$\mathcal{N}_0$-Foundation: Towards the Age of Tactile Intelligence

NeoteAI Team & Fudan TEAI Team

We present $\mathcal{N}_0$-Foundation, a paradigm for tactile-enabled embodied manipulation, which integrates tactile sensing hardware, large-scale multimodal data, tactile representation learning, and standardized evaluation. First, we engineer the underlying tactile infrastructure for scalable data collection and practical deployment, including a vision-based tactile sensor, a Tactile Universal Manipulation Interface ($\mathcal{N}_0$-TacUMI), and a synchronized visual–tactile data collection system supporting both multiple robot embodiments and $\mathcal{N}_0$-TacUMI-based demonstrations. Leveraging this infrastructure, we construct NeoData, which contains more than 30,000 hours of synchronized visual and tactile demonstrations, spanning six embodiments, 450 tasks, and billions of paired RGB and tactile frames collected through a mixture of real-robot teleoperation and $\mathcal{N}_0$-TacUMI demonstrations. To facilitate open research, we further release OpenNeoData, a 5,000-hour open-source subset of NeoData that spans six embodiments, more than 250 tasks, and over 200 skills. The dataset addresses a central limitation of existing manipulation corpora: contact states, local forces, and incipient slip are rarely observable from vision alone, yet they are critical for deformable-object manipulation, precise assembly, delicate force control, and sustained surface interaction. Capitalizing on the large-scale, heterogeneous tactile measurements within NeoData, we propose NeoForce, a visuo-tactile representation learning model that uses dense three-axis force fields as unified, hardware-agnostic supervision to learn transferable tactile representations across different sensor designs for downstream embodied policies. To enable systematic evaluation of tactile embodied models built upon our infrastructure, datasets and tactile representations, we further propose a comprehensive benchmark, which combines the real-world NeoReal suite and the simulated NeoSim suite for standardized evaluation of contact-rich manipulation. Experiments across both suites show that policies benefit from the physical contact state rather than from the device-specific appearance of the tactile signal. These results indicate that large-scale tactile data and supervision make a universal tactile representation possible, which in turn enables embodied policies to perform fine-grained contact-rich manipulation. We release the dataset, the representation, and the benchmark, aiming at supporting future work on tactile-enabled embodied manipulation.

Date: July 25, 2026
Website: http://research.neoteai.com/n0-foundation
Dataset: https://huggingface.co/datasets/NeoteAIEmbodied/OpenNeoData

NeoteAI

Institute of Trustworthy Embodied AI, Fudan University

1 Introduction

Learning generalizable manipulation policies from large-scale demonstration data has become a central paradigm in embodied intelligence. Recent progress in vision-language-action (VLA) models has demonstrated that scaling diverse multimodal demonstrations substantially improves the generalization and robustness of visuomotor policies [3, 52, 54]. Complementing direct policy learning, World Action Models learn action-conditioned environment dynamics by modeling how the future world evolves under robot actions, showing promising generalization across real-world tasks and environments [34, 68, 80]. Progress in both paradigms is increasingly driven by the scale and diversity of available manipulation data. Real-robot datasets range from teleoperated single-embodiment corpora [45, 66] to cross-embodiment collections gathered across multiple robots and institutions [28, 52]. Portable collection systems such as the UMI [10] further enable scalable in-the-wild demonstration collection beyond fixed robot workcells and have inspired a growing family of embodiment-agnostic interfaces [21, 38, 75, 85]. Meanwhile, large-scale egocentric videos provide

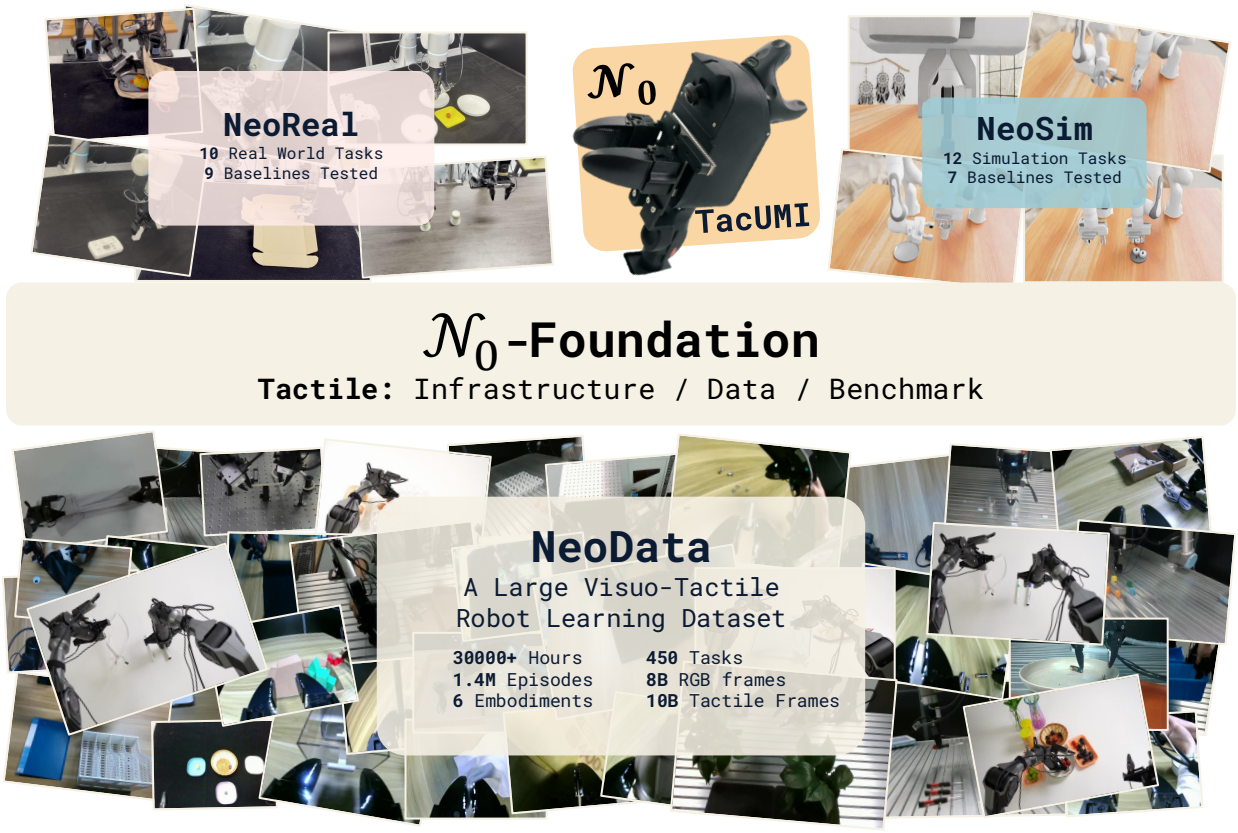


Figure 1 $\mathcal{N}_0$**-Foundation**, a tactile-centric foundation for embodied manipulation that unifies hardware, data, tactile representation and evaluation. At its core is NeoData, a large-scale visuo-tactile robot learning dataset with more than 30,000 hours of interaction, 450 tasks, 1.4+M episodes, 8B RGB frames, and 10B tactile frames gathered from a mixture of real-robot embodiments and $\mathcal{N}_0$-TacUMI collection. On top of the data we evaluate tactile-based policies on the proposed real-world NeoReal suite and the simulated NeoSim suite.

abundant observations of human–object interactions, offering complementary behavioral and semantic priors for embodied learning [12, 24, 55].

Despite this progress, two limitations remain. First, most large-scale embodied datasets are dominated by visual observations. However, vision alone is often insufficient for contact-rich manipulation, where success depends on local force, friction, incipient slip, and contact transitions that are weakly observable from external or wrist-mounted cameras. These limitations affect a broad range of task families, including deformable-object manipulation, precise insertion and assembly, delicate force control, sustained surface interaction, and bimanual coordination. Second, the tactile modality remains fragmented. Camera-based visuo-tactile sensors [30, 84], capacitive arrays [44, 56], and fiber-based sensors [74] produce signals in different hardware-specific formats. This heterogeneity prevents tactile demonstrations collected on one device from being reused directly on another, which complicates policy training over pooled tactile data and ultimately prevents tactile datasets and models from benefiting from scale.

To address these limitations, $\mathcal{N}_0$-Foundation, pronounced as Neo-Foundation, establishes a tactile-centric foundation for embodied manipulation. As summarized in Fig. 1, $\mathcal{N}_0$-Foundation establishes a scalable paradigm for tactile-enabled embodied learning that connects scalable tactile data infrastructure, large-scale multimodal data acquisition, unified representation learning, and standardized policy evaluation. $\mathcal{N}_0$-Foundation is organized around four technical pillars: tactile infrastructure, a large-scale dataset named NeoData, tactile representation learning with NeoForce, and a comprehensive evaluation protocol. First, we build the underlying tactile infrastructure required for scalable collection and practical deployment. It includes durable vision-based tactile sensors, $\mathcal{N}_0$-TacUMI, and a synchronized visual–tactile data collection

system supporting both real-robot demonstrations across multiple embodiments and portable UMI-based demonstrations. Built on this infrastructure, NeoData contains more than 30,000 hours of synchronized visual and tactile demonstrations, spanning six embodiments, 450+ tasks, and billions of paired RGB and tactile frames. The corpus combines real-robot teleoperation with UMI demonstrations, jointly providing physically grounded robot executions and scalable human-operated trajectories. By recording tactile signals alongside visual observations, NeoData captures contact states, local forces, and incipient slip that are difficult to infer from vision alone. To support open research, we further release OpenNeoData, a 5,000-hour open-source subset spanning six embodiments, more than 250 tasks, and over 200 skills. Based on the large-scale and heterogeneous tactile data in NeoData, we introduce a unified, hardware-agnostic tactile representation learning formulation that uses dense three-axis force fields as a common physical supervision space across different tactile sensor designs. Under this formulation, we develop NeoForce, a visuo-tactile representation model that learns transferable and temporally structured tactile representations for downstream embodied policies. Finally, we provide standardized evaluation for contact-rich manipulation by combining the real-world NeoReal suite with the simulated NeoSim suite.

In summary, this report makes the following five contributions:

- **An overall tactile-centric paradigm for embodied manipulation.** We establish $\mathcal{N}_0$-Foundation, a comprehensive embodied paradigm that unifies a scalable tactile infrastructure, large-scale tactile data acquisition, hardware-agnostic tactile representation learning, and standardized policy evaluation.
- **A scalable tactile-based embodied infrastructure.** We develop a comprehensive data collection hardware and pipeline, including durable vision-based tactile sensors and $\mathcal{N}_0$-TacUMI, a sensor-integrated UMI device that enables efficient tactile manipulation data collection at scale.
- **A large-scale cross-embodiment visuo-tactile dataset.** We release *NeoData*, a large-scale cross-embodiment corpus that combines real-robot teleoperation and UMI demonstrations, covers six embodiments and 450+ tasks, and augments RGB observations with synchronized tactile sensing for contact-rich manipulation.
- **A unified force-based tactile representation.** We develop *NeoForce*, a tactile-oriented model that unifies versatile tactile signals into a hardware-agnostic representation via dense three-axis force fields, providing a transferable interface for tactile application across embodied models.
- **A standardized evaluation protocol.** We comprehensively evaluate tactile-aware policies on the proposed real-world *NeoReal* and simulated *NeoSim* test suites, demonstrating the effectiveness of large-scale tactile learning and unified tactile representation learning.

Collectively, these contributions establish a scalable path for incorporating tactile perception into large-scale embodied learning, enabling policies to acquire fine-grained contact-rich manipulation capabilities beyond vision alone.

2 Related Work

2.1 Embodiment Hardware and Data Collection

Embodied manipulation systems couple the robot body, sensing stack, and demonstration interface. Conventional teleoperation records executable robot actions directly but binds collection to particular arms and grippers. UMI [10] instead uses a handheld gripper and relative end-effector commands to decouple demonstration collection from physical robots, and subsequent systems extend this principle to mobile, dexterous, and hardware-independent platforms [21, 38, 75]. Portable visuo-tactile interfaces (Touch in the Wild [93], FreeTacMan [69], ViTaMIn-B [32], TacUMI [8], and exUMI [77]) further attach contact sensors to handheld or wearable collectors. TAMEn [70] additionally combines a cross-morphology wearable interface, MoCap/VR tracking, online feasibility checking, and tactile recovery teleoperation in a closed-loop collection engine. Yet these systems generally target a few robot platforms and retain device-specific tactile streams. NeoData combines robot-free collection with multi-robot teleoperation and maps six embodiments to a common action and observation schema.

2.2 Embodied Manipulation Datasets

Large-scale datasets such as BridgeData V2 [66], Open X-Embodiment [52], and DROID [28] establish data scale and embodiment diversity as central ingredients of general-purpose robot learning, but their observations remain dominated by vision and proprioception. RH20T is a multimodal exception, providing more than 110k real-robot sequences over 147 tasks with synchronized vision, force, audio, and action streams, although fingertip tactile signals appear in only one of its seven configurations [14]. At a substantially larger scale, AgiBot World provides more than one million teleoperated trajectories spanning 217 tasks and 2,976.4 hours. It uses over 100 homogeneous mobile dual-arm robots equipped with RGB-based visuo-tactile sensors, and applies human-in-the-loop quality verification [57]. The portable visuo-tactile datasets above improve contact coverage but remain at the scale of 10^2–10^4 demonstrations and split across heterogeneous tactile encodings, ranging from low-resolution taxel arrays to camera-based RGB streams. Recent corpora such as XenseRobotics-TacVerse [63] and Daimon-Infinity [11] substantially expand the scale and duration of robot-free visuo-tactile collection, but contain no demonstrations executed on physical robot platforms. These datasets are also typically gathered under a single acquisition mode, either teleoperation or a robot-free handheld device, and seldom document any per-episode quality control. Most closely related, OmniVTA [91] contains 21,879 visuo-tactile-action trajectories over 86 tasks and 126 objects and is the only prior corpus that pairs robot-based and handheld collection, though still without verified quality control. As Tab. 1 summarizes, NeoData scales this convergence to 1.4M+ episodes over 450 tasks and more than 30,000 hours. It combines on-robot teleoperated demonstrations with scalable robot-free collection across six embodiment types, standardizes every finger on camera-based tactile RGB frames that are further converted into a unified three-axis force representation, and passes each episode through automatic and human quality control.

Dataset	Traj. / tasks	Time	Embodiment Types	Tactile signal	Control Type	Verified QC
Large-scale robot manipulation corpora						
BridgeData V2 [66]	60k / 13	–	1	–	Tele	–
Open X-Embodiment [52]	1M+ / –	–	22	–	Tele	–
DROID [28]	76k / 86	350 h	1	–	Tele	–
RH20T [14]	110k / 147	–	4	taxel	Tele	Human
AgiBot World [57]	1M+ / 217	2,976.4 h	1	RGB	Tele	Human
Robot-free collection interface						
UMI [10]	1.4k / –	12 h	2	–	RF	–
Visuo-tactile manipulation corpora						
Touch in the Wild [93]	2.7k+ / 43	31 h	1	taxel	RF	–
FreeTacMan [69]	10k+ / 50	27.8 h	2	RGB	RF	–
ViTaMIn-B [32]	0.8k / 4	2.9 h	2	RGB	RF	–
TacUMI [8]	30k / 1	0.5 h	1	RGB	RF	–
exUMI [77]	1.7k / 9	5 h	1	RGB	RF	–
XenseRobotics-TacVerse [63]	8.2k / 128	200 h	1	RGB	RF	–
Daimon-Infinity [11]	276k / -	1,209 h	1	RGB	RF	–
NeoData (ours)	**1.4M+ / 450+**	**>30,000 h**	6	RGB	**Tele+RF**	**Auto+Human**

Table 1 Representative manipulation datasets in three families. Traj. / tasks: counts as stated by each paper (BridgeData V2: skills; TacUMI: frames), rounded to thousands (k). Time: total interaction time in hours, estimated from frame counts and rates where not directly reported. Embodiment Types: number of distinct embodiment platforms ($N\times$ identical arms count once). Tactile signal: released contact modality, grouped as camera-based RGB or taxel arrays. Control Type: collection mode, teleoperation (Tele), robot-free handheld (RF), or both (Tele+RF). Verified QC: documented per-episode quality control. "–": not reported or not applicable. Red = column best, blue = second best.

2.3 Robot Learning in Representation and Manipulation

Robot learning increasingly combines scalable policy learning with reusable perception. RT-1 [3] and Diffusion Policy [9] demonstrate effective action modeling for real-world manipulation, and VLA models such as

$\pi_{0.5}$ scale this recipe into generalist policies [54]. Recent embodied foundation models extend this scaling thesis. Xiaomi-Robotics-0 couples flow-matching action generation with asynchronous inference for real-time bimanual control [5], Xiaomi-Robotics-1 pre-trains on more than 100k hours of real-world trajectories collected with handheld UMI devices [73], LingBot-VLA 2.0 curates roughly 60,000 hours of robot and egocentric human data spanning 20 robot configurations and whole-body action spaces [72], and the HY-Embodied series builds Mixture-of-Transformers (MoT) and Mixture-of-Experts (MoE) vision language models that support downstream VLA training [65, 67], while concurrent systems unify understanding, generation, and action or simplify the VLA stack [4, 79]. A complementary line uses video and world modeling for control. LingBot-Vision pretrains dense spatial representations via masked boundary modeling [16], LingBot-Video scales MoE video pretraining on robot-oriented footage [43], causal video-action models such as LingBot-VA 2.0 predict scene evolution jointly with actions [33, 86], world-action models integrate world prediction into policy learning [39, 51, 82], and interactive or 3D world models serve as simulators and embodied data engines [18, 64], with Xiaomi-Robotics-U0 substantially improving the out-of-distribution success rate of $\pi_{0.5}$ through synthesized embodied data [36]. Across these foundation-model efforts, however, observations and pretraining corpora remain visual, linguistic, and proprioceptive, while contact is neither sensed nor standardized, and TouchGuide's tactile contact model, which steers pretrained visuomotor policies at inference time, remains a rare exception [88].

On the perception side, masked reconstruction, joint-embedding prediction, and self-supervised visual backbones provide transferable representation-learning objectives [2, 53]. Tactile observations are harder to reuse, because raw images from sensors such as GelSight [84] and DIGIT [30] entangle contact with sensor-specific optics, geometry, and elastomer appearance. Prior work reduces this dependence through canonical force-based pretraining [71], sensor-invariant contact geometry [20], action-aware tactile latents [76, 77], or continuous deformation representations [91]. NeoData instead uses a calibrated field of x-, y-, and z-axis forces as the dataset-level interchange format, and NeoForce learns over this shared physical interface so that heterogeneous tactile demonstrations can support a common manipulation policy.

2.4 Embodied Manipulation Benchmarks

Simulation suites such as RLBench [25], Meta-World [81], CALVIN [46], ManiSkill2 [19], and LIBERO [37] standardize multi-task, language-conditioned, and lifelong manipulation, and household-scale successors further extend task and scene diversity [31, 48, 49, 87]. Perturbation studies show, however, that near-saturated scores on such suites can reflect overfitting to the specific task configurations seen during training rather than task understanding [15, 92], motivating capability-level diagnosis of generalist policies [17]. Evaluation is therefore anchoring itself to physical hardware. FurnitureBench [23] standardizes long-horizon real-world assembly, SIMPLER [35] builds simulated twins whose scores predict real performance, and RoboDojo [7] and DuoBench [27] couple simulated and real task suites under shared protocols, which is the paired sim-and-real methodology that we adopt. All of these benchmarks, however, observe only vision and proprioception, so contact quality is neither sensed nor scored.

Tactile manipulation benchmarks remain scarce because simulation must reproduce both contact mechanics and sensor deformation. Taxim [59] and TacSL [1] simulate the appearance of specific vision-based tactile sensors, UniVTAC [6] unifies multi-sensor tactile simulation with an eight-task benchmark, ContactWorld [89] evaluates visuo-tactile world models on twelve simulated tasks and one real task, and SoftVTBench [26] finds that policies often succeed while violating deformation limits, a failure that integrating tactile sensing substantially reduces. These efforts substantially expand tactile evaluation within their own scope, but their data remains tied to a particular sensor or confined to simulation, so the resulting protocols do not transfer across tactile hardware or to physical deployment. In our work, both NeoReal and NeoSim are paired for policy evaluation, so that tactile feedback and representation choices can be compared reproducibly without tying evaluation to the appearance of any particular sensor.

Figure 2 Overview of $\mathcal{N}_0$-TacUMI. A single handheld unit records complementary streams: a fisheye camera captures the wide-angle wrist view, two camera-based tactile sensors capture contact-rich deformation at the fingertips, an infrared tracker with base stations provides a low-drift six-degree-of-freedom pose, and a magnetic encoder measures the gripper aperture through a rack-and-pinion transmission. The inset shows the layered structure of a tactile sensor, comprising a glass panel, an elastic sensing layer, the main body with an embedded RGB camera, and a controller board.

3 Data Acquisition Hardware

3.1 Overview

NeoData spans six embodiments, comprising five robot platforms (Franka Research 3, Piper, ARX X5, UR5e, and Flexiv Rizon 4s robots) and our handheld $\mathcal{N}_0$-TacUMI device. To collect comprehensive, high-quality tactile demonstrations, we adopt two complementary acquisition settings. For the five robot embodiments, tactile sensors are mounted directly on the physical grippers, and demonstrations are collected through teleoperation on the corresponding target platforms. In parallel, $\mathcal{N}_0$-TacUMI is a human-operated handheld parallel gripper developed following the Universal Manipulation Interface philosophy [10], enabling scalable demonstration collection beyond fixed robot workcells. All six embodiments share aligned tactile sensing, visual observations, and action conventions, allowing NeoData to integrate physically grounded robot executions and scalable TacUMI demonstrations within a unified data schema.

3.2 Tactile Sensor

Our tactile sensor follows the camera-based visuo-tactile principle [30, 40–42, 84], in which contact information is recovered from the optical distortion of a deformable sensing surface. As illustrated in the inset of Fig. 2, the sensor is a layered assembly. A thin wear-resistant glass panel protects the contact surface against abrasion during repeated interaction. Beneath it, an elastic sensing layer deforms under external load and carries a textured pattern whose displacement encodes the applied force. An RGB camera embedded in the main body observes the underside of the sensing layer, while a compact controller board handles illumination, image acquisition, and data transmission.

When the finger contacts an object, local pressure and shear deform the sensing layer, and the camera captures the resulting texture displacement as a tactile image $T_t \in \mathbb{R}^{H\times W\times 3}$. The tactile signal is therefore a visual measurement of deformation. Let Φ denote the elastomer force-to-deformation transfer function, which maps a spatial force distribution F to a deformation field, and let $\mathcal{C}$ denote the imaging process of the

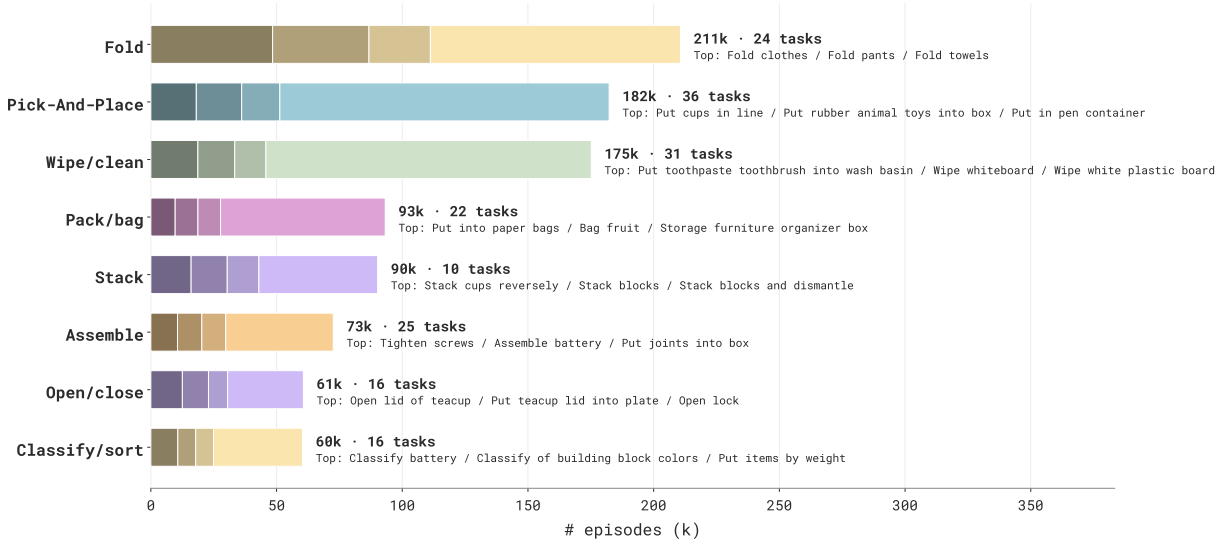


Figure 3 Task coverage of NeoData. The more than 450 tasks are organized into eight categories, and for each category the most frequent constituent tasks are shown by episode count. The categories span deformable-object skills, high-precision assembly, everyday transport, surface interaction, and semantic manipulation.

internal camera. The observed tactile image can be written as

$$T_t = \mathcal{C}\big(\Phi(F_t)\big), \tag{1}$$

so that recovering the underlying force field F_t from T_t amounts to inverting the composed transduction and imaging operators. We learn this inverse mapping explicitly and use its output as the unified tactile representation, as detailed in Section 5.

3.3 $\mathcal{N}_0$-TacUMI

Building on the tactile sensor, we develop $\mathcal{N}_0$-TacUMI for robot-free data collection, shown in Fig. 2. The device consists of a parallel gripper equipped with two tactile sensors, a 160° FOV fisheye camera that captures the wide-angle wrist view, an infrared tracker for spatial localization, and a magnetic encoder for physical aperture measurement. The tracker provides the six-degree-of-freedom gripper pose in a global tracking frame, while the magnetic encoder reads the instantaneous finger separation through a rack-and-pinion transmission.

$\mathcal{N}_0$-TacUMI records synchronized wrist images, left and right tactile images, gripper aperture, and relative end-effector motion. Following the UMI convention [10, 29], actions are represented by the end-effector pose and the gripper width. This convention avoids binding the demonstration to a specific robot base frame and makes the handheld trajectories compatible with the teleoperated robot data described in Section 4. The detailed observation and action definitions are provided in Section 4.2.

4 NeoData

4.1 Overview

We introduce NeoData, one of the largest tactile-enabled real-world manipulation datasets constructed to date. As summarized in Fig. 1, the corpus aggregates more than 30,000 hours of interaction into 1.4M episodes and 3.3B timesteps, yielding 8B RGB frames paired with 10B tactile frames. The data span six embodiments, including the handheld $\mathcal{N}_0$-TacUMI device and five robot platforms: ARX X5, UR5e, Flexiv Rizon 4s, Piper and Franka Research 3 robots, and were collected by nearly 100 human operators. Every episode contains synchronized visual and tactile streams, so the contact signal is available throughout the manipulation process rather than only at discrete grasp events. To facilitate open research, we release

OpenNeoData, a 5,000-hour open-source subset of NeoData that spans six embodiments, more than 250 tasks, and over 200 skills, providing an accessible entry point to the full corpus.

Tab. 1 situates NeoData among representative manipulation datasets. Compared with large visual corpora, NeoData adds dense tactile observations at comparable scale and covers a broad set of contact-rich skills. Compared with existing tactile datasets and collection systems, it provides larger episode volume, greater embodiment diversity, and a mixture of real-robot and UMI demonstrations. The dataset is therefore designed to support policy learning in settings where tactile feedback is not incidental but central to the manipulation objective.

4.2 Data Format

Every NeoData episode is stored as a synchronized sequence of observations and actions. For the five robot embodiments, the observation at timestep t contains an external third-person RGB image I_t^{ext}, a wrist RGB image I_t^{wrist}, and the left and right tactile images, denoted as T_t^l and T_t^r, respectively:

$$o_t^{\text{robot}} = \left(I_t^{\text{ext}},\ I_t^{\text{wrist}},\ T_t^l,\ T_t^r\right), \tag{2}$$

where each image stream is temporally aligned and the tactile images are recorded at a native resolution of 640×360. The robot action a_t^{robot} contains the end-effector command in the tool center point frame and the absolute joint command:

$$a_t^{\text{robot}} = \left(\Delta x_t^{\text{tcp}},\ q_{t+1}\right), \tag{3}$$

where Δx_t^{tcp} denotes the commanded Cartesian end-effector motion and q_{t+1} denotes the target joint configuration for the next control step. We represent rotations using the continuous 6D representation to avoid the discontinuities of Euler-angle and quaternion parameterizations, as well as the gimbal-lock issue associated with Euler angles. This format preserves both the operational action used by task policies and the absolute robot state needed for embodiment-specific replay or analysis.

For $\mathcal{N}_0$-TacUMI data, the observation o_t^{UMI} contains a fisheye wrist image I_t^{fisheye} and the left and right tactile images:

$$o_t^{\text{UMI}} = \left(I_t^{\text{fisheye}},\ T_t^l,\ T_t^r\right). \tag{4}$$

The corresponding action records the relative end-effector motion between the current frame and a future frame, together with the target gripper aperture:

$$a_t^{\text{UMI}} = \left(\Delta x_{t\to t+k},\ g_{t+k}\right), \tag{5}$$

where $\Delta x_{t\to t+k}$ denotes the relative motion between frame t and $t+k$, and g_{t+k} denotes the gripper aperture at frame $t+k$. This relative action convention follows UMI and supports transfer from handheld demonstrations to robot embodiments.

4.3 Dataset Composition

Fig. 4 summarizes the embodiment, object, scene, and duration distributions, while Fig. 3 summarizes task coverage. $\mathcal{N}_0$-TacUMI contributes 57% of all trajectories, reflecting the scalability of handheld collection for dexterous and repetitive tasks. The five robot embodiments provide the remaining trajectories and anchor the corpus in executable robot behavior. This mixture allows NeoData to combine the efficiency of UMI demonstrations with the embodiment fidelity of robot teleoperation. The object distribution is likewise long-tailed, with cups and boxes as the most frequent object categories, followed by test tubes and building blocks.

The task distribution deliberately emphasizes tactile-relevant manipulation. High volume categories include folding, pick-and-place, wiping or cleaning, stacking, assembling, inserting, pouring, opening, closing, and classification. These categories cover deformable-object manipulation, precise mating, delicate grasping, sustained surface contact, and manipulation under ambiguous visual states. Scenes include tabletop, workbench, laundry, kitchen, lab bench, and storage environments, with objects spanning containers, tools, deformables, kitchenware, electronics, and lab items. Episode durations peak near 80 seconds and include a long tail of multistep interactions, which supports learning over both short contact primitives and extended task sequences.

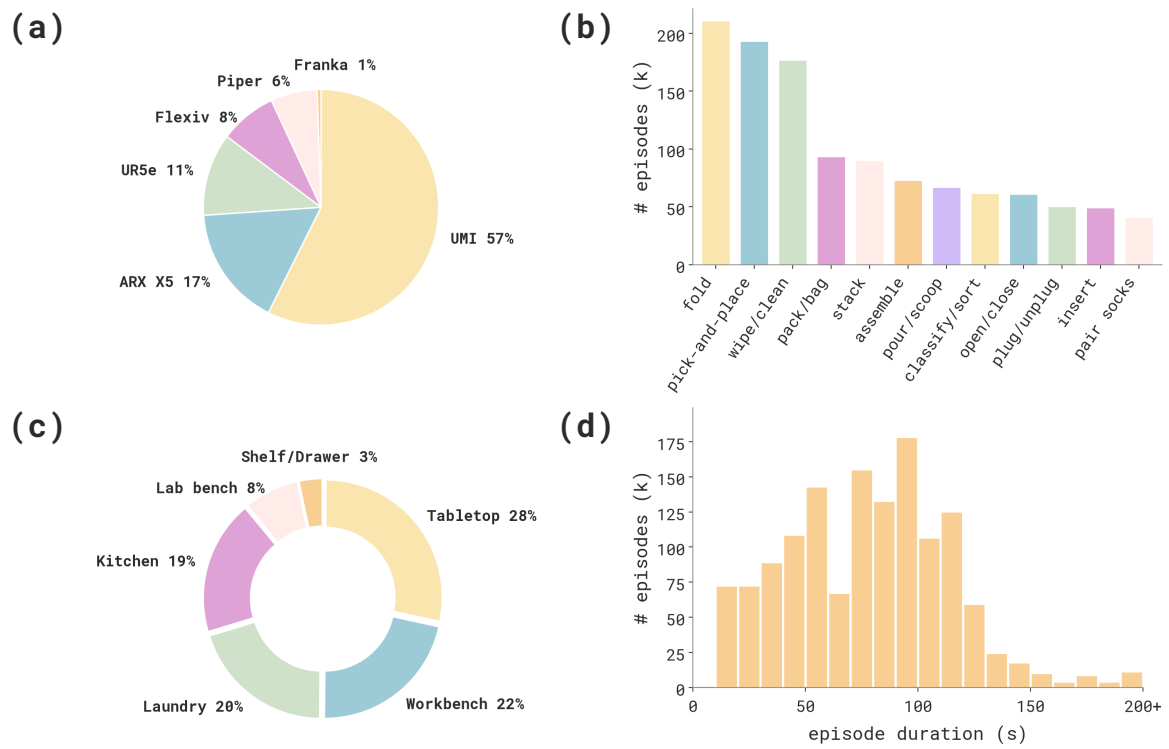


Figure 4 NeoData statistics. (a) Trajectories per embodiment: $\mathcal{N}_0$-TacUMI contributes 57% of the data, followed by ARX X5 (17%), UR5e (11%), Flexiv Rizon 4s (8%), Piper (6%), and Franka Research 3 (1%). (b) Episode counts for the twelve most frequent object categories, led by cups (33.8k), boxes (32.6k), test tubes (22.3k), and building blocks (21.1k). (c) Scene distribution across tabletop, workbench, laundry, kitchen, lab bench, and shelf or drawer settings. (d) Episode duration histogram, which peaks near 80 seconds with a long right tail.

4.4 Data Curation and Preprocessing

Fusing real-robot and $\mathcal{N}_0$-TacUMI demonstrations requires effectively integrating cross-embodiment data and unifying it into a consistent format. This motivates a multi-step cleaning pipeline that turns heterogeneous raw recordings into a complete, high-quality corpus suitable for model training. As illustrated in Fig. 5, we summarize this process into a staged quality selection pipeline of five complementary checks.

Data Completeness. The first check verifies that each episode contains the required visual, tactile, action, and metadata streams. Episodes that are missing a required camera, tactile, action, or metadata stream are removed so that each retained sequence provides a complete multimodal record.

Motion Quality. The second check screens for motion artifacts that would destabilize policy training. We estimate trajectory smoothness along each episode and flag low-quality regions such as abrupt tracking pulses and tracking-lost segments, and episodes dominated by these artifacts are removed.

Episode Quality. The third check normalizes the pace of the demonstrations and verifies that each one actually completes its task. For each task, we compute the median episode duration m and the difference between the 75th- and 25th-percentile durations, denoted as $r = q_{75} - q_{25}$. We retain only episodes whose duration falls within $[m - 3r, m + 3r]$. This removes demonstrations that are completed abnormally fast or slow relative to the typical duration range of the corresponding task. Episodes that terminate before the task is completed are also removed.

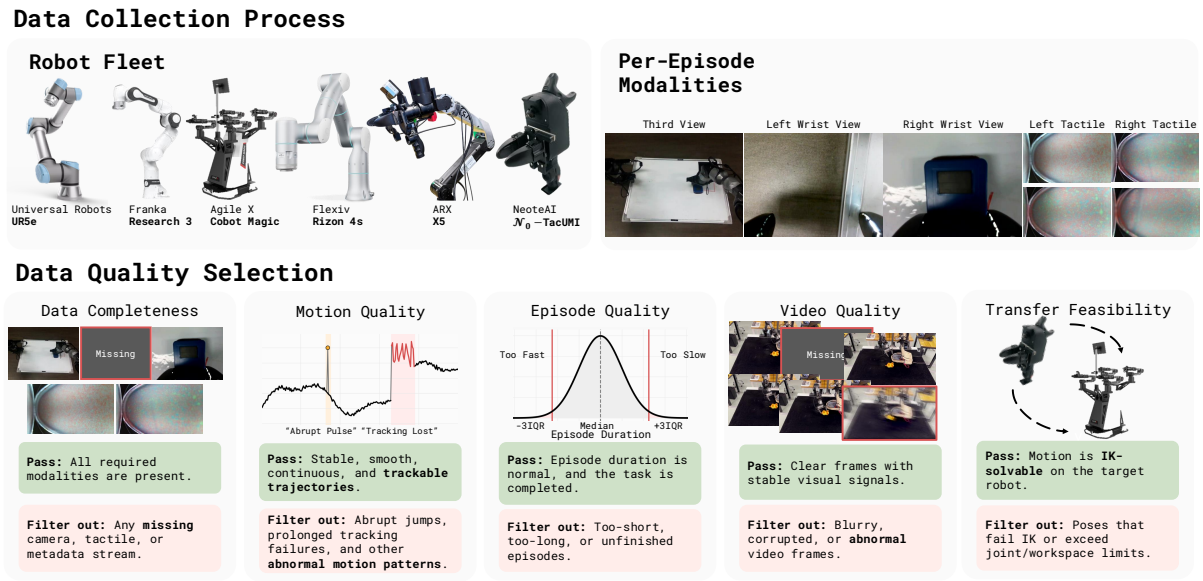


Figure 5 The NeoData curation pipeline. Raw demonstrations are stored as synchronized multimodal episodes and pass through quality checks for stream completeness, motion quality, episode quality, video validity, and transfer feasibility before entering the released corpus.

Video Quality. The fourth check validates the recorded streams and repairs the defects that remain recoverable. We inspect every view for corrupted content, severe blur, dropped frames, and abnormal illumination. In addition, we align the asynchronously recorded modalities to a unified timeline with a strict one to one correspondence between video, tactile and action frames. Recording errors such as swapped camera mappings, mislabeled left and right arms, and abnormal end-effector poses are repaired when possible, for example by remapping the cameras or recovering the pose through forward kinematics, and are otherwise discarded. We finally trim near-static segments, whose nearly identical observations correspond to inconsistent actions and cause the model to pause unnecessarily at inference time.

Transfer Feasibility. A large fraction of NeoData is collected robot-free, so the recorded motions are not guaranteed to be physically realizable on the target robots. The final check therefore runs inverse kinematics on each trajectory and rejects poses that fail to solve or that exceed the joint and workspace limits of the intended embodiment, which preserves the scalability of $\mathcal{N}_0$-TacUMI collection while keeping every retained demonstration executable after transfer. Episodes that pass all five checks constitute the released corpus, complete in modality, well conditioned for training, visually valid, and executable on the target embodiment.

4.5 Data Annotation

We introduce an automatic annotation pipeline that assigns every manipulation episode a four-level hierarchy. The levels are task (L3), subtask (L2), action (L1), and atomic segment (L0), as shown in Fig. 6. The hierarchy separates semantic structure from temporal localization. The template defines what should occur in a task, while signal-based segmentation determines when each event occurs in a specific episode.

The pipeline has two stages. In the first stage, a vision-language model (VLM) summarizes representative episodes into a time-independent task template, which records the task objective, ordered semantic phases, and action-level substeps. A human expert verifies the template to remove semantic ambiguity before it is applied to the full task category. In the second stage, each episode is segmented by fusing action-based and visual event boundaries. The model then labels the resulting fixed intervals according to the verified template, and deterministic code assembles the L0, L1, L2, and L3 temporal extents. This design keeps temporal endpoints grounded in measured signals while using language models for semantic labeling, which reduces temporal drift on long manipulation videos. By segmenting each episode into labeled subtasks,

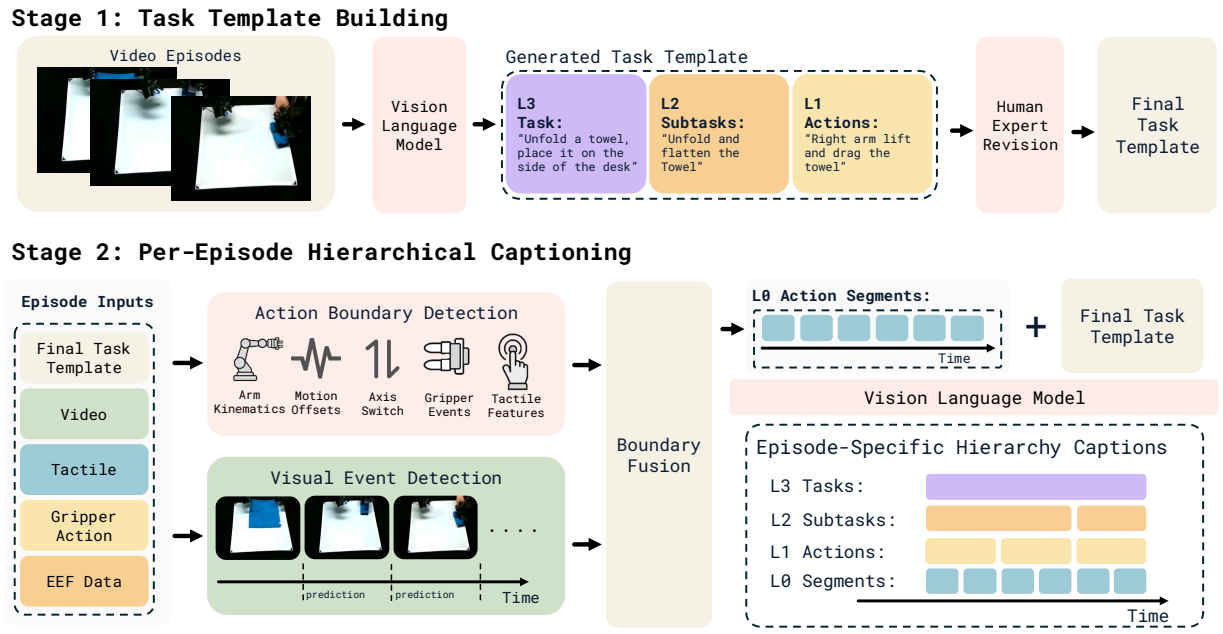


Figure 6 The NeoData hierarchical annotation pipeline. A VLM first proposes a task template from representative episodes, a human expert verifies the template, and each episode is then segmented by signal-based boundaries before the fixed intervals are labeled into a four-level hierarchy.

this pipeline supplies the temporal structure needed for training long-horizon, multi-stage, and multi-task embodied policies.

5 Tactile Representation

This section proposes a tactile representation that allows tactile information to act on embodied manipulation policies, describing every tactile observation as a dense three-axis force field over the sensing surface and thereby expressing contact as a physical quantity rather than as the appearance of a particular device. We first present the motivation and the design principles behind the representation, then define it, describe the data pipeline that produces force fields, and present NeoForce, the model used to learn temporally structured tactile features from these force fields.

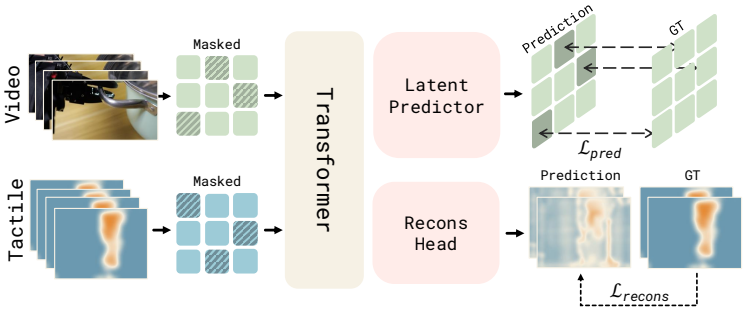


Figure 7 The structure of NeoForce. A chunk of RGB frames and tactile force fields is patchified into visual and tactile tokens and processed by a shared transformer backbone. A reconstruction head decodes force fields, while a latent prediction head predicts teacher latent outputs that serve as ground truth in the latent space.

5.1 Motivation

Unlike visual hardware, which records a consistent image space based on electronic pixels, tactile hardware is fragmented in the form of the signal it emits. Optical gel sensors [30, 84], capacitive arrays [44, 56], and piezoresistive skins [60] differ in units,

resolution, and sensing geometry, so downstream representations and models built on any single signal form are bound to the device that produced it and cannot be carried elsewhere. Scaling tactile manipulation therefore requires converting these heterogeneous signals into one universal form that the tactile community can share. To provide a transferable interface, a universal tactile representation should satisfy three principles.

Physical grounding. The representation should be anchored to what touch physically delivers rather than to an appearance-level proxy. Across various manipulation scenarios, what an end-effector delivers to an object is completely described by its physical effect at the contact interface, so its behavior is expressed by an intrinsic attribute rather than by how the contact is measured through the hardware.

Cross-sensor unification. The representation should be a common space that different tactile devices can utilize. Signals from any sensor with an appropriate calibration procedure should be mappable into the same target space, which makes heterogeneous tactile data directly comparable and jointly trainable.

Spatially rich semantics. Complex contact is spatially structured. The location, shape, and extent of the contact patch, together with how intensity varies inside it, distinguish a full grasp from an edge contact and a stable hold from incipient slip, so the representation must preserve this distribution rather than summarize it.

5.2 Definition

Following the three principles above, we adopt force as the physical attribute that the tactile representation encodes, since it describes the effect of contact independently of the sensor that measured it. At every point of the sensing surface we record the local force as two orthogonal in-plane components tangent to the tactile plane, which capture the shear that drives sliding and grasping, together with one normal component, which captures the pressure exerted during pressing. Sampling this three-dimensional force on a high-resolution lattice yields a dense force field that expresses the complex contact taking place over the sensing surface and retains the spatial distribution of contact intensity. Concretely, for one tactile sensor at time t the force field is

$$F_t \in \mathbb{R}^{H\times W\times 3}, \qquad F_t(u,v) = \big(f_x(u,v),\, f_y(u,v),\, f_z(u,v)\big), \tag{6}$$

where each spatial location (u, v) carries two tangential components f_x, f_y along the x and y axes that encode shear, and one normal component f_z along the z axis that encodes pressure. For a parallel gripper, the left and right tactile fields are represented as F_t^l and F_t^r and may be concatenated into a six-channel tensor. This definition gives downstream embodied models a physically interpretable tactile input that is independent of raw gel appearance.

5.3 Implementation

The force representation is produced directly from raw visuo-tactile images. Let T_t denote the raw tactile image from one sensor. We learn a dense mapping

$$\hat{F}_t = g_\theta(T_t), \qquad g_\theta : \mathbb{R}^{H\times W\times 3} \to \mathbb{R}^{H\times W\times 3}, \tag{7}$$

where $\hat{F}_t$ is the estimated force field. The model is trained on paired visuo-tactile data, and the construction of this dataset is described in the supplementary material.

The resulting force field serves as the tactile representation of the sensor, so that given a raw tactile image, the model outputs a unified three-axis force representation of the contact. This representation also provides the target that NeoForce builds upon, since the model encodes the low-level image observation of the current sensor into a force field. Detailed calibration hardware, data split, model architecture, and preprocessing parameters are provided in Appendix D.

Objective	MAE ↓	RMSE ↓	mIoU ↑
Reconstruction	0.070	0.095	**0.968**
+ latent pred	**0.066**	**0.089**	0.966

Table 2 Ablation of the NeoForce training objective on force field reconstruction. Adding latent prediction preserves reconstruction fidelity while training the representation to model temporal contact evolution.

5.4 NeoForce: A Unified Force Tactile Representation Model

NeoForce learns a temporally structured visuo-tactile representation on top of the force fields, which serves as a structured tactile representation that downstream embodied manipulation models can consume. As shown in Fig. 7, the model receives synchronized chunks of RGB observations and tactile force fields. Each modality is embedded into patch tokens, the tokens are fused by a shared transformer backbone, and temporal context is aggregated across the input chunk. This design allows the representation to encode not only the current contact state but also the recent evolution of contact. The model is supervised jointly by a reconstruction objective in the force space and a latent-space objective, so that it learns to recover physically meaningful force fields while also capturing temporally structured tactile features.

The reconstruction head predicts the force field and contact mask from the fused tokens. Let $\hat{F}$ and $\hat{M}$ denote the predicted force field and contact mask, and let F and M denote the corresponding targets. The reconstruction loss is

$$\mathcal{L}_{\text{recons}} = \lambda_f \left\| (\hat{F} - F) \odot M \right\|_{\text{H}} + \lambda_g \left\| \bar{\hat{F}} - \bar{F} \right\|_{\text{H}} + \lambda_m \mathcal{L}_{\text{mask}}(\hat{M}, M), \tag{8}$$

where $\|\cdot\|_{\text{H}}$ is the Huber norm, $\bar{\cdot}$ denotes the spatial mean, and $\mathcal{L}_{\text{mask}}$ supervises contact segmentation.

The latent prediction head supervises the model in a shared latent space to capture the temporal correlations of tactile signals across the input chunk, using a teacher model to provide the ground truth in that latent space. The teacher receives the complete input and produces latent targets, while the student receives masked inputs and predicts those targets. Let Z_τ denote the teacher latent output and $\hat{Z}_s$ denote the student prediction. The latent prediction loss is

$$\mathcal{L}_{\text{pred}} = H\left(Z_\tau^v, \hat{Z}_s^v\right) + H\left(Z_\tau^t, \hat{Z}_s^t\right) + \lambda_x H\left(Z_\tau^{v\leftrightarrow t}, \hat{Z}_s^{v\leftrightarrow t}\right) + \lambda_p H\left(Z_{\tau,\text{patch}}, \hat{Z}_{s,\text{patch}}\right), \tag{9}$$

where Z^v and Z^t denote the visual and tactile latents. The cross-modal latent $Z^{v\leftrightarrow t}$ is obtained by predicting the tactile latent from the visual context and the visual latent from the tactile context, so that this term aligns the two modalities by forcing each to be recoverable from the other. The patch term Z_{patch} supervises masked patch-level latents and encourages local contact prediction under masking.

The complete objective is

$$\mathcal{L} = \mathcal{L}_{\text{recons}} + \mathcal{L}_{\text{pred}} + \lambda_c \mathcal{L}_{\text{con}}, \tag{10}$$

where $\mathcal{L}_{\text{con}}$ is a contrastive alignment term between visual and tactile embeddings. We evaluate the reconstruction quality of NeoForce in Section 5.5, and study how its force representation should be consumed by downstream policies in Section 6.2.

5.5 Evaluations on Tactile Representation

We evaluate NeoForce to figure out two essential problems, namely the effectiveness of the force-based representation on encoding contact intensity over space and time, and whether the distribution of the force space it learns stays consistent once the manipulation task varies. We first ablate the latent prediction objective against a reconstruction-only baseline and score held-out contacts with mean absolute error (MAE) and root mean squared error (RMSE) on the reconstructed force, together with mean intersection over union (mIoU) on the predicted contact region, so that errors in force magnitude and errors in contact localization are read separately. We then group the reconstructions by skill type and compare them with the ground truth force fields produced by the pipeline of Section 5.3.

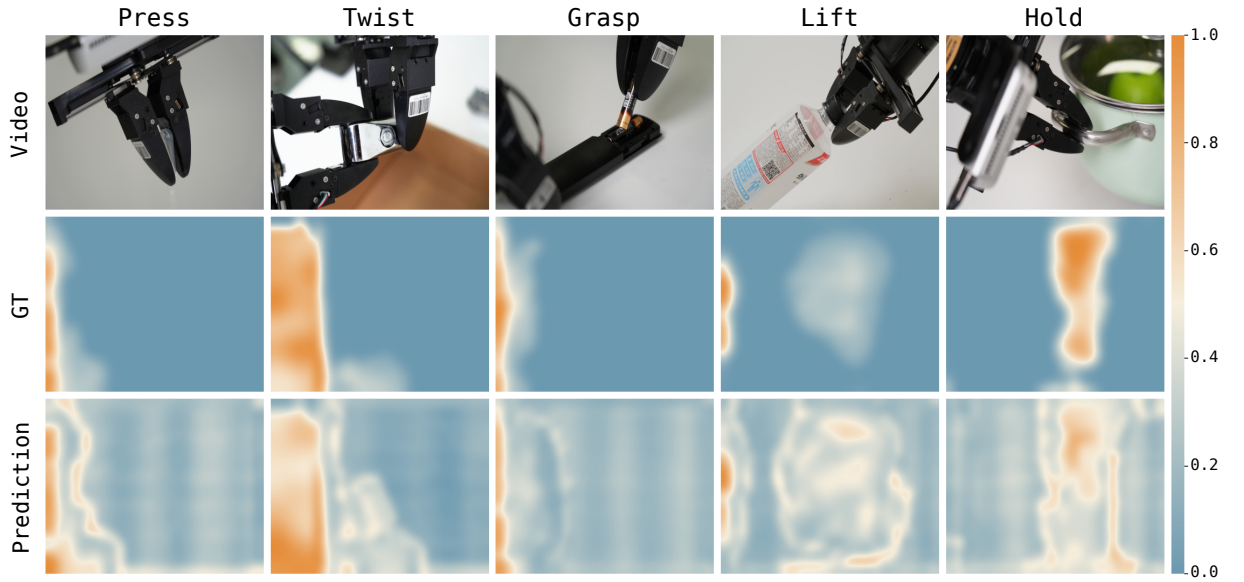


Figure 8 Force fields reconstruction of NeoForce on five representative contact behaviors: press, twist, grasp, lift, and hold. For each column, the top row shows the RGB frame, the middle row shows the ground truth force fields produced by the preprocessing pipeline of this section, and the bottom row shows the field reconstructed by NeoForce. The colorbar encodes normalized force magnitude.

Finding 1. Latent prediction captures the spatial correlation of tactile signals and thereby improves the representation space. As reported in Tab. 2, the reconstruction-only objective already reaches an MAE of 0.070 and an RMSE of 0.095 together with a contact region overlap of 0.968 mIoU. Adding latent prediction lowers the MAE to 0.066 and the RMSE to 0.089, while the overlap stays essentially unchanged at 0.966 mIoU. Based on this contrast between the two metric groups, the gain is concentrated in the magnitude of the reconstructed force rather than in the localization of the contact region. Compared with the pure reconstruction objective, predicting masked patch-level and cross-modal latents requires the model to infer the contact at an unobserved location. The surrounding contact area, the neighboring frames of the chunk, and the visual context covering the same region jointly constrain that missing location, which drives the representation to convey how contact intensity is distributed across space and time. The large-scale training on NeoData also contributes to the estimate, since it endows the shared backbone with a strong visual prior over object geometry and material. Cross-modal prediction carries this prior into the tactile tokens, so the tactile branch knows the contact region and estimates its intensity quantitatively more precisely. We therefore adopt the combined objective, whose downstream benefit we confirm on NeoReal in Section 6.2. *In short, latent prediction improves force estimation at no cost to contact localization, because a visual prior learned at scale helps the representation encode contact intensity quantitatively across space and time.*

Finding 2. Grounding on force yields a tactile representation that stays consistent across tasks. Fig. 8 visualizes the effect of the representation across skills. Across press, wrist, grasp, lift, and hold, which differ substantially in contact geometry and duration, NeoForce focuses on the region where the interaction takes place and extracts the properties that characterize it, including the intensity of the applied force, the geometry of the contact patch, and its spatial extent. In detail, press, wrist, and grasp yield localized peaks with sharp boundaries and near-zero surroundings, while lift and hold yield broader sustained distributions, and the reconstructed field consistently tracks the fine-grained intensity of the ground truth over a wide dynamic range. This consistency follows from the choice of force as the representation space. Such representation places the contact effects of the individual skills, together with their spatial distributions, into the same space, which is defined by physics rather than by the device or the behavior that produced the signal. *In short, anchoring the representation to force is what produces consistency across heterogeneous skills, rather than skill-specific fitting.*

Figure 9 The NeoReal suite. NeoReal comprises 10 real-world contact-rich tasks executed on robots equipped with tactile fingers: Cardboard Box Folding, Bag Packing, Cable Winding, Board Wiping, and Cup Stacking (top row); Socket Plugging, Board Insertion, Bottle Standing, Fruit Collection, and Towel Folding (bottom row).

Taken together, these results support the design of Section 5.1. The force-based representation encodes the universal force attribute of touch and generalizes over a wide range of embodied manipulation behaviors, which makes it a stable carrier of tactile interaction. We expect this representation to generalize further, so that a tactile signal carries hardware-agnostic information that remains consistent across sensors and that downstream manipulation policies can consume directly.

6 Evaluations on Tactile-aware Manipulation Tasks

We evaluate tactile-aware manipulation on contact-rich tasks where success depends on force regulation, fine-grained contact state, deformable-object handling, delicate insertion, and tactile feedback as an efficient control signal. The evaluation includes NeoReal for physical robot evaluation and NeoSim for reproducible simulation evaluation. For both NeoReal and NeoSim, we first describe the task set and protocol, and then report the evaluation of representative policy families.

6.1 NeoReal

NeoReal is designed to measure policy performance under fine-grained physical interaction, evaluating how policies behave under real-world tactile interaction and feedback. The suite contains 10 tasks selected from frequent and tactile-relevant skills in NeoData. These tasks cover deformable-object shaping, force-guided mating, delicate grasping, sustained surface contact, cable routing, stacking, insertion, and bimanual folding. As shown in Fig. 9, the task set includes Cardboard Box Folding, Bag Packing, Cable Winding, Board Wiping, Cup Stacking, Socket Plugging, Board Insertion, Bottle Standing, Fruit Collection, and Towel Folding.

Each task specifies a standardized initial state distribution, reset protocol, and binary success criterion. The tasks are executed on robots equipped with the tactile fingers described in Section 3.2. NeoReal therefore serves as the primary physical evaluation suite for determining whether tactile feedback improves manipulation performance in deployment. Per-task descriptions and success criteria are provided in Appendix E.

6.2 Evaluation on NeoReal

Setup. We evaluate representative policy families on NeoReal to validate their ability on contact-rich tactile manipulation tasks. The baselines include the imitation learning policies ACT [90] and Diffusion Policy [9]. For VLA, we evaluate $\pi_{0.5}$ [54], Xiaomi-Robotics-0 [5], InternVLA-A1 [4], and StarVLA-α [79]. We further include the world action models Fast-WAM [83] and GigaWorld Policy [78], together with the VA baseline LingBot-VA [33]. For NeoReal, policies are pretrained on NeoData and then post-trained on each task, and

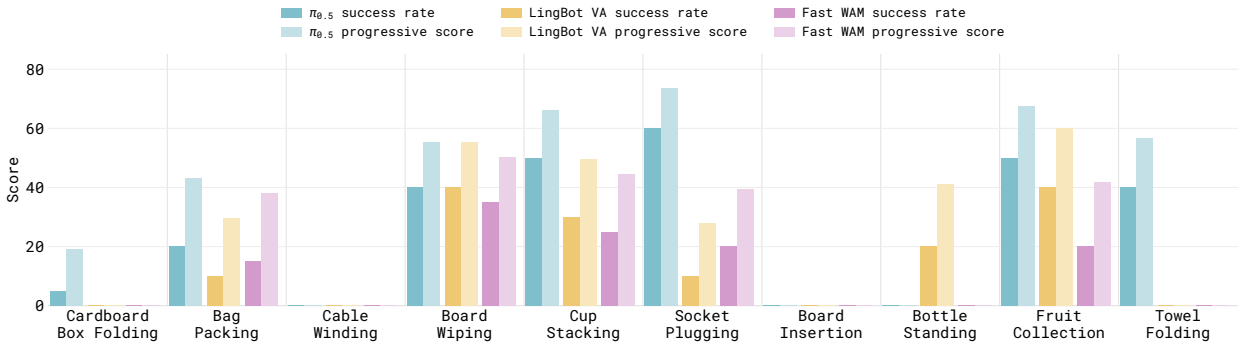


Figure 10 Task-level results on the NeoReal benchmark for the three strongest policies, $\pi_{0.5}$, LingBot-VA, and Fast-WAM. For each of the ten tasks, the six bars show, per policy, the binary success rate over 20 trials (solid) and the task progressive score (light) side by side. The progressive score credits partial completion of the per-task milestones defined in Appendix E.2 and averages over the same 20 trials, so it is never lower than the success rate.

Model (based on $\pi_{0.5}$)	Box	Bag	Cable	Wipe	Cups	Socket	Board	Bottle	Fruit	Towel	Avg
	success rate / progressive score										
No tactile	5/19.0	20/43.0	0/0.0	40/55.2	50/66.0	60/73.5	0/0.0	0/0.0	50/67.5	40/56.8	26.5/38.1
Tactile image concatenation	5/17.0	15/38.0	0/0.0	45/58.2	45/61.0	60/71.0	**5/19.0**	5/20.8	50/69.0	45/59.5	27.5/41.4
Tactile image action expert conditioning	5/19.0	**25/48.0**	0/0.0	45/60.7	**55/70.0**	55/70.0	**5/19.0**	5/20.8	**55/71.0**	50/64.5	30.0/44.3
NeoForce representation action expert conditioning	**15/34.5**	**25/48.0**	0/0.0	**50/63.2**	**55/70.0**	**65/76.0**	**5/19.0**	**10/29.8**	50/67.5	**50/66.5**	**32.5/47.5**

Table 3 Effect of tactile integration strategy on NeoReal, using $\pi_{0.5}$ as a fixed backbone. Each cell reports *success rate / progressive score* as in Fig. 10, and the No tactile row reproduces the $\pi_{0.5}$ result there. Column abbreviations denote Cardboard Box Folding (Box), Bag Packing (Bag), Cable Winding (Cable), Board Wiping (Wipe), Cup Stacking (Cups), Socket Plugging (Socket), Board Insertion (Board), Bottle Standing (Bottle), Fruit Collection (Fruit), and Towel Folding (Towel).

unless otherwise stated NeoReal reports success rate over 20 randomized rollouts per task. Full training and evaluation settings are provided in Appendices E and F.

Fig. 10 reports both the binary success rate and the progressive score of the three strongest policies, $\pi_{0.5}$, LingBot-VA, and Fast-WAM, on the 10 NeoReal tasks, where the progressive score credits partial completion of the per-task milestones defined in Appendix E.2. Under real contact dynamics, the tasks are markedly more challenging than in simulation, and $\pi_{0.5}$ reaches the highest average success rate (26.5%) by handling the sustained contact and force-guided tasks such as Board Wiping, Cup Stacking, and Socket Plugging, while LingBot-VA and Fast-WAM trail and solve only a subset of them. The progressive score consistently exceeds the success rate for every policy, and the gap widens on the contact-rich tasks, where on Socket Plugging $\pi_{0.5}$ moves from a 60% success rate to a 73.5% progressive score, and on Bag Packing from 20% to 43.0%, which indicates that the policies frequently reach intermediate milestones such as grasping and alignment even when they fail to complete the task. The progressive score also preserves the overall ranking, with $\pi_{0.5}$ still leading at an average of 38.1%. Under both metrics, every policy retains substantial room for improvement on these tactile tasks, and the shortfall is most pronounced where touch has to act as staged feedback, such as verifying a secure grasp before lifting or a seated fit before releasing. We believe that policies driven by visual observations and proprioceptive state alone remain limited on fine and precise manipulation, since neither modality exposes the contact state to perceive, nor the signal needed to correct an action while it is still in progress.

Furthermore, we use NeoReal to analyze how tactile information should be integrated with the policy. Using $\pi_{0.5}$ as a fixed backbone, we compare four variants, namely no tactile input, raw tactile image concatenation, raw tactile image conditioning for the action expert, and NeoForce force representation conditioning for the action expert. As reported in Tab. 3, every tactile variant outperforms the vision-only policy, confirming that contact-rich tasks benefit from information about force and contact state. Conditioning the action expert on tactile images is more effective than simple concatenation, because the contact signal is supplied directly to the module that generates actions. The NeoForce representation improves further over raw tactile images, since it expresses contact as a compact force field rather than as device appearance. The average progressive score

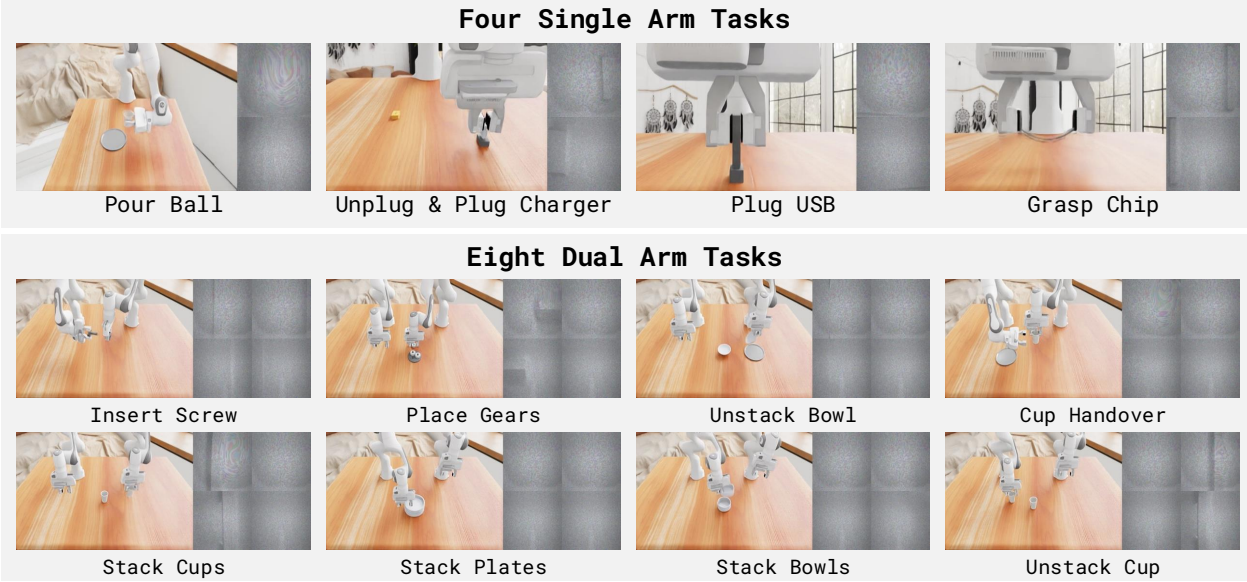


Figure 11 Demonstrations from NeoSim. NeoSim contains 12 tasks, including 4 single-arm tasks and 8 dual-arm tasks. For each task, the left panel shows the simulated RGB scene and the right panel visualizes the particle gel images rendered from the tactile sensors.

follows the same ordering, rising from 38.1% for the vision-only baseline to 47.5% for NeoForce conditioning, and although individual tasks fluctuate, with the raw tactile variants for instance dropping on Bag Packing and Cup Stacking before the force representation recovers them, the aggregate trend confirms that the gains also appear as more complete milestone progress and not only as additional binary successes. *In short, a policy benefits from the physical contact representation rather than from the device-specific appearance, and a sensor-agnostic force field provides a universal interface that delivers such a representation.*

6.3 NeoSim

NeoSim is constructed in the simulation environment and provides a reproducible arena for tactile-aware evaluation at scale. Following UniVTAC [6], NeoSim performs high-precision tactile data collection in simulation and renders per-contact force fields in the same x, y, and z format as the unified representation in Section 5. This design allows policies that consume force-based tactile inputs to be evaluated in simulation with the same representation used for real tactile data.

The suite contains 12 contact-rich tasks, including 4 single-arm tasks and 8 dual-arm tasks. The single-arm tasks are Pour Ball, Unplug and Plug Charger, Plug USB, and Grasp Chip, which stress force regulation during pouring, insertion, and delicate grasping. The dual-arm tasks are Insert Screw, Place Gears, Unstack Bowl, Cup Handover, Stack Cups, Stack Plates, Stack Bowls, and Unstack Cup, which further require bimanual coordination, stable contact maintenance, handover, and nesting. These tasks span rigid objects, thin-walled containers, fragile parts, deformable contact, and tight-clearance assembly.

NeoSim is intended to complement NeoReal by providing controlled, reproducible, and large-scale evaluation that is difficult to obtain on physical hardware. It enables controlled comparisons across policy families, tactile input choices, and representation choices under large-scale contact dynamics. A detailed per-task description of NeoSim, together with rendered examples of RGB observations and tactile force fields, is provided in Appendix F.

6.4 Evaluation on NeoSim

We evaluate on NeoSim how current policy families handle contact-rich manipulation in a reproducible simulated setting. Using the same policy families and NeoData pretraining as above, NeoSim reports success rate over 100 randomized rollouts per task. Tab. 4 reports the NeoSim results. $\pi_{0.5}$ leads with a mean success

Policy	Single arm				Dual arm								Mean
	USB	Chip	Charger	Pour	Bowl U	Cup S	Screw	Gears	Cup H	Plate S	Bowl S	Cup U	
$\pi_{0.5}$ [54]	**74**	86	**23**	**92**	3	**22**	**26**	**18**	13	**96**	**93**	3	**45.8**
Xiaomi-Robotics-0 [5]	62	49	5	52	0	0	0	0	0	0	42	**71**	23.4
InternVLA-A1 [4]	36	12	0	47	0	0	0	0	0	0	8	0	8.6
StarVLA-α [79]	72	**93**	0	32	17	0	8	0	**25**	0	0	31	23.2
Fast-WAM [83]	0	0	0	0	0	0	0	0	0	0	0	0	0.0
GigaWorld Policy [78]	34	71	3	0	0	12	0	0	0	0	0	9	10.8
LingBot-VA [33]	31	77	0	37	**18**	20	18	0	0	93	67	24	32.1

Table 4 Task level success rates over 100 randomized rollouts on NeoSim. Column abbreviations denote Insert USB (USB), Grasp Chip (Chip), Unplug and Plug Charger (Charger), Pour Ball (Pour), Bowl Unstack (Bowl U), Cup Stack (Cup S), Insert Screw (Screw), Place Gears (Gears), Cup Handover (Cup H), Plate Stack (Plate S), Bowl Stack (Bowl S), and Cup Unstack (Cup U). Mean averages over the twelve task success rates in each row.

rate of 45.8%, followed by LingBot-VA at 32.1%, while Xiaomi-Robotics-0 and StarVLA-α form a middle group near 23% and the remaining baselines stay below 11%. The dual-arm tasks are consistently more difficult than the single-arm tasks. Averaged over the four single-arm tasks, every policy scores higher than it does over the eight dual-arm tasks, and the drop is severe for most of them, from 49.3% to 10.1% for StarVLA-α, from 42.0% to 14.1% for Xiaomi-Robotics-0, and from 23.8% to 1.0% for InternVLA-A1. Only $\pi_{0.5}$ and LingBot-VA retain substantial bimanual competence, at 34.3% and 30.0%, respectively, and they do so almost entirely on Plate Stack and Bowl Stack, where the two arms can act in a loosely coupled manner. The tasks that demand sustained mutual contact remain nearly unsolved for every policy, with the best score reaching only 18% on Place Gears and 25% on Cup Handover, and on each of these two tasks at most two of the seven policies succeed at all. Fast-WAM fails on all twelve tasks despite being one of the three strongest policies on NeoReal, which indicates that its competence does not transfer to the simulated contact dynamics. These results expose contact-rich bimanual manipulation as the main bottleneck for current policy families that rely primarily on visual observations, which reveal little about whether a grasp is secure and whether contact is stably maintained. *In short, tactile signals make grasp security and contact stability directly observable to the policy, which is exactly the information that contact-rich bimanual manipulation currently fails on.* This in turn strengthens the necessity of equipping robots with tactile hardware and of policies that consume a tactile representation.

7 Conclusion

We presented $\mathcal{N}_0$-Foundation, an integrated resource for tactile-enabled embodied manipulation that spans hardware, data, representation, and evaluation. The report contributes NeoData, a large-scale multimodal dataset with more than 30,000 hours of synchronized visual and tactile demonstrations across six embodiments and 450+ tasks, collected through a mixture of real-robot teleoperation and UMI demonstrations. It also contributes NeoForce, which convert sensor signals into tactile representation for downstream embodied models. Finally, we propose a standardized evaluation that combines the real-world NeoReal suite and the simulated NeoSim suite for standardized evaluation of contact-rich manipulation. The experiments indicate that tactile feedback improves contact-rich manipulation, that latent supervision improves temporal tactile representation learning, and that the NeoForce force representation provides a compact physical alternative to raw tactile images.

Several directions remain open. The unified representation currently targets parallel-jaw visuo-tactile fingers, and extending it to dexterous hands and non-camera-based transducers would broaden its applicability. Scaling policy learning to the full multimodal corpus and tightening the connection between NeoSim and real deployment are natural next steps. NeoData, together with its representation and benchmark, provides a foundation for more capable and transferable tactile-aware manipulation policies.

Contributors

Data Infrastructure. Xiufeng Song, Rui Li, Wenjie Zhou, Yutao Fan, Heng Zhou, Tianyu Yang, Longjie Su, Li Kang, Zhemeng Zhang, Kun Huang, Haotian Wang, Shitao Lin, Shunlin Lu, Yiran Qin.
Data Preprocess and Annotation. Rui Li, Xiufeng Song, Tianyu Yang, Silong Dai, Longjie Su.
Evaluation. Bruno N.Y. Chen, Jiongwei Lu, Yifan Wang, Shengqi Xu, Yutao Fan, Zipei Ma, Xiaofei Wei, Zhemeng Zhang, Heng Zhou, Silong Dai, Li Kang, Longjie Su, Boyu Mi, Yanjun Li, Chenxi Wu, Nan Min, Guojin Zhong, Haoyu Zhao, Xiufeng Song, Rui Li, Xin Wang, Yiran Qin.
Hardware. Hu Luo, Liuchuang Tang, Yu Han, Hua Zhu, Yutao Ling, Qiuhui Huang, Fanjie Wang, Lianjie Jiang, Hao Chen, Yicheng Yang, Yinxiao Lu.
Academic Supervision. Ziyi Ye, Guoxiang Dong, Xiaosong Jia, Wenming Chen.
Project Lead. Shunlin Lu, Yiran Qin, Shihao Zhao, Daoguo Dong, Zuxuan Wu, Yu-Gang Jiang.

References

[1] Iretiayo Akinola, Jie Xu, Jan Carius, Dieter Fox, and Yashraj Narang. Tacsl: A library for visuotactile sensor simulation and learning. *T-RO*, 2025.

[2] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. *CVPR*, 2023.

[3] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Karol Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

[4] Junhao Cai, Zetao Cai, Jiafei Cao, Yilun Chen, Zeyu He, Lei Jiang, Hang Li, Hengjie Li, Yang Li, Yufei Liu, et al. Internvla-a1: Unifying understanding, generation and action for robotic manipulation. *arXiv preprint arXiv:2601.02456*, 2026.

[5] Rui Cai, Jun Guo, Xinze He, Piaopiao Jin, Jie Li, Bingxuan Lin, Futeng Liu, Wei Liu, Fei Ma, Kun Ma, et al. Xiaomi-robotics-0: An open-sourced vision-language-action model with real-time execution. *arXiv preprint arXiv:2602.12684*, 2026.

[6] Baijun Chen, Weijie Wan, Tianxing Chen, Xianda Guo, Congsheng Xu, Yuanyang Qi, Haojie Zhang, Longyan Wu, Tianling Xu, Zixuan Li, Yizhe Wu, Rui Li, Xiaokang Yang, Ping Luo, Wei Sui, and Yao Mu. Univtac: A unified simulation platform for visuo-tactile manipulation data generation, learning, and benchmarking. *arXiv preprint arXiv:2602.10093*, 2026.

[7] Tianxing Chen, Yue Chen, Zixuan Li, Junyuan Tang, Kailun Su, Haoran Lu, Weijie Wan, Baijun Chen, Songling Liu, Haowen Yan, Honghao Su, Zhiyang Dou, Kaixuan Wang, Dandan Zhang, Yunze Liu, Yan Qin, Qiwei Liang, Qiwei Wu, Zijian Lin, Wenwei Lin, Yuran Wang, Minghua He, Tianshu Wu, Ruihai Wu, Jingquan Zhou, Kai-Chong Lei, Haibao Yu, Yuanfeng Ji, Weiyang Jin, Guanyu Lin, Xiaofan Li, Qi Xiong, Renjing Xu, Zhongyu Li, Wenhao Chai, Enze Xie, Ziwei Wang, Yao Mu, Hao Dong, Wojciech Matusik, Mingyu Ding, Wenbo Ding, Ping Luo, and Masayoshi Tomizuka. RoboDojo: a unified Sim-and-Real benchmark for comprehensive evaluation of generalist robot manipulation policies. *arXiv preprint arXiv:2607.04434*, 2026.

[8] Tailai Cheng, Kejia Chen, Lingyun Chen, Liding Zhang, Yue Zhang, Yao Ling, Mahdi Hamad, Zhenshan Bing, Fan Wu, Karan Sharma, et al. Tacumi: A multi-modal universal manipulation interface for contact-rich tasks. *arXiv preprint arXiv:2601.14550*, 2026.

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

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

[11] Daimon Robotics. Daimon-Infinity. https://modelscope.cn/datasets/daimonrobotics/Daimon-Infinity/, 2026.

[12] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The EPIC-KITCHENS dataset. *ECCV*, 2018.

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. *CVPR*, 2009.

[14] Hao-Shu Fang, Hongjie Fang, Zhenyu Tang, Jirong Liu, Chenxi Wang, Junbo Wang, Haoyi Zhu, and Cewu Lu. RH20T: A comprehensive robotic dataset for learning diverse skills in one-shot. *arXiv preprint arXiv:2307.00595*, 2023.

[15] Senyu Fei, Siyin Wang, Junhao Shi, Zihao Dai, Jikun Cai, Pengfang Qian, Li Ji, Xinzhe He, Shiduo Zhang, Zhaoye Fei, Jinlan Fu, Jingjing Gong, and Xipeng Qiu. LIBERO-Plus: in-depth robustness analysis of Vision-Language-Action models. *arXiv preprint arXiv:2510.13626*, 2025.

[16] Zelin Fu, Bin Tan, Changjiang Sun, Shaohui Liu, Kecheng Zheng, Yinghao Xu, Xing Zhu, Yujun Shen, and Nan Xue. Vision pretraining for dense spatial perception. *arXiv preprint arXiv:2607.05247*, 2026.

[17] Ning Gao, Jinliang Zheng, Xing Gao, Haoxiang Ma, Hanqing Wang, Yukai Wang, Jiantong Chen, Zanxin Chen, Shujie Zhang, Mingda Jia, Xuekun Jiang, Zihou Zhu, Xinyu Li, Shuai Wang, Hao Li, Wenzhe Cai, Yuqiang Yang, Xudong Xu, Zhaoyang Lyu, Yao Mu, Tai Wang, Jiangmiao Pang, Jia Zeng, Weinan Zhang, and Chunhua Shen. EBench: elemental diagnosis of generalist mobile manipulation policies. *arXiv preprint arXiv:2606.18239*, 2026.

[18] Zelin Gao, Qiuyu Wang, Jiapeng Zhu, Jingye Chen, Zichen Liu, Qingyan Bai, Jiahao Wang, Yufeng Yuan, Hanlin Wang, Yichong Lu, Ka Leong Cheng, Haojie Zhang, Jian Gao, Tianrui Feng, Yuzheng Liu, Yao Yao, Yinghao Xu, Xing Zhu, Yujun Shen, and Hao Ouyang. Infinite worlds with versatile interactions. *arXiv preprint arXiv:2607.07534*, 2026.

[19] Jiayuan Gu, Fanbo Xiang, Xuanlin Li, Zhan Ling, Xiqiang Liu, Tongzhou Mu, Yihe Tang, Stone Tao, Xinyue Wei, Yunchao Yao, Xiaodi Yuan, Pengwei Xie, Zhiao Huang, Rui Chen, and Hao Su. Maniskill2: A unified benchmark for generalizable manipulation skills. *ICLR*, 2023.

[20] Harsh Gupta, Yuchen Mo, Shengmiao Jin, and Wenzhen Yuan. Sensor-invariant tactile representation. *arXiv preprint arXiv:2502.19638*, 2025.

[21] Huy Ha, Yihuai Gao, Zipeng Fu, Jie Tan, and Shuran Song. Umi on legs: Making manipulation policies mobile with manipulation-centric whole-body controllers. *arXiv preprint arXiv:2407.10353*, 2024.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CVPR*, 2016.

[23] Minho Heo, Youngwoon Lee, Doohyun Lee, and Joseph J. Lim. Furniturebench: Reproducible real-world benchmark for long-horizon complex manipulation. *IJRR*, (10-11), 2025.

[24] Ryan Hoque, Peide Huang, David J. Yoon, Mouli Sivapurapu, and Jian Zhang. EgoDex: Learning dexterous manipulation from large-scale egocentric video. *ICLR*, 2026.

[25] Stephen James, Zicong Ma, David Rovick Arrojo, and Andrew J. Davison. Rlbench: The robot learning benchmark & learning environment. *RA-L*, (2), 2020.

[26] Bowen Jing, Mingxin Wang, Ruiyang Hao, Chenchen Ge, Hanwen Shen, Junjie He, Yang Cui, Yiming Hou, Weitao Zhou, Jiawei Wang, Minglei Li, Dandan Zhang, Ding Zhao, Houde Liu, Xiaofan Li, Si Liu, Ping Luo, and Haibao Yu. SoftVTBench: a Safety-Aware Visuo-Tactile benchmark for physically constrained robotic manipulation of deformable objects. *arXiv preprint arXiv:2607.04234*, 2026.

[27] Tobias Jülg, Seongjin Bien, Simon Hilber, Yannik Blei, Pierre Krack, Maximilian Li, Sven Parusel, Rudolf Lioutikov, Florian Walter, and Wolfram Burgard. DuoBench: a reproducible benchmark for bimanual manipulation in simulation and the real world. *arXiv preprint arXiv:2606.11901*, 2026.

[28] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024.

[29] Torsten Kröger, Bernd Finkemeyer, and Friedrich M Wahl. Manipulation primitives—a universal interface between sensor-based motion control and robot programming. *Robotic systems for handling and assembly*, 2011.

[30] Mike Lambeta, Po-Wei Chou, Stephen Tian, Brian Yang, Benjamin Maloon, Victoria Rose Most, Dave

Stroud, Raymond Santos, Ahmad Byagowi, Gregg Kammerer, et al. Digit: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation. *RA-L*, (3), 2020.

[31] Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Michael Lingelbach, Jiankai Sun, Mona Anvari, Minjune Hwang, Manasi Sharma, Arman Aydin, Dhruva Bansal, Samuel Hunter, Kyu-Young Kim, Alan Lou, Caleb R. Matthews, Ivan Villa-Renteria, Jerry Huayang Tang, Claire Tang, Fei Xia, Silvio Savarese, Hyowon Gweon, C. Karen Liu, Jiajun Wu, and Li Fei-Fei. BEHAVIOR-1K: A benchmark for embodied AI with 1,000 everyday activities and realistic simulation. *CoRL*, 2022.

[32] Chuanyu Li, Chaoyi Liu, Daotan Wang, Shuyu Zhang, Lusong Li, Zecui Zeng, Fangchen Liu, Jing Xu, and Rui Chen. Vitamin-b: A reliable and efficient visuo-tactile bimanual manipulation interface. *arXiv preprint arXiv:2511.05858*, 2025.

[33] Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Fei Han, Mingrui Yu, Zelin Gao, Nan Xue, Xing Zhu, Yujun Shen, and Yinghao Xu. Causal world modeling for robot control. *arXiv preprint arXiv:2601.21998*, 2026.

[34] Shuang Li, Yihuai Gao, Dorsa Sadigh, and Shuran Song. Unified video action model. *arXiv preprint arXiv:2503.00200*, 2025.

[35] Xuanlin Li, Kyle Hsu, Jiayuan Gu, Oier Mees, Karl Pertsch, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, Sergey Levine, Jiajun Wu, Chelsea Finn, Hao Su, Quan Vuong, and Ted Xiao. Evaluating real-world robot manipulation policies in simulation. *CoRL*, 2024.

[36] Xinghang Li, Jun Guo, Qiwei Li, Long Qian, Hang Lai, Yueze Wang, Hongyu Yan, Jiahang Cao, Xi Chen, Jingen Qu, Jiaxi Song, Nan Sun, Hanye Zhao, Futeng Liu, Wanli Peng, Heyun Wang, Yunhong Wang, Caoyu Xia, Jack Zhao, Diyun Xiang, Hangjun Ye, Heng Qu, Huaping Liu, and Jason Li. Xiaomi-robotics-u0: Unified embodied synthesis with world foundation model. *arXiv preprint arXiv:2607.11643*, 2026.

[37] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: benchmarking knowledge transfer for lifelong robot learning. *NeurIPS*, 2023.

[38] Kehui Liu, Chuyue Guan, Zhongjie Jia, Ziniu Wu, Xin Liu, Tianyu Wang, Shuai Liang, Pengan Chen, Pingrui Zhang, Haoming Song, et al. Fastumi: A scalable and hardware-independent universal manipulation interface with dataset. *arXiv preprint arXiv:2409.19499*, 2024.

[39] Yunfan Lou, Yifan Ye, Yankai Fu, Jun Cen, Xiaowei Chi, Yaoxu Lyu, Peidong Jia, Sirui Han, Zhihe Lu, and Shanghang Zhang. Dream-tac: A unified tactile world action model for contact-rich robot manipulation. *arXiv preprint arXiv:2606.08737*, 2026.

[40] Hu Luo, Xingyu Zhang, Yangyang Xu, Jian Yu, Xin Ma, and Wen-Ming Chen. A novel flexible multidimensional force platform for assessing regional plantar pressure and shear stress under the foot during gait. *IEEE Sensors Journal*, 24(7):11723–11733, 2024.

[41] Hu Luo, Yuxin Ma, Zesheng Wang, Jiewen Li, Xin Ma, and Wen-Ming Chen. Efficacy of varying sensing features for enhanced performance of deep-learning-informed multidimensional force platform. *Pattern Recognition*, page 112327, 2025.

[42] Hu Luo, Yangyang Xu, Qiuhui Huang, Xiaohui Wang, Zesheng Wang, Jian Yu, Yunchao Zhu, Xin Ma, and Wen-Ming Chen. A quantitative benchmark of deep learning-based decoupling methods for multidimensional force platform used in dynamic gait analysis. *T-IM*, 74:1–16, 2025.

[43] Shuailei Ma, Jiaqi Liao, Xinyang Wang, Jingjing Wang, Chaoran Feng, Zijing Hu, Chong Bao, Zichen Xi, Yuqi Gan, Weisen Wang, Yanhong Zeng, Qin Zhao, Zifan Shi, Wei Wu, Hao Ouyang, Qiuyu Wang, Shangzhan Zhang, Jiahao Shao, Yipengjing Sun, Liangxiao Hu, Lunke Pan, Nan Xue, Kecheng Zheng, Yinghao Xu, Xing Zhu, Yujun Shen, and Ka Leong Cheng. Scaling mixture-of-experts video pretraining for embodied intelligence. *arXiv preprint arXiv:2607.07675*, 2026.

[44] Perla Maiolino, Marco Maggiali, Giorgio Cannata, Giorgio Metta, and Lorenzo Natale. A flexible and robust large scale capacitive tactile system for robots. *IEEE Sensors Journal*, (10), 2013.

[45] Ajay Mandlekar, Danfei Xu, Josiah Wong, Soroush Nasiriany, Chen Wang, Rohun Kulkarni, Li Fei-Fei, Silvio Savarese, Yuke Zhu, and Roberto Martín-Martín. What matters in learning from offline human demonstrations for robot manipulation. *CoRL*, 2021.

[46] Oier Mees, Lukás Hermann, Erick Rosete-Beas, and Wolfram Burgard. CALVIN: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *RA-L*, (3), 2022.

[47] Mayank Mittal, Calvin Yu, Qinxi Yu, Jingzhou Liu, Nikita Rudin, David Hoeller, Jia Lin Yuan, Ritvik Singh, Yunrong Guo, Hammad Mazhar, Ajay Mandlekar, Buck Babich, Gavriel State, Marco Hutter, and Animesh Garg. Orbit: A unified simulation framework for interactive robot learning environments. *RA-L*, (6), 2023.

[48] Soroush Nasiriany, Abhiram Maddukuri, Lance Zhang, Adeet Parikh, Aaron Lo, Abhishek Joshi, Ajay Mandlekar, and Yuke Zhu. Robocasa: Large-scale simulation of household tasks for generalist robots. *RSS*, 2024.

[49] Soroush Nasiriany, Sepehr Nasiriany, Abhiram Maddukuri, and Yuke Zhu. RoboCasa365: a Large-Scale simulation framework for training and benchmarking generalist robots. *arXiv preprint arXiv:2603.04356*, 2026.

[50] 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. *arXiv preprint arXiv:2411.04776*, 2024.

[51] Dantong Niu, Zhuoyang Liu, Zekai Wang, Boning Shao, Zhao-Heng Yin, Anirudh Pai, Yuvan Sharma, Stefano Saravalle, Ruijie Zheng, Jing Wang, et al. T-rex: Tactile-reactive dexterous manipulation. *arXiv preprint arXiv:2606.17055*, 2026.

[52] Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, et al. Open x-embodiment: Robotic learning datasets and rt-x models. *ICRA*, 2024.

[53] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[54] Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. $\pi_{0.5}$: A vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

[55] Ryan Punamiya, Simar Kareer, Zeyi Liu, Josh Citron, Ri-Zhao Qiu, Xiongyi Cai, Alexey Gavryushin, Jiaqi Chen, Davide Liconti, Lawrence Y. Zhu, Patcharapong Aphiwetsa, Baoyu Li, Aniketh Cheluva, Pranav Kuppili, Yangcen Liu, Dhruv Patel, Aidan Gao, Hye-Young Chung, Ryan Co, Renee Zbizika, Jeff Liu, Xiaomeng Xu, Haoyu Xiong, Geng Chen, Sebastiano Oliani, Wenkai Xuan, Chenyu Yang, Xi Wang, James Fort, Richard Newcombe, Josh Gao, Jason Chong, Garrett Matsuda, Aseem Doriwala, Marc Pollefeys, Robert Katzschmann, Xiaolong Wang, Shuran Song, Judy Hoffman, and Danfei Xu. EgoVerse: An egocentric human dataset for robot learning from around the world. *arXiv preprint arXiv:2604.07607*, 2026.

[56] Alexander Schmitz, Perla Maiolino, Marco Maggiali, Lorenzo Natale, Giorgio Cannata, and Giorgio Metta. Methods and technologies for the implementation of large-scale robot tactile sensors. *T-RO*, (3), 2011.

[57] Modi Shi, Yuxiang Lu, Huijie Wang, Chengen Xie, and Qingwen Bu. Introducing AgiBot World Colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. https://opendrivelab.com/AgiBot-World/, 2025.

[58] Mike Zheng Shou, Stan Weixian Lei, Weiyao Wang, Deepti Ghadiyaram, and Matt Feiszli. Generic event boundary detection: A benchmark for event segmentation. *ICCV*, 2021.

[59] Zilin Si and Wenzhen Yuan. Taxim: An example-based simulation model for gelsight tactile sensors. *RA-L*, (2), 2022.

[60] Stefano Stassi, Valentina Cauda, Giancarlo Canavese, and Candido Fabrizio Pirri. Flexible tactile sensing based on piezoresistive composites: A review. *Sensors*, (3), 2014.

[61] 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. *arXiv preprint arXiv:2310.17274*, 2023.

[62] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *CVPR*, 2015.

[63] TacVerse. TacVerse datasets. https://huggingface.co/TacVerse/datasets, 2026.

[64] Team HY-World, Chenjie Cao, Xuhui Zuo, Zhenwei Wang, Yisu Zhang, Junta Wu, Zhenyang Liu, Yuning Gong, Yang Liu, Bo Yuan, Chao Zhang, Coopers Li, Dongyuan Guo, Fan Yang, Haiyu Zhang, Hang Cao, Jianchen Zhu, Jiaxin Lin, Jie Xiao, Jihong Zhang, Junlin Yu, Lei Wang, Lifu Wang, Lilin Wang, Linus, Minghui Chen, Peng He, Penghao Zhao, Qi Chen, Rui Chen, Rui Shao, Sicong Liu, Wangchen Qin, Xiaochuan Niu, Xiang Yuan, Yi Sun, Yifei Tang, Yifu Sun, Yihang Lian, Yonghao Tan, Yuhong Liu, Yuyang Yin, Zhiyuan Min, Tengfei Wang, and Chunchao Guo. Hy-world 2.0: A multi-modal world model for reconstructing, generating, and simulating 3d worlds. *arXiv preprint arXiv:2604.14268*, 2026.

[65] Tencent Robotics X, HY Vision Team, Xumin Yu, Zuyan Liu, Ziyi Wang, He Zhang, Yongming Rao, Fangfu Liu, Yani Zhang, Ruowen Zhao, Oran Wang, Yves Liang, Haitao Lin, Minghui Wang, Yubo Dong, Kevin Cheng, Bolin Ni, Rui Huang, Han Hu, Zhengyou Zhang, Linus, and Shunyu Yao. Hy-embodied-0.5: Embodied foundation models for real-world agents. *arXiv preprint arXiv:2604.07430*, 2026.

[66] Homer Rich Walke, Kevin Black, Tony Z Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, et al. Bridgedata v2: A dataset for robot learning at scale. *CoRL*, 2023.

[67] Ziyi Wang, Xumin Yu, Yongming Rao, Yonggen Ling, Yunheng Li, Oran Wang, Mingqi Gao, Yuchen Zhou, Yves Liang, Zuyan Liu, Yani Zhang, Rui Huang, Xiaoran Xu, Bowen Yuan, Yifu Yuan, Xu Tan, He Zhang, Yufei Huang, Shenghao Zhang, Hongsheng Wu, Han Hu, and Zhengyou Zhang. Hy-embodied-vlm-1.0: Efficient physical-world agents. *arXiv preprint arXiv:2607.12894*, 2026.

[68] Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. Unleashing large-scale video generative pre-training for visual robot manipulation. *ICLR*, 2024.

[69] Longyan Wu, Checheng Yu, Jieji Ren, Li Chen, Yufei Jiang, Ran Huang, Guoying Gu, and Hongyang Li. Freetacman: Robot-free visuo-tactile data collection system for contact-rich manipulation. *arXiv preprint arXiv:2506.01941*, 2025.

[70] Longyan Wu, Jieji Ren, Chenghang Jiang, Junxi Zhou, Shijia Peng, Ran Huang, Guoying Gu, Li Chen, and Hongyang Li. TAMEn: Tactile-aware manipulation engine for closed-loop data collection in contact-rich tasks. *arXiv preprint arXiv:2604.07335*, 2026.

[71] Tianhao Wu, Jinzhou Li, Jiyao Zhang, Mingdong Wu, and Hao Dong. Canonical representation and force-based pretraining of 3d tactile for dexterous visuo-tactile policy learning. *ICRA*, 2025.

[72] Wei Wu, Fangjing Wang, Fan Lu, He Sun, Shi Liu, Yunnan Wang, Yibin Yan, Yong Wang, Shuailei Ma, Xinyang Wang, Yibin Liu, Shuai Yang, Tianxiang Zhou, Kejia Zhang, Lei Zhou, Cheng Su, Nan Xue, Bin Tan, Han Zhang, Youchao Zhang, Fei Liao, Xing Zhu, Yujun Shen, and Kecheng Zheng. From foundation to application: Improving vla models in practice. *arXiv preprint arXiv:2607.06403*, 2026.

[73] Xiaomi Robotics Team, Jun Guo, Piaopiao Jin, Jason Li, Peiyan Li, Yingyan Li, Futeng Liu, Wanli Peng, Optimus Qin, Yifei Su, Nan Sun, Qiao Sun, Runze Suo, Heyun Wang, Yunhong Wang, Rujie Wu, Caoyu Xia, Lina Zhang, Jack Zhao, Guoliang Chen, Wenlong Chen, Xinze He, Bin Li, Qing Li, Zhuorong Li, Heng Qu, Wenxuan Song, Diyun Xiang, Yifan Xie, Peiran Xu, Hangjun Ye, Wen Ye, Han Zhao, and Quanyun Zhou. Xiaomi-robotics-1: Scaling vision-language-action models with over 100k hours of real-world trajectories. *arXiv preprint arXiv:2607.15330*, 2026.

[74] Chengyi Xing, Hao Li, Yi-Lin Wei, Tian-Ao Ren, Tianyu Tu, Yuhao Lin, Elizabeth Schumann, Wei-Shi Zheng, and Mark R Cutkosky. Taccap: A wearable fbg-based tactile sensor for seamless human-to-robot skill transfer. *arXiv preprint arXiv:2503.01789*, 2025.

[75] Mengda Xu, Han Zhang, Yifan Hou, Zhenjia Xu, Linxi Fan, Manuela Veloso, and Shuran Song. Dexumi: Using human hand as the universal manipulation interface for dexterous manipulation. *arXiv preprint arXiv:2505.21864*, 2025.

[76] Shengqi Xu, Guojin Zhong, Yang Liu, Fanjie Wang, Hu Luo, Hanyu Zhou, Weiyao Zhang, Ziyi Ye, Zuxuan Wu, and Yu-Gang Jiang. Seeing touch from motion: A unified modality-aware visuo-tactile policy with tactile motion correlation. *arXiv preprint arXiv:2606.29941*, 2026.

[77] Yue Xu, Litao Wei, Pengyu An, Qingyu Zhang, and Yong-Lu Li. exumi: Extensible robot teaching system with action-aware task-agnostic tactile representation. *arXiv preprint arXiv:2509.14688*, 2025.

[78] Angen Ye, Boyuan Wang, Chaojun Ni, Guan Huang, Guosheng Zhao, Hao Li, Hengtao Li, Jie Li, Jindi Lv, Jingyu Liu, et al. Gigaworld-policy: An efficient action-centered world–action model. *arXiv preprint arXiv:2603.17240*, 2026.

[79] Jinhui Ye, Ning Gao, Senqiao Yang, Jinliang Zheng, Zixuan Wang, Yuxin Chen, Pengguang Chen, Yilun Chen, Shu Liu, and Jiaya Jia. Starvla-α: Reducing complexity in vision-language-action systems. *arXiv preprint arXiv:2604.11757*, 2026.

[80] Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, Ayaan Malik, Kyungmin Lee, William Liang, Nadun Ranawaka, Jiasheng Gu, Yinzhen Xu, Guanzhi Wang, Fengyuan Hu, Avnish Narayan, Johan Bjorck, Jing Wang, Gwanghyun Kim, Dantong Niu, Ruijie Zheng, Yuqi Xie, Jimmy Wu, Qi Wang, Ryan Julian, Danfei Xu, Yilun Du, Yevgen Chebotar, Scott Reed, Jan Kautz, Yuke Zhu, Linxi Jim Fan, and Joel Jang. World action models are zero-shot policies. *arXiv preprint arXiv:2602.15922*, 2026.

[81] 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. *CoRL*, 2019.

[82] Chengbo Yuan, Zicheng Zhang, Mingjie Zhou, Wendi Chen, Yi Wang, Zhuoyang Liu, Dantong Niu, Shuo Wang, Hui Zhang, Wenkang Zhang, et al. Ftp-1: A generalist foundation tactile policy across tactile sensors for contact-rich manipulation. *arXiv preprint arXiv:2606.13102*, 2026.

[83] Tianyuan Yuan, Zibin Dong, Yicheng Liu, and Hang Zhao. Fast-wam: Do world action models need test-time future imagination? *arXiv preprint arXiv:2603.16666*, 2026.

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

[85] Qiyuan Zeng, Chengmeng Li, Jude St John, Zhongyi Zhou, Junjie Wen, Guorui Feng, Yichen Zhu, and Yi Xu. Activeumi: Robotic manipulation with active perception from robot-free human demonstrations. *arXiv preprint arXiv:2510.01607*, 2025.

[86] Qihang Zhang, Lin Li, Luyao Zhang, Shuai Yang, Yiming Luo, Shuaiting Li, Ruilin Wang, Junke Wang, Jiahao Shao, Gangwei Xu, Jiaming Zhou, Yishu Shen, Yudong Jin, Fangyi Xu, Shuailei Ma, Jiaqi Liao, Guanxing Lu, Zifan Shi, Yongkun Wen, Yujie Zhao, Weixuan Tang, Xinyang Wang, Chaojian Li, Jiapeng Zhu, Ka Leong Cheng, Nan Xue, Xing Zhu, Yujun Shen, and Yinghao Xu. Native video-action pretraining for generalizable robot control. *Technical Report*, 2026.

[87] Shiduo Zhang, Zhe Xu, Peiju Liu, Xiaopeng Yu, Yuan Li, Qinghui Gao, Zhaoye Fei, Zhangyue Yin, Zuxuan Wu, Yu-Gang Jiang, et al. Vlabench: A large-scale benchmark for language-conditioned robotics manipulation with long-horizon reasoning tasks. *ICCV*, 2025.

[88] Zhemeng Zhang, Jiahua Ma, Xincheng Yang, Xin Wen, Yuzhi Zhang, Boyan Li, Yiran Qin, Jin Liu, Can Zhao, Li Kang, Haoqin Hong, Zhenfei Yin, Philip Torr, Hao Su, Ruimao Zhang, and Daolin Ma. Touchguide: Inference-time steering of visuomotor policies via touch guidance. *RSS*, 2026.

[89] Zhiyuan Zhang, Pokuang Zhou, Kaidi Zhang, Adeesh Desai, Temitope Amosa, Davood Soleymanzadeh, Jiuzhou Lei, Minghui Zheng, and Yu She. ContactWorld: What matters in vision-tactile world models for contact-rich manipulation. *arXiv preprint arXiv:2606.13877*, 2026.

[90] Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *RSS*, 2023.

[91] Yuhang Zheng, Songen Gu, Weize Li, Yupeng Zheng, Yujie Zang, Shuai Tian, Xiang Li, Ce Hao, Chen Gao, Si Liu, Haoran Li, Yilun Chen, Shuicheng Yan, and Wenchao Ding. OmniVTA: Visuo-tactile world modeling for contact-rich robotic manipulation. *arXiv preprint arXiv:2603.19201*, 2026.

[92] Xueyang Zhou, Yangming Xu, Guiyao Tie, Yongchao Chen, Guowen Zhang, Duanfeng Chu, Pan Zhou, and Lichao Sun. LIBERO-PRO: towards robust and fair evaluation of Vision-Language-Action models beyond memorization. *arXiv preprint arXiv:2510.03827*, 2025.

[93] Xinyue Zhu, Binghao Huang, and Yunzhu Li. Touch in the wild: Learning fine-grained manipulation with a portable visuo-tactile gripper. *NeurIPS*, 2026.

A NeoData Information

Scale and Episodes. NeoData comprises approximately 1.4M episodes totaling more than 30,000 hours of interaction, 3.3B timesteps, 8B RGB frames, and 10B tactile frames, collected across six embodiments by more than 90 human operators. The mean episode duration is about 80 seconds, and the distribution of episode lengths is approximately unimodal, as reported in Fig. 4. The object-frequency distribution is long-tailed, with cups and boxes followed by test tubes and building blocks among the most frequent categories. Task frequencies are summarized separately in Fig. 3, where a small set of high-volume skills, led by cloth folding, coexists with a broad tail of rarer tasks.

Data Definition. Every episode is a synchronized, temporally aligned stream of observations and actions. UMI observations contain one fisheye wrist RGB image and two visuo-tactile images from the left and right fingers, and all three streams are stored at a native resolution of 640×360. Robot embodiment observations additionally include external RGB views when available. Following Section 4.2, actions are stored as relative end-effector commands and gripper width for UMI data, and as tool center point commands together with absolute joint commands for robot embodiment data. Alongside the raw tactile images, each episode also carries the derived NeoForce force field at 640×360 resolution.

Task Types. The corpus spans more than 450 tasks, ranging from everyday household activities such as folding clothes, folding pants, wiping, and pick-and-place, to delicate contact-rich industrial operations such as small part assembly, screw tightening, USB and charger plugging, and accessory classification. As summarized in Fig. 3, the most common skill categories are folding, pick-and-place, and wiping or cleaning, followed by a long tail of contact-rich primitives including stacking, assembling, pouring, inserting, plugging, and pressing. The tasks are distributed across six scene types and a diverse object set grouped into containers, food and lab items, tools, deformables, kitchenware, and electronics. Their scene and leading object distributions are shown in Fig. 4.

B $\mathcal{N}_0$-TacUMI Details

Collection Device. The robot-free portion of NeoData is gathered with the $\mathcal{N}_0$-TacUMI device introduced in Section 3.3, a handheld parallel gripper carrying two tactile sensors, an infrared tracker for six-degree-of-freedom pose, and a magnetic encoder for finger separation. All streams are recorded synchronously at 30 fps against a common clock, and the same sensors and schema are reused on the teleoperated robots so that the two regimes are directly interoperable.

Action. At each timestep the tracker reports the gripper pose $P_t \in \mathrm{SE}(3)$ and the magnetic encoder reports the gripper width g_t. Rather than storing absolute poses as targets, we record the relative transform to a frame k steps ahead, $\Delta P_t = P_t^{-1} P_{t+k}$, and define the action as $a_t = (\Delta P_t, g_{t+k})$, pairing the relative end-effector motion with the commanded gripper width. This relative pose convention is invariant to the world frame and supports transfer across embodiments.

Wrist Images. The wrist view is captured by the fisheye RGB camera mounted on the handheld gripper, at 640×360 and 30 fps, temporally aligned with the pose and tactile streams. It provides the egocentric visual context used as a policy input and as the visual stream that NeoForce fuses with the tactile force field.

Tactile Images. Each finger embeds a camera-based visuo-tactile sensor (Section 5) that images the deformation of its sensing surface. The left and right tactile images T_t^l, T_t^r are recorded at 640×360 and 30 fps in lockstep with the wrist image and the action. During preprocessing, each tactile image is mapped through the learned inverse model g_θ into the 640×360 force field, yielding the unified tactile representation stored alongside the raw frames.

C Annotation Pipeline Details

We build a semi-automatic hierarchical annotation pipeline that labels every robot manipulation video with a four-level, temporally nested structure of task (L3), subtask (L2), action (L1), and atomic segment (L0). The pipeline runs in two stages, a task template generation stage and a per-episode annotation stage.

Task Template Construction For each task category, we first use the video summarization capability of a VLM, Gemini-3.5-Flash, to generate a time-independent task template from a few representative episodes. The template specifies the L3 overall objective shared by the task, an ordered set of L2 semantic phases, and the fixed L1 substeps under each phase. A human expert then reviews and confirms this template. This review is the only manual quality control step in the entire pipeline, and it removes the semantic wording drift that would otherwise arise across different episodes of the same task.

Episode Segmentation After the template is confirmed, we annotate each episode in a boundary-first, label-second order. The lowest level L0 atomic intervals are first delimited by two parallel branches that operate on two complementary modalities. The action signal branch locates a set of physically grounded atomic action boundaries from the kinematic signals of the end effector and gripper together with our self-collected tactile signals. These include the grasp boundary, whose criterion is that the gripper transitions from open to closed while the tactile signal rises sharply, indicating that an object is now held, with the remaining boundaries using analogous criteria, as well as the release boundary, the motion start and stop boundaries, the axis switch boundary, the vertical reversal boundary, and the contact and separation boundaries. Each boundary is aligned precisely to the moment at which the mechanical state changes. The visual boundary branch applies unsupervised Generic Event Boundary Detection (GEBD) [58], a method that perceives event boundaries from the visual stream. We subsample the video at 4 fps, extract ResNet-50 [22] features pretrained on ImageNet [13], and at each timestep compare the predictability, measured by cosine distance, between the average features of the 5 preceding and the 5 following frames, taking local maxima at points of abrupt semantic change as boundaries. This branch captures perceptual action transitions that the control signals cannot encode, such as the transition from folding to wiping that carries no clear mechanical discontinuity. The two boundary sets are then fused and deduplicated, and the total number of segments is capped at 24 while fine granularity is preserved, yielding a temporally fixed and more accurately partitioned set of L0 atomic intervals.

Hierarchical Labeling Finally, we let the same VLM, Gemini-3.5-Flash, label only these pre-segmented intervals. It assigns each L0 segment to the corresponding L1 and L2 level in the template and generates the associated description, after which deterministic code assembles the temporal extents of the L1, L2, and L3 levels from the bottom up. The temporal endpoints at all levels come from signal computation rather than VLM inference, so this design guarantees the accuracy of the temporal intervals at every level and fundamentally avoids the temporal drift of VLMs on long videos.

D Force Field Conversion and NeoForce Details

D.1 Real Force Calibration Data

We collect real tactile data and calibrate the applied force to build the training data for the tactile model. The calibration corpus is produced by a controlled annotation and curation pipeline rather than by uncalibrated in the wild demonstrations. As the acquisition source, a six-degree-of-freedom robot arm presses an indenter against the tactile surface while a six axis force and torque sensor records the reference load under a quasi-static protocol, in which the indenter is held stationary once it reaches a target depth so that the tactile image and the measured force are captured in a stable state. Under automated control, the indenter traverses the entire sensing plane at penetration depths from 1 to 6 mm in 1 mm increments, and shear supervision is enriched by tilting the indenter through rotation angles from 1° to 20° to cover diverse horizontal loads. Each curated sample takes the form of an optical flow image paired with a pixel aligned multi dimensional force field, where the two in-plane components are combined into a resultant shear stress and the normal

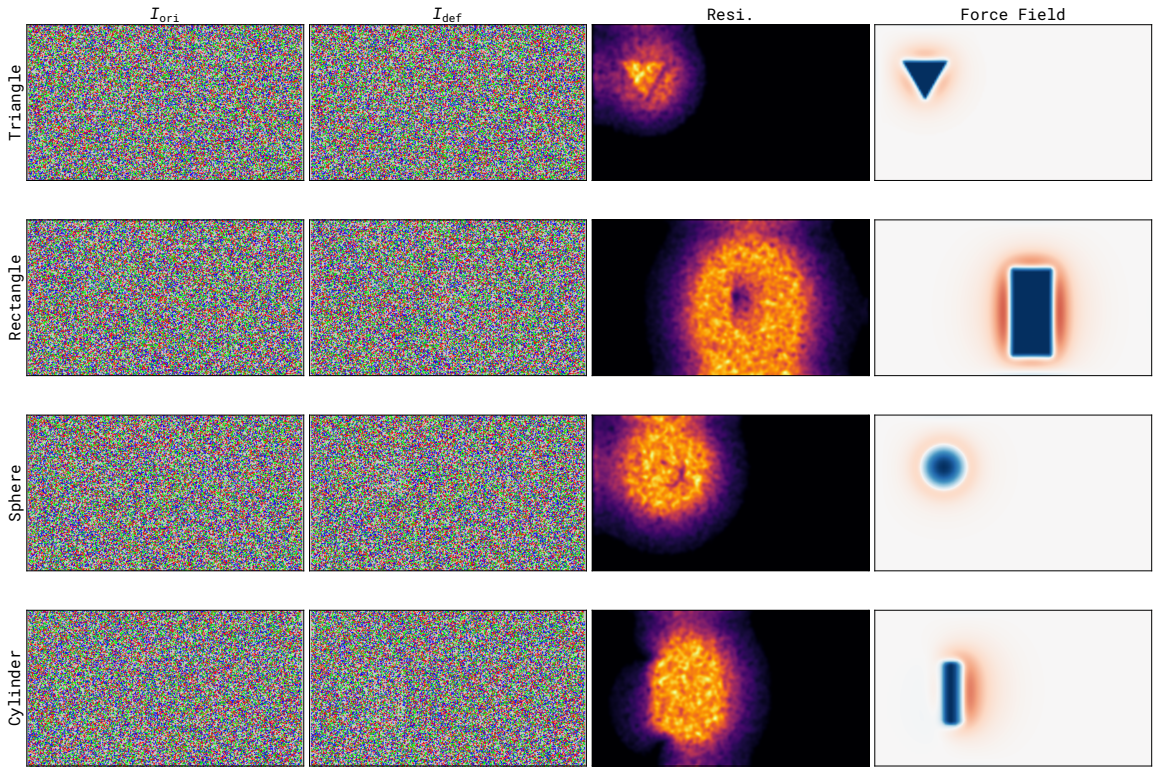
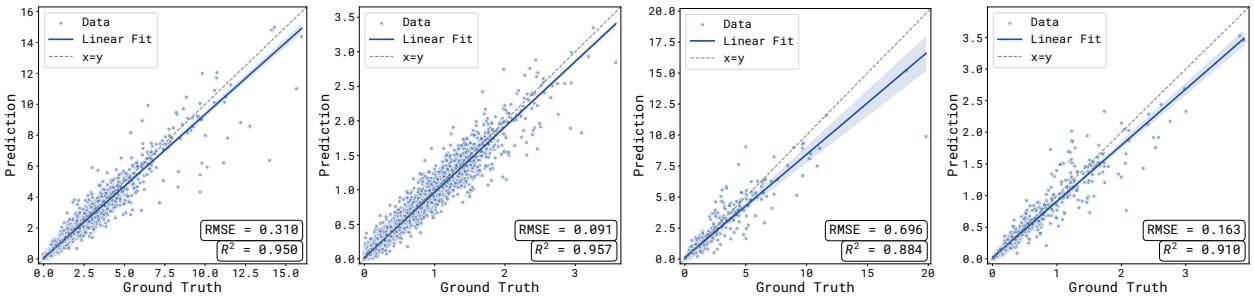


Figure 12 Simulated visuo-tactile data generation for four indenter shapes: triangle, rectangle, sphere, and cylinder. From left to right, the columns show the undeformed reference image (I_{ori}), the deformed image after loading (I_{def}), the residual between them (Resi.) that visualizes how the gel particles move at the contact region, and the resulting three-axis force fields. All four shapes are converted into pixel-aligned force fields, showing that the simulation pipeline produces consistent force supervision across diverse contact geometries.

component records pressure. The pipeline yields approximately 4,300 optical flow and multi dimensional force pairs, which are randomly shuffled and split into training, validation, and test partitions with an 80%, 10%, and 10% ratio, with the test partition held out exclusively for final evaluation.

D.2 Tactile Simulation Augmentation

To enlarge the diversity of contact geometries beyond what physical indentation can cover, we augment the real calibration corpus with simulated visuo-tactile data. The goal of the simulation is to synthesize supervised pairs that map visual deformation to a contact force field. Taking the relative contact position between the object and the sensing layer as the boundary condition, and combining the material and geometric parameters of the sensing layer, we drive an Abaqus finite element solver to obtain the distributed force and displacement fields. The displacement field then deforms the reference particle texture of the sensing surface through a renderer, producing an image pair before and after loading. Formally, each simulated sample is a triplet $(I_{\text{ori}}, I_{\text{def}}, F)$, where $I_{\text{ori}} \in \mathbb{R}^{1440\times2560\times3}$ is the undeformed reference image, $I_{\text{def}} \in \mathbb{R}^{1440\times2560\times3}$ is the deformed image rendered under the solved displacement field, and $F \in \mathbb{R}^{480\times640\times3}$ is the pixel aligned force field whose three channels store the x shear, the y shear, and the normal pressure. Since F shares the image pixel coordinates, the spatial mapping is fixed by the physical size of the sensing layer, with a physical length of 28.0 mm spanning 2560 pixels and a physical width of 16.0 mm spanning 1440 pixels. As shown in Fig. 12, the residual between I_{ori} and I_{def} visualizes the particle motion induced by contact, and the pipeline converts this deformation into a force field for each of the four indenter shapes, so that the augmented data follows the same force field form as the real calibration corpus.



(a) Pressure (real) **(b)** Shear (real) **(c)** Pressure (sim) **(d)** Shear (sim)

Figure 13 Evaluation of the force field conversion model. Each panel regresses the predicted force against the ground truth with a linear fit and the $x = y$ reference line, reporting RMSE and R^2, for pressure and shear in the real setting and on the simulated data.

D.3 Force Field Conversion Model

The tactile image to force field converter adopts GoogLeNet [62] as the backbone, whose inception modules aggregate multiscale features at a controlled computational cost, and modifies the output layer to regress the three-axis per-pixel force field that aligns with the shear and pressure targets. The converter is trained on a mixture of the real calibration corpus and the simulation augmented data described above, so that it benefits from both physically measured contact and the broader shape diversity of simulation. Following the reference training protocol, the model is implemented in PyTorch and optimized with AdamW using a batch size of 25, an initial learning rate of 0.001 decayed by a factor of 10 every 10 epochs, for a total of 50 epochs. The objective combines an L_1 per-pixel force regression loss with an L_2 contact segmentation loss, each normalized by the number of valid contact pixels so that the loss reflects the average per-pixel error and is robust to the varying size of the contact region. As reported in Fig. 13, the converter regresses force accurately in the real setting, reaching $R^2 = 0.950$ with RMSE $= 0.310$ for pressure and $R^2 = 0.957$ with RMSE $= 0.091$ for shear, and it remains reliable on the simulated data, reaching $R^2 = 0.884$ with RMSE $= 0.696$ for pressure and $R^2 = 0.910$ with RMSE $= 0.163$ for shear. The consistently high linearity and low error across both domains and both force components confirm that the converter performs effective force regression and can serve as the annotation model that labels the large-scale tactile corpus with physically meaningful force fields.

D.4 NeoForce Experiment Details

We train NeoForce on 20,000 real-robot and $\mathcal{N}_0$-TacUMI demonstrations spanning 30 tasks and evaluate it on 2,500 held-out episodes, which endows the model with broad contact-rich manipulation priors. Each training sample is a chunk of synchronized RGB observations and tactile force fields, and the chunk length is set to 4 in our experiments. The visual and tactile streams are patchified independently and then jointly processed by a shared transformer backbone, for which we adopt a ViT-B initialized from DINOv2 pretrained weights, on top of which sit two prediction heads. The reconstruction head decodes force fields and contact masks, while the latent prediction head is supervised by an exponential moving average teacher whose output serves as the ground truth in latent space. Given masked inputs, the student predicts these teacher latents for visual tokens, tactile tokens, cross-modal alignment, and masked patch tokens. The complete training objective follows Section 5. We train NeoForce on 8× NVIDIA A100 GPUs for 100,000 steps with a learning rate of 2×10^{-5}.

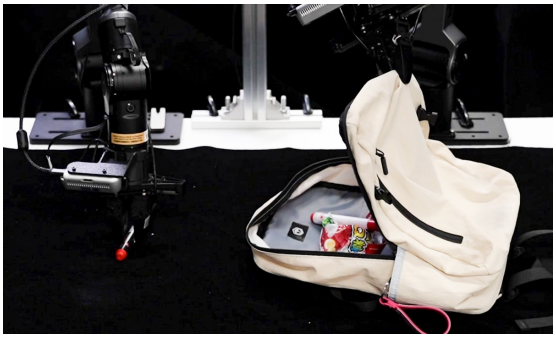
E NeoReal Task Suite

For each task, we provide a representative benchmark image and describe the manipulation skills and tactile capabilities it evaluates. NeoReal contains 10 real-world contact-rich manipulation tasks executed on physical robot setups equipped with the tactile fingers described in Section 3.2. Each task uses a standardized initial

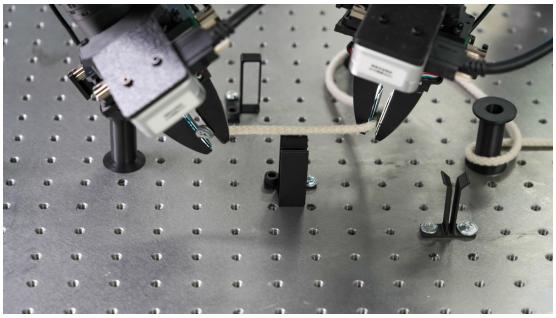
state distribution and reset protocol, and success is judged by a binary task-specific criterion such as complete insertion, stable stacking, correct routing, object placement, or final shape quality. For each task we show the real benchmark image extracted from the NeoReal task suite slide and describe the manipulation skill it exercises.

E.1 Data Demonstrations

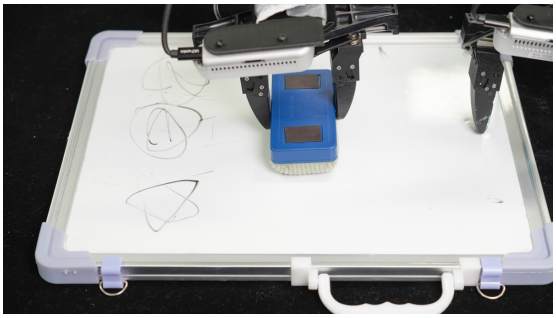
Cardboard Box Folding. The robot manipulates a cardboard box blank and folds it into the target box shape. The task stresses contact-rich bending, crease following, and force controlled pressing so that the cardboard seats correctly without tearing or collapsing.



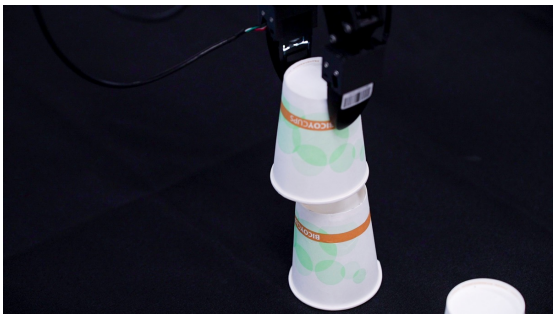
Bag Packing. The robot picks up the scattered items on the table, packs them into the bag, and pulls the bag zipper closed. The task combines cluttered-object grasping, placement inside a compliant container, and contact-rich zipper closing that requires sensing the pulling resistance along the zipper track.



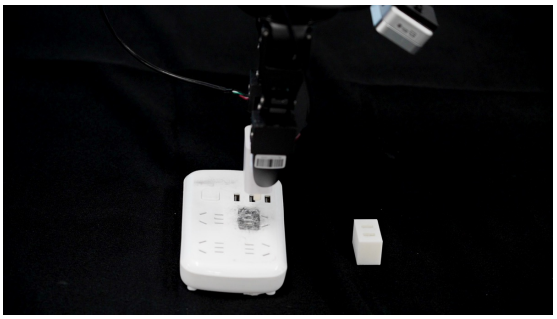
Cable Winding. The robot routes and winds a cable around the designated fixtures. The task requires maintaining cable tension, following the target path, and avoiding missed posts, tangles, or crossing errors.



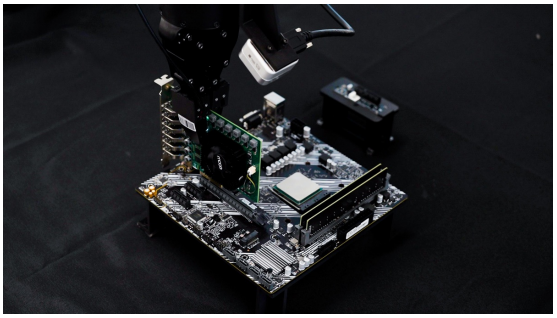
Board Wiping. The robot presses a wiping tool against a whiteboard and cleans the marked region. The task tests sustained surface contact, stable normal force, and coverage of the target area without losing contact or pushing the board out of place.



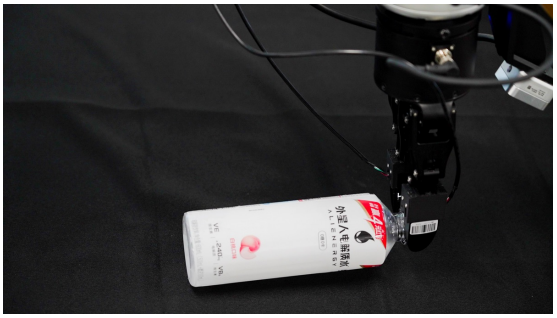
Cup Stacking. The robot stacks lightweight cups into a stable nested arrangement. The cups are thin-walled and easily deformed, so the policy must align them while limiting contact force during insertion and seating.



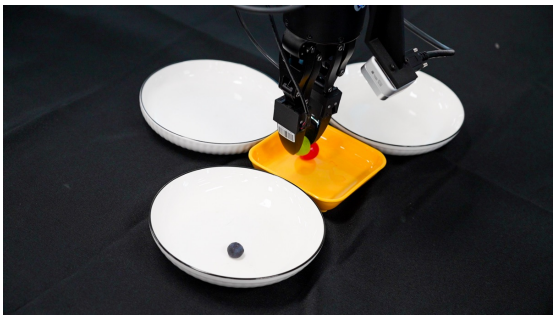
Socket Plugging. The robot grasps a plug, aligns it with the socket, and inserts it along the constrained direction. The task requires tactile detection of edge contact and insertion resistance to avoid jamming or wrong-orientation insertion.



Board Insertion. The robot inserts a circuit board into the target slot or fixture. This precise assembly task stresses small-clearance alignment, low-force insertion, and contact feedback to prevent bending or damage.



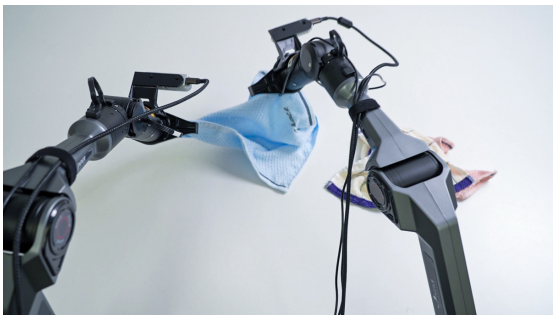
Bottle Standing. The robot recovers a tipped bottle to an upright, stable pose. The task combines curved-surface grasping, center-of-mass management, and controlled release so that the bottle remains standing after the action.



Fruit Collection. The robot picks and places fruit-like objects into the target dish. The task evaluates gentle grasping, transport, and release under a narrow force window that avoids both dropping and visible object damage.

Task	Milestone 1	Milestone 2	Milestone 3	Milestone 4
Cardboard Box Folding	Grasp and lift the box blank (20)	Fold one side of the box (30)	Press and fold the other side of the box (30)	Box holds the target shape without tearing or collapsing (20)
Bag Packing	Grasp the scattered items and put them in the bag (50)	Grasp the zipper pull (20)	Pull the zipper closed along the track (30)	
Cable Winding	Grasp the cable (20)	Route it to the first fixture under tension (30)	Wind around the remaining posts along the target path (30)	Complete routing with no missed post, tangle, or crossing (20)
Board Wiping	Grasp the wiping tool (15)	Grasp the blackboard to make it stable (25)	Sweep across the marked region while keeping contact (30)	Cover and clean the full target region (30)
Cup Stacking	Grasp the first cup (20)	Align it over the second cup (30)	Insert and seat it with limited contact force, no visible deformation (30)	Nested cups form a stable stack (20)
Socket Plugging	Grasp the plug (20)	Align it with the socket in the correct orientation (30)	Insert along the constrained direction without jamming (30)	Plug fully seated in the socket (20)
Board Insertion	Grasp the circuit board (20)	Align it with the slot at small clearance (30)	Insert with low force and contact feedback (30)	Board fully seated without bending or damage (20)
Bottle Standing	Grasp the tipped bottle on its curved surface (25)	Lift and reorient it toward upright (30)	Place it with the center of mass over the base (25)	Release in a controlled manner; bottle stays standing (20)
Fruit Collection	Grasp the fruit within the safe force window (30)	Lift without dropping or damaging it (25)	Transport it over the target dish (20)	Release it into the dish (25)
Towel Folding	Bimanual grasp of the towel from the initial state (20)	Flatten and spread the towel (25)	Align the target edges (25)	Fold the towel into the target shape (30)

Table 5 Task progressive score for the ten NeoReal tasks. Each task is decomposed into ordered milestones that follow its description in Appendix E. The points in parentheses are awarded when the milestone is reached and accumulate along the row, summing to 100 at full task completion. A milestone is credited only after all earlier milestones in the same row have been achieved, so the progressive score of a trial equals the total points of the last milestone it reaches.



Towel Folding. Two arms manipulate a towel from an unstructured initial state and fold it into the target shape. The task requires deformable-object grasping, edge alignment, flattening, and coordinated bimanual folding.

E.2 NeoReal Task Progressive Score

The binary success criterion in Appendix E scores a trial as 1 only when the full task goal is reached, which discards all information about how far a policy has progressed before failing. To provide a finer view of partial competence, we additionally define a task progressive score for every NeoReal task. Each task is decomposed into ordered sub-goals that follow its manipulation description, and a fixed number of points is awarded at each key milestone, so that the points accumulate monotonically as the task advances and sum to 100 at full completion. Tab. 5 lists the decomposition and the points assigned to each milestone, one row per task. Unless otherwise noted, a milestone is credited only if all earlier milestones in the same row have been achieved, and the reported progressive score of a trial is the total points of the last milestone reached.

Per-execution scoring. Following Tab. 5, every rollout is scored by the furthest milestone it reaches. We walk through the ordered milestones of the task and accumulate their points until the execution stops making progress or triggers one of the failure conditions in Section E.4, and the progressive score of that rollout is the total accumulated points. A rollout that completes the task passes all milestones and therefore scores the full 100 points, whereas a rollout that fails partway is credited only for the milestones it has already achieved. The per-task progressive score reported in Fig. 10 is the mean of these per-rollout scores over the 20 evaluation trials. A successful trial always attains all milestones and contributes the maximum 100 points, while partially completed trials still contribute positive points, so the progressive score of a task is always at least as large as its binary success rate and exposes the partial competence that the success rate alone hides.

E.3 NeoReal Training Settings

NeoReal policies follow a pretrain then post train recipe. In the pretraining stage, each policy is pretrained on over 400,000 episodes from NeoData to acquire broad contact-rich manipulation priors. In the post-training stage, the pretrained policy is specialized to each NeoReal task using 300 demonstrations that are collected specifically for that task.

All policies and baselines are trained on 8× NVIDIA A100 GPUs. Both the third-person external view and the wrist view are provided as visual inputs, and actions are parameterized as end-effector poses for all tasks. Single-arm tasks are executed on a Flexiv Rizon 4s arm, while dual-arm tasks are executed on Piper or ARX X5 arms.

E.4 NeoReal Evaluation Settings

Each task is evaluated over 20 trials. At the start of every trial we reset the robot arm to an initial pose sampled within a bounded range, so that the evaluation carries a controlled degree of randomization and tests generalization over the starting configuration. The positions of the manipulated assets are likewise randomized within their own ranges, adding spatial generalization to the object layout.

During a trial, we count the run as an execution failure if the arm reaches a joint limit, moves out of the workspace boundary, enters a self-locking configuration, or becomes stuck against or collides with the environment. If the arm repeats the same action more than 5 times without progress, we count it as a task progression failure. If the task is not completed within 5 minutes, we count it as a timeout failure. A trial is scored as a success only when the arm completes all milestones of the task defined in the task progressive score of Appendix E.2 without triggering any of the above failures. We report the mean success rate across the 20 trials.

F NeoSim Task Suite

This appendix describes the NeoSim benchmark, covering its simulation platform, the 12 contact-rich tasks (4 single-arm and 8 dual-arm), and the training and evaluation settings. For each task we show a rendered example, with the simulated RGB scene on the left and the corresponding particle gel images on the right following the format of Fig. 11, and describe the manipulation skill it exercises and the role that tactile feedback plays.

F.1 Simulation Platform

NeoSim is built on UniVTAC [6], a visuo-tactile simulation platform based on Isaac Sim and Isaac Lab [47] whose tactile layer is provided by TacEx [50]. The sensor gel is simulated as a finite element soft body with incremental potential contact and outputs tactile RGB images, marker motion, and gel depth maps. On top of this platform, we additionally simulate a markerless particle gel surface, where a dense speckle coating is advected by the finite element surface displacement, whose tangential component is taken from marker motion and whose normal component is taken from the gel depth map. We also derive per-vertex contact forces from the contact solver and interpolate them into the dense force field that serves as the tactile ground truth in NeoSim.

Each task is instantiated from a modular configuration file that specifies the task assets, the initial object layout, the randomization ranges, and the recorded observation streams. Initial object poses are randomized per episode within task-specific ranges. Success is judged by task-specific geometric criteria. For insertion tasks, for example, a trial counts as successful only when the axial alignment, the radial offset, and the insertion depth all reach their thresholds.

F.2 Single-arm Tasks

The four single-arm tasks are executed with a single arm and stress force regulation, insertion precision, and damage-free grasping.

Pour Ball. The arm grasps a cup filled with small balls and pours them onto a target plate. The cup is compliant and the balls shift during motion, so the policy must maintain a stable grasp force and modulate the pouring angle, and the particle gel rendering reveals how the contents load the fingers, which vision alone cannot observe.

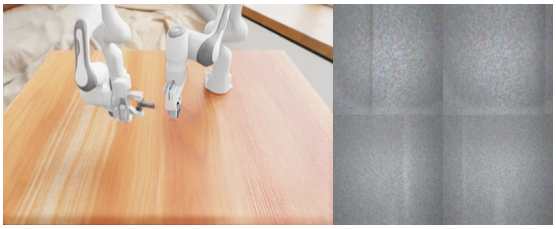
Unplug and Plug Charger. The arm disconnects a charger connector from a socket and inserts it again. Both the release and the insertion depend on sensing the seating force and the small resistance at the moment of connection, making this a canonical force-guided mating task.

Plug USB. A precision peg-in-hole task in which the arm aligns and inserts a USB connector into its port. The tight-clearance means that success hinges on tactile detection of edge contact and small misalignment before the connector jams.

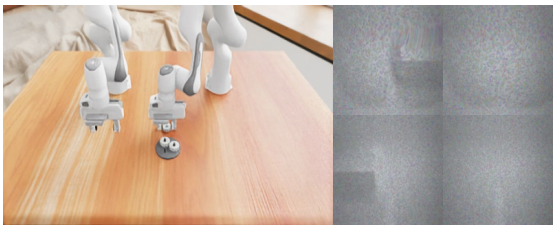
Grasp Chip. The arm grasps a thin, fragile chip and lifts it without crushing or dropping it. This task defines a narrow force window, since too little force lets the chip slip while too much force damages it, so it directly probes fine-grained force limiting.

F.3 Dual-Arm Tasks

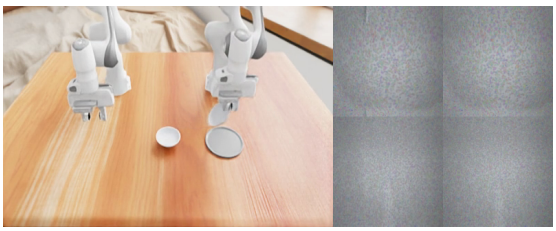
The eight dual-arm tasks require two arms and additionally test bimanual coordination and stable contact during handover, insertion, and stacking.



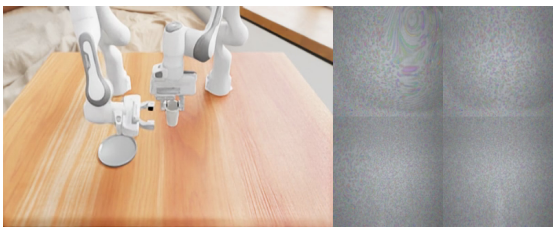
Insert Screw. One arm stabilizes a sleeve while the other inserts and threads a screw into it. Coordinated force sensing on both arms is needed to keep the parts aligned and to detect proper seating during threading.



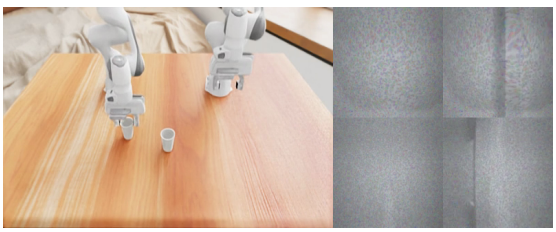
Place Gears. The arms pick up gears and thread them onto holder pegs. Tactile feedback confirms that each gear is seated flush against the peg rather than resting on an edge, a state that is visually ambiguous.



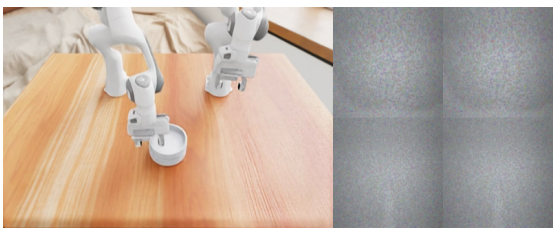
Unstack Bowl. One arm grips the rim of the top bowl in a nested stack and lifts it clear while the other stabilizes the remaining stack. Detecting the thin rim through touch is decisive, since the contact surface is small and easily missed by vision.



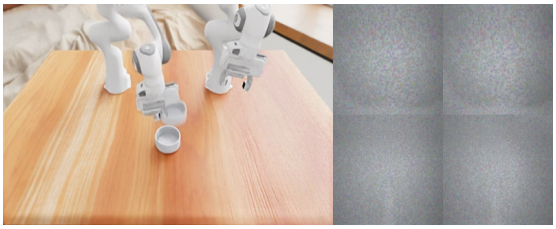
Cup Handover. One arm grasps a cup and hands it to the second arm, which then places it. The handover moment requires both arms to sense the shared contact so that the cup is neither dropped nor crushed during the transfer.



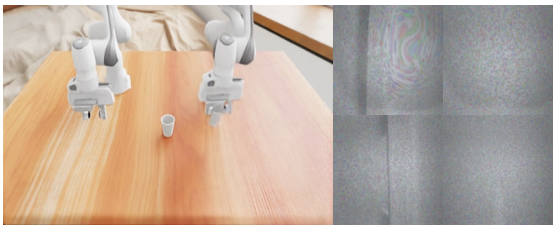
Stack Cups. Two separated cups are brought together and nested into a centered stack. The policy must align the cups and sense the seating contact as one cup enters the other.



Stack Plates. Two separated plates are combined into a centered stack. The thin, flat geometry makes tactile confirmation of centered, flush contact important for a stable stack.



Stack Bowls. Two separated bowls are brought together for centered nesting. As with plates, tactile alignment ensures the bowls seat concentrically rather than off-center.



Unstack Cup. One arm manipulates the top cup of a nested cup pair while the other stabilizes the base. The tight fit of nested cups increases the separation force the policy must sense and overcome, mirroring Unstack Bowl at a smaller scale.

F.4 NeoSim Training Settings

Demonstrations are collected automatically by scripted expert policies built on the cuRobo motion planner [61]. An episode is kept only when the planner succeeds and the final state passes the geometric success check, and collection continues until 100 successful episodes are obtained per task. All evaluated policies are trained on the same 100 demonstrations for each task. Every raw episode stores synchronized external and wrist RGB views, the two finger tactile streams, and both joint and end-effector states, so the demonstrations can be converted into the native input format of each policy, such as joint space actions for some baselines and end-effector poses combined with joint states for others. Each task-and-policy combination is trained on 4× NVIDIA A100 GPUs.

F.5 NeoSim Evaluation Settings

Each task is evaluated over 100 rollouts generated from 100 fixed random seeds that are disjoint from the seeds used for demonstration collection, so all policies face the same set of initial configurations. A rollout terminates when the task succeeds, when an early stop condition is triggered, or when the step limit is reached, and this limit is set per-task according to its difficulty. The reported metric is the mean success rate over the 100 rollouts. Policy inference runs on NVIDIA RTX 4090 GPUs and communicates with the simulator through a remote inference interface, which decouples policy serving from simulation and allows evaluations to be deployed across multiple machines under an identical environment configuration.