarial Networks. *IEEE Robotics and Automation Letters*, 8(8):4481–4488, August 2023. ISSN 2377-3766. doi: 10.1109/LRA.2023.3284370.

[17] Zhiqian Lan, Yuxuan Jiang, Ruiqi Wang, Xuanbing Xie, Rongkui Zhang, Yicheng Zhu, Peihang Li, Tianshuo Yang, Tianxing Chen, Haoyu Gao, et al. Autobio: A simulation and benchmark for robotic automation in digital biology laboratory. *arXiv preprint arXiv:2505.14030*, 2025.

[18] Chuanyu Li, Renjun Dang, Xiang Li, Zhiyuan Wu, Jing Xu, Hamidreza Kasaei, Roberto Calandra, Nathan Lepora, Shan Luo, Hao Su, and Rui Chen. ManiSkill-ViTac 2025: Challenge on Manipulation Skill Learning With Vision and Tactile Sensing, 2024-11-19. URL http://arxiv.org/abs/2411.12503.

[19] Haozhan Li, Yuxin Zuo, Jiale Yu, Yuhao Zhang, Zhaohui Yang, Kaiyan Zhang, Xuekai Zhu, Yuchen Zhang, Tianxing Chen, Ganqu Cui, et al. Simplevla-rl: Scaling vla training via reinforcement learning. *arXiv preprint arXiv:2509.09674*, 2025.

[20] Minchen Li, Zachary Ferguson, Teseo Schneider, Timothy R Langlois, Denis Zorin, Daniele Panozzo, Chenfanfu Jiang, and Danny M Kaufman. Incremental potential contact: intersection-and inversion-free, large-deformation dynamics. *ACM Trans. Graph.*, 39(4):49, 2020.

[21] Qixiu Li, Yaobo Liang, Zeyu Wang, Lin Luo, Xi Chen, Mozheng Liao, Fangyun Wei, Yu Deng, Sicheng Xu, Yizhong Zhang, et al. Cogact: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation. *arXiv preprint arXiv:2411.19650*, 2024.

[22] Yuyang Li, Wenxin Du, Chang Yu, Puhao Li, Zihang Zhao, Tengyu Liu, Chenfanfu Jiang, Yixin Zhu, and Siyuan Huang. Taccel: Scaling Up Vision-based Tactile Robotics via High-performance GPU Simulation, 2025-04-17. URL http://arxiv.org/abs/2504.12908.

[23] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36:44776–44791, 2023.

[24] 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, September 2025.

[25] Jiaming Liu, Hao Chen, Pengju An, Zhuoyang Liu, Renrui Zhang, Chenyang Gu, Xiaoqi Li, Ziyu Guo, Sixiang Chen, Mengzhen Liu, et al. Hybridvla: Collaborative diffusion and autoregression in a unified vision-language-action model. *arXiv preprint arXiv:2503.10631*, 2025.

[26] Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. Rdt-1b: a diffusion foundation model for bimanual manipulation. *arXiv preprint arXiv:2410.07864*, 2024.

[27] Yibin Liu, Zhixuan Liang, Zanxin Chen, Tianxing Chen, Mengkang Hu, Wanxi Dong, Congsheng Xu, Zhaoming Han, Yusen Qin, and Yao Mu. Hycodepolicy: Hybrid language controllers for multimodal monitoring and decision in embodied agents. *arXiv preprint arXiv:2508.02629*, 2025.

[28] Ajay Mandlekar, Soroush Nasiriany, Bowen Wen, Iretiayo Akinola, Yashraj Narang, Linxi Fan, Yuke Zhu, and Dieter Fox. Mimicgen: A data generation system for scalable robot learning using human demonstrations. *arXiv preprint arXiv:2310.17596*, 2023.

[29] Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters*, 7(3):7327–7334, 2022.

[30] Yao Mu, Tianxing Chen, Shijia Peng, Zanxin 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.

[31] Yao Mu, Tianxing Chen, Zanxin Chen, Shijia Peng, Zhiqian Lan, Zeyu Gao, Zhixuan Liang, Qiaojun Yu, Yude Zou, Mingkun Xu, et al. Robotwin: Dual-arm robot benchmark with generative digital twins. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 27649–27660, 2025.

[32] Soroush Nasiriany, Abhiram Maddukuri, Lance Zhang, Adeet Parikh, Aaron Lo, Abhishek Joshi, Ajay Mandlekar, and Yuke Zhu. Robocasa: Large-scale simulation of everyday tasks for generalist robots. In *Robotics: Science and Systems (RSS)*, 2024.

[33] Duc Huy Nguyen, Tim Schneider, Guillaume Duret, Alap Kshirsagar, Boris Belousov, and Jan Peters. TacEx: GelSight Tactile Simulation in Isaac Sim – Combining Soft-Body and Visuotactile Simulators, 2024-11-07. URL http://arxiv.org/abs/2411.04776.

[34] NVIDIA. Isaac Sim. URL https://github.com/isaac-sim/IsaacSim.

[35] Wilbert Pumacay, Ishika Singh, Jiafei Duan, Ranjay Krishna, Jesse Thomason, and Dieter Fox. The colosseum: A benchmark for evaluating generalization for robotic manipulation. *arXiv preprint arXiv:2402.08191*, 2024.

[36] ViTai Robotics. Vitai-sdk-release. https://github.com/ViTai-Tech/ViTai-SDK-Release, 2026.

[37] Zilin Si, Gu Zhang, Qingwei Ben, Branden Romero, Zhou Xian, Chao Liu, and Chuang Gan. DIFFTACTILE: A PHYSICS-BASED DIFFERENTIABLE TACTILE SIMULATOR FOR CONTACT-RICH ROBOTIC MANIPULATION. 2024.

[38] Balakumar Sundaralingam, Siva Kumar Sastry Hari, Adam Fishman, Caelan Garrett, Karl Van Wyk, Valts Blukis, Alexander Millane, Helen Oleynikova, Ankur Handa, Fabio Ramos, Nathan Ratliff, and Dieter Fox. cuRobo: Parallelized Collision-Free Minimum-Jerk Robot Motion Generation, November 2023.

[39] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024.

[40] RDT Team. Rdt2: Enabling zero-shot cross-embodiment generalization by scaling up umi data, September 2025. URL https://github.com/thu-ml/RDT2.

[41] Yufei Wang, Zhou Xian, Feng Chen, Tsun-Hsuan Wang, Yian Wang, Katerina Fragkiadaki, Zackory Erickson, David Held, and Chuang Gan. Robogen: Towards unleashing infinite data for automated robot learning via generative simulation. *arXiv preprint arXiv:2311.01455*, 2023.

[42] Junjie Wen, Yichen Zhu, Jinming Li, Zhibin Tang, Chaomin Shen, and Feifei Feng. Dexvla: Vision-

language model with plug-in diffusion expert for general robot control. *arXiv preprint arXiv:2502.05855*, 2025.

[43] Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, et al. Sapien: A simulated part-based interactive environment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11097–11107, 2020.

[44] Zhengtong Xu, Raghava Uppuluri, Xinwei Zhang, Cael Fitch, Philip Glen Crandall, Wan Shou, Dongyi Wang, and Yu She. UniT: Data Efficient Tactile Representation with Generalization to Unseen Objects, April 2025.

[45] 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, April 2025.

[46] Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on robot learning*, pages 1094–1100. PMLR, 2020.

[47] Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on robot learning*, pages 1094–1100. PMLR, 2020.

[48] Wenzhen Yuan, Siyuan Dong, and Edward H Adelson. Gelsight: High-resolution robot tactile sensors for estimating geometry and force. *Sensors*, 17(12):2762, 2017.

[49] Chaofan Zhang, Shaowei Cui, Jingyi Hu, Tianyu Jiang, Tiandong Zhang, Rui Wang, and Shuo Wang. TacFlex: Multimode Tactile Imprints Simulation for Visuotactile Sensors With Coating Patterns. *IEEE Transactions on Robotics*, 41:3965–3985, 2025. ISSN 1941-0468. doi: 10.1109/TRO.2025.3576970. URL https://ieeexplore.ieee.org/document/11024236/.

[50] Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware, April 2023.

APPENDIX

TABLE IV
TASK DESCRIPTIONS FOR UNIVTAC BENCHMARK.

Task	**Description**
Lift Bottle	A bottle rests on the ground plane facing a vertical wall. The robot grasps the bottle and lifts it vertically, keeping its final base within 5 cm of the wall.
Pull-out Key	A key sits in a slot with random initial rotation. The robot rotates the key until sensing mechanical resistance, then pulls it straight out.
Lift Can	One of three cylindrical cans (4, 5, and 6 cm diameter) lie horizontally on the ground plane. The robot grasps the can and lifts it vertically without slippage.
Put Bottle in Shelf	A bottle stands upright before a shelf. The robot grasps the bottle, then positions it into the shelf cavity.
Insert Hole	An cube with inclined hole (60° or 120° orientation) lies on the ground plane. The robot explores the hole's geometry through contact, determines its orientation, and inserts a test tube into the hole.
Insert HDMI	A HDMI connector held by the robot has random rotational offset. The robot aligns and inserts the connector into a fixed slot under rotational uncertainty.
Insert Tube	A narrow hole (2.05 cm diameter) appears on an inclined surface. The robot inserts a 2.0 cm diameter test tube from a near-center starting position through the tight clearance.
Classify Texture	Two cylindrical objects share similar visual appearance but differ in surface texture. The robot first performs tactile contact to perceive texture, then classifies each cylinder and places it at the pre-specified goal region for its class.