PRISM: Precision and contact-rich Real-world Industrial Skill dataset with Multimodal sensing

Tengbo Yu[1,2] Jiahao Wu[2] Hanning Wang[1] Rui Chen[3] Chuanhou Liu[4] Chuang Sun[5] Hangxin Liu[1†]

***Abstract*— Recent progress in robotic learning has been fueled by large-scale datasets collected in everyday environments. However, most existing datasets emphasize short-horizon, low-contact tasks such as pick-and-place, and therefore do not capture the precision control, force/torque or tactile regulation, and multimodal feedback required for industrial assembly. To address this gap, we introduce PRISM, a large-scale multimodal dataset for contact-rich industrial operations. The dataset spans more than 25 manipulation tasks (e.g., *electronic components plug/unplug, conveyor-based sorting*) and covers diverse mechanical constraints. PRISM includes more than 5,000 trajectories totaling 45 hours of teleoperated demonstrations, recorded using synchronized multi-view RGB-D, force/torque, tactile, and robot-state measurements. In contrast to datasets collected in household or laboratory settings, PRISM provides a realistic benchmark for multimodal perception and control under high-precision industrial constraints, and serves as a foundation for contact-rich, generalizable manipulation in real-world manufacturing environments. The dataset is open-sourced at: https://tengbo-yu.github.io/PRISM/**

I. Introduction

Robotic learning has recently undergone a paradigm shift driven by large-scale datasets collected in diverse everyday environments [1]. By leveraging broad visual coverage and imitation learning at scale, robots have demonstrated encouraging generalization for short-horizon, low-contact manipulation skills [2–5]. Such datasets have powered substantial advances in modern robotics, allowing learning-based methods to perform strongly on a wide range of problems, from motion planning [6–9] and task planning [10–13] to locomotion [14, 15] and anomaly detection [10, 16, 17]. However, this progress has not yet translated to industrial assembly, where task success depends on tight tolerance, persistent contact, and precise force/torque or tactile regulation. In such settings, vision alone is often insufficient: minute misalignments, frictional sticking, compliance, and insertion jamming can be visually ambiguous but are immediately expressed through force/torque or tactile signals and other contact cues [18–20]. Consequently, industrial manipulation demands learning not only from perception, but also how to control interaction reliably throughout long-horizon procedures.

A key bottleneck is the mismatch between existing robot learning datasets and the requirements of industrial operations. Many widely used datasets emphasize primitive skills (e.g., pick-and-place, in-hand manipulation, lifting, and stacking) [21–23] that are typically short-horizon and weakly contact-dependent [21, 23, 24], and therefore do not provide the high-resolution contact information needed for precision assembly. As highlighted in prior work, such primitive-focused datasets "failed to provide the critical data necessary for precise, contact-rich manipulation," while long-horizon assembly and disassembly require accurate contact measurements—especially force/torque and tactile—that visual sensing alone cannot reliably capture [25–28]. Recent efforts, such as REASSEMBLE, begin to address long-horizon and contact-rich scenarios by incorporating richer sensing streams such as multi-view cameras and six-axis force/torque measurements, demonstrating the importance of multimodal supervision for tight-tolerance manipulation [29, 30]. However, the remaining gap is not only dataset scale. Existing resources rarely combine multi-embodiment robot platforms, synchronized multimodal sensing, industrial contact-rich procedures, and diverse teleoperation interfaces within one unified dataset. This combination is essential for learning policies that must transfer across robot bodies, sensing modalities, operators, and contact conditions in production-like settings.

To bridge this gap, we present PRISM, a large-scale multimodal dataset for contact-rich industrial operations. PRISM covers 25 manipulation tasks, including `installing bearings`, `automotive material sorting` and so on, spanning diverse mechanical constraints representative of real industrial processes. We collect over 5,000 teleoperated trajectories totaling more than 45 hours, recorded with synchronized multi-view RGB-D, six-axis force/torque, and robot state streams, with tactile sensing included for a subset of episodes. We provide a comparison of commonly used robot learning datasets and their properties in Table I. Compared with existing datasets collected in household or laboratory environments, PRISM offers a realistic benchmark for learning multimodal policies under high-precision industrial constraints, emphasizing contact-rich interactions and stringent tolerances. Crucially, PRISM is designed around three dataset principles aligned with industrial requirements: higher capture standards, broader multimodal coverage, and large-scale diversity. First, we emphasize high-fidelity sensing, accurate calibration, and strict temporal synchronization to ensure that subtle contact events and force/torque transitions, as well as tactile transitions when available, are captured reliably. Second, we provide comprehensive multimodal observations that support learning contact-rich

† Corresponding author. Email: hx.liu@pku.edu.cn.
[1] State Key Laboratory of General Artificial Intelligence, School of Intelligence Science and Technology, Peking University, Emails: tengboyu1@gmail.com. [2] Delta Intelligence. [3] PKU-Wuhan Institute for Artificial Intelligence. [4] Hubei Humanoid Robot Innovation Center Co., Ltd. [5] China Academy of Information and Communications Technology.

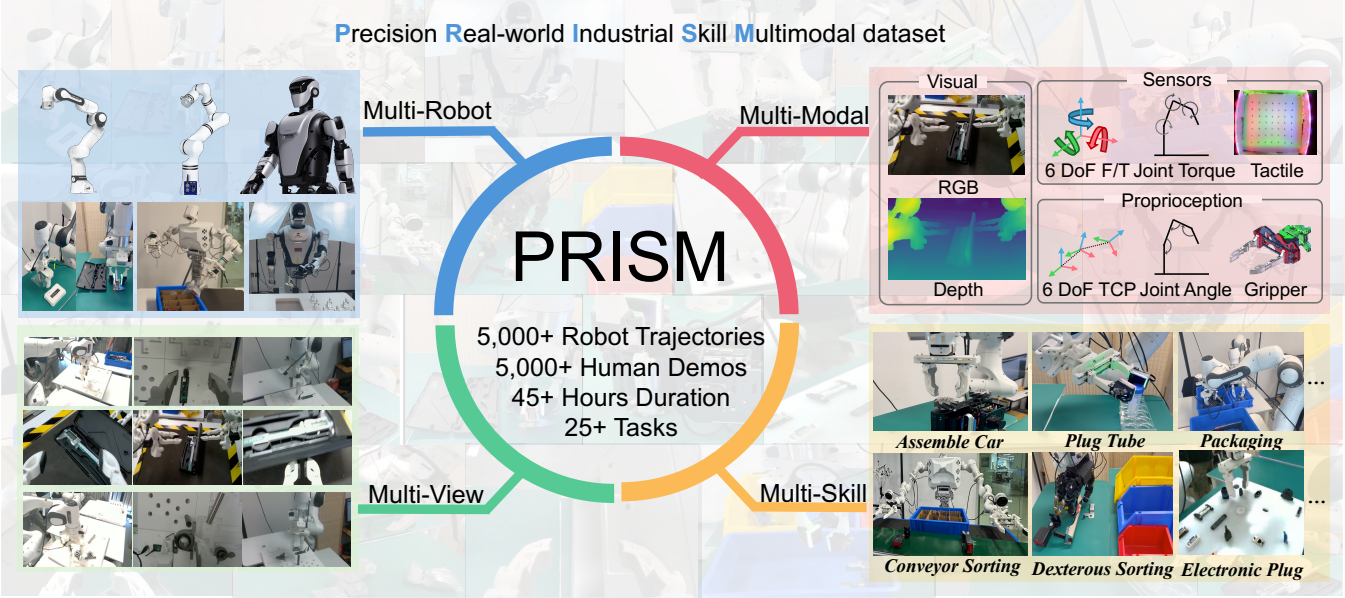


Fig. 1: **Overview of PRISM dataset.** We collect data using multiple robot platforms in diverse environments. Each manipulation episode includes synchronized multimodal observations, including visual, force/torque, and proprioceptive signals, with tactile signals available for a subset of episodes. For each episode, the full manipulation process is captured with well-calibrated multi-view cameras. The dataset spans a wide range of manipulation skills, and each episode is paired with a corresponding language description. In total, we provide more than 5,000 robot trajectories and 5,000 paired human demonstrations, covering over 45 hours of interaction across more than 25 tasks.

representations and controllers beyond purely visual policies. Third, we scale demonstrations across multiple tasks, robots, data collection platforms, and constraint structures to encourage generalization across industrial operation families. By offering a rich, multimodal dataset, PRISM fosters the development of adaptive and versatile robotic systems capable of tackling the challenges of long-horizon, contact-rich manipulation. To summarize, our contributions are as follows:

1) **Industrial realism with contact-rich constraints.** We introduce PRISM, a multimodal dataset centered on high-precision industrial operations with diverse mechanical constraints.
2) **High-fidelity synchronized multimodal data.** PRISM provides time-aligned multi-view RGB-D, six-axis force/torque, robot state streams, and tactile signals for a subset of episodes for contact-rich learning.
3) **Scale for generalizable industrial manipulation.** PRISM contains more than 5,000 trajectories and more than 45 hours of demonstrations, enabling systematic training and evaluation of multimodal perception and control under industrial constraints.

II. Related Work

A. Large-Scale Datasets in Robotic Manipulation

In recent years, the field of robot learning has undergone a significant paradigm shift, moving from small-scale, task-specific datasets [36, 37] toward large-scale, cross-embodiment data [8, 32, 38]. To achieve stronger generalization and train universal robot foundation models, the community has introduced several milestone datasets by integrating data from various robotic platforms.

For instance, Open X-Embodiment [21] aggregates hundreds of thousands of demonstration trajectories, aiming to cover diverse environments and tasks through sheer data volume. Research such as RT-1 [8] and BridgeData V2 [32] has further demonstrated that models trained on massive datasets can exhibit impressive zero-shot transfer capabilities, adapting to unseen instructions and scenarios. The success of these datasets is primarily attributed to their scale, which enables models to learn general visuomotor priors. This allows them to perform object manipulation and relocation tasks within unstructured, real-world settings.

B. High-Precision Datasets for Industrial Manipulation

Recent advancements in robot learning have been driven by the emergence of large-scale datasets collected across diverse environments. Foundation initiatives such as Open X-Embodiment [21], RT-1 [8], and BridgeData V2 [32] have aggregated extensive demonstrations to enable general-purpose robotic control. These datasets typically focus on unstructured settings, such as household or kitchen environments, where robots perform fundamental, short-horizon tasks like picking, placing, and object rearrangement.

To address the limitations of primitive skills, research has shifted towards long-horizon and contact-rich manipulation tasks. FurnitureBench [33] has made progress in addressing long-horizon tasks. However, despite the extended temporal horizon, furniture assembly generally involves loose mating parts, lacking the tight tolerances of precision manufacturing.

To tackle high-precision challenges, recent benchmarks

Dataset	# Demos	Sensors	Robot	Collection Method	Contact Rich
BC-Z [31]	263k	1 RGB Camera, Robot Proprioception	Everyday Robots	VR teleoperation	✗
RT-1 [8]	130k	1 RGB Camera, Robot Proprioception	Everyday Robots	VR teleoperation	✗
BridgeDatav2 [32]	60.1k	4 RGB Cameras, 1 Depth Camera, Robot Proprioception	WidowX	VR teleoperation	✗
FurnitureBench [33]	5.1k	2 RGB Cameras, Robot Proprioception	Franka	VR teleoperation	✓
DROID [23]	76k	3 RGBD Cameras, Robot Proprioception	Franka Emika Panda Research 3	VR teleoperation	✓
RH20T [24]	110k	8-10 RGBD Cameras, 1 Microphone, F/T Sensor	Multiple	haptic teleoperation	✓
Assembly101 [34]	4.3k	8 RGB Cameras, 4 Mono Cameras	N/A	Human demonstration	✓
FAILURE [35]	229	1 RGBD Camera, 1 Microphone	Baxter	Scripted	✗
REASSEMBLE [29]	4k	3 RGB cameras, Robot Proprioception, 1 Event camera, 3 Microphones, F/T Sensor	Franka Emika Panda Research 3	Haptic teleoperation	✓
PRISM	5k	2-4 RGBD Cameras, Robot Proprioception, F/T Sensors, **Tactile Sensors**	Multiple robots with **3 different types end-effector**	**Exoskeleton, Tracker, VR teleoperation**	✓

TABLE I: **Datasets comparison.** We compare several commonly used datasets based on the number of demonstrations, the sensors used during data collection, the robotic platform, the data collection method, and whether the dataset contains contact-rich tasks.

Conf.	Robot	Gripper	Teleoperation
Cfg 1	Franka	3D Printed	Tracker
Cfg 2	Franka	Visuotactile Gripper	Tracker
Cfg 3	Franka	Visuotactile Dexterous Hand	Tracker
Cfg 4	Franka	3D Printed	Exoskeleton
Cfg 5	Realman	3D Printed	Exoskeleton
Cfg 6	Realman	3D Printed	VR
Cfg 7	LEJU	Robotiq-85	VR

TABLE II: Hardware specification of different configurations.

like REASSEMBLE [29] and RH20T [24] have adopted standardized industrial protocols, such as the NIST Task Board. These works emphasize that visual sensing alone is often insufficient for tasks like gear meshing or connector insertion, necessitating the use of high-frequency force-torque sensors and proprioception. While incorporating such modalities improves physical feedback, it introduces hardware dependencies and complicates data collection, often limiting dataset scale (e.g., REASSEMBLE contains approximately 4,000 demonstrations).

In contrast, our work bridges the gap between large-scale generalist datasets and high-precision industrial needs. We posit that vision-based policies can implicitly learn the contact dynamics required for tight-tolerance assembly when supported by substantial data scale and hardware diversity. To this end, we introduce a dataset comprising 5,000 trajectories covering a wide range of industrial task scenarios, collected across multiple robotic platforms including Franka, Realman, and LEJU. Furthermore, to enable rigorous evaluation of interaction quality, we incorporate a subset of data equipped with tactile sensors, serving as a high-fidelity benchmark for fine-grained testing and validation.

III. DATASETS

We introduce Precision and contact-rich Real-world Industrial Skill dataset with Multimodal sensing (PRISM), a dataset for contact-rich industrial manipulation. Fig. 1 shows an overview, and Table I compares PRISM with representative public datasets.

A. *Properties of PRISM*

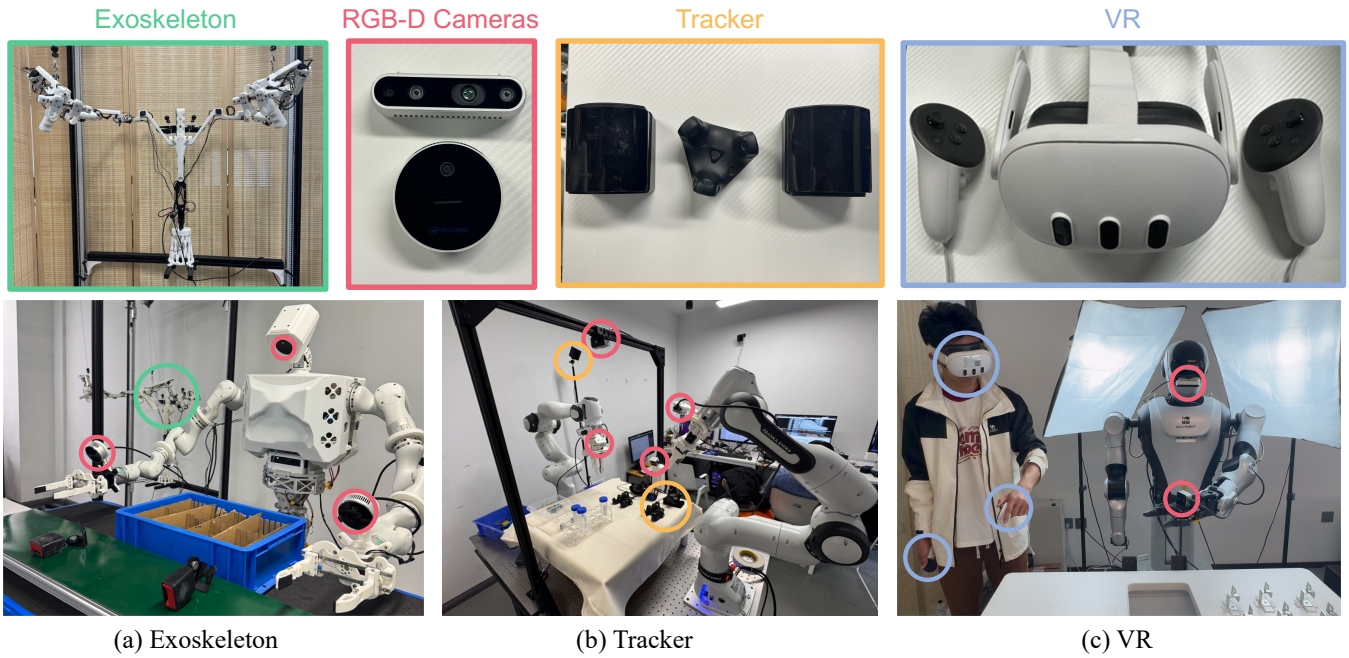
PRISM is designed for general contact-rich industrial manipulation under high-precision constraints. We emphasize diversity, multimodal sensing, scale, and compositional task structure.

Modal	Size	Frequency
RGB image	$540 \times 960 \times 3$	15 Hz
Depth image	540×960	15 Hz
Visuotactile image	256×256	30 Hz
Robot joint angle	6 / 7	15 Hz
Robot joint torque	6 / 7	15 Hz
Gripper Cartesian pose	6 / 7	15 Hz
Gripper width	1	15 Hz
6DoF F/T	6	100 Hz

TABLE III: Data information of different modals.

a) Diversity: PRISM spans contact-rich tasks on the NIST Assembly Task Board [39] and production-oriented tasks such as packaging, installation, and conveyor-based sorting. These tasks cover different mechanical constraints in industrial manipulation. We collect data with 3 robot platforms and 3 gripper types, allowing similar task families to appear under different kinematics, grasping affordances, and contact patterns. Hardware details are shown in Table II.

We further vary teleoperation interfaces, environments, and trajectories. Demonstrations are collected with 3 tele-



(a) Exoskeleton (b) Tracker (c) VR

Fig. 2: **Illustration of our data collection platform.** Including three different teleoperation platforms with multi-view RGB-D cameras.

operation modalities—exoskeleton, tracker, and VR-based control—as shown in Fig. 2. We change tabletop surfaces, backgrounds, lighting, object placements, and object poses across sessions. Data are collected by 8 volunteers, adding natural variation in motion style, contact timing, and corrective behaviors. We also include episodes with intentional human perturbations during manipulation to improve robustness to non-ideal contact events.

b) Multi-Modal: Each trajectory in PRISM is recorded with synchronized multimodal observations to capture both geometric perception and contact interaction signals that are critical for contact-rich industrial manipulation. Specifically, we log RGB-D images from the vision system to provide scene appearance and depth geometry, and visuotactile images from tactile sensors/end-effectors to directly observe local contact patterns, deformation, and slip cues. To characterize interaction dynamics, we record the 6-axis end-effector force/torque wrench as well as joint torques, enabling learning of force- and contact-sensitive control policies. In addition, we provide proprioception, which encompasses joint angles/torques, end-effector cartesian pose, and gripper states, offering a complete description of the robot's internal state for state-feedback control and multimodal sensor fusion. All information is collected at the highest frequency supported by our data collection platform and saved with corresponding timestamps, and the details are given in Table III

c) Scale: PRISM is collected at a substantial scale to support data-hungry multimodal learning. The dataset contains over 5,000 robot trajectories, together with an equal number of paired human demonstrations, totaling approximately 27 million images across vision and visuotactile streams. Fig. 3 summarizes the dataset statistics on the manipulation time for each episode in our dataset. Most episodes have durations ranging from 25 to 35 seconds, reflecting the sustained contact and multi-step nature of industrial operations. With its combination of large trajectory count, high-volume multimodal visual data and coverage across multiple industrial settings and robot embodiments, PRISM constitutes one of the most comprehensive multimodal datasets available for contact-rich manipulation in real-world industrial scenarios.

d) Compositionality: PRISM is organized to support learning at multiple stages by treating complex industrial procedures as compositions of reusable sub-tasks. Many real-world operations are naturally multi-step, where each step is meaningful on its own yet also contributes to a larger goal. For example, during the assembly of an autonomous patrol vehicle, installing the wheels, mounting the hubs, plugging the camera and installing the radar can be considered a single end-to-end task when viewed as the complete assembly process; at the same time, each step can be treated as an individual task with its own objectives, constraints and success criteria. This compositional structure allows users to (i) train models on fine-grained skills and evaluate them in isolation, (ii) study how performance scales with increasing procedural length, and (iii) combine learned components to solve longer-horizon industrial workflows under consistent sensing and annotation.

B. Data Collection and Processing

Unlike most existing datasets that rely on a single teleoperation interface, we collect demonstrations using three complementary teleoperation methods (Exoskeleton-, Tracker-, and VR-based). This design not only broadens the coverage of human control styles and interaction behaviors, but also

enables a controlled comparison of how the same task, collected under different teleoperation modalities, differ in data quality (e.g., smoothness, precision, contact stability and failure rates) and how these differences translate into downstream learning performance. We further detail the hardware configurations, multimodal synchronization and calibration, and post-processing steps that ensure consistent, high-fidelity multimodal records across robots, end-effectors and collection sessions.

a) Collection: Fig. 2 illustrates the three teleoperation platforms used to collect PRISM. (a) Exoskeleton teleoperation platform. This platform is a bimanual robot built with two Realman RM75-6F arms mounted on a torso with 3 waist DoFs, resulting in 17 DoFs for the full body, leading to a wider operation space. The system is equipped with 6-axis and end-effector force/torque sensors, two parallel-jaw grippers, a first-person head-mounted camera, and two wrist cameras. During data collection, we simultaneously record both the robot states and the exoskeleton joint angles, enabling paired human-robot motion traces for the same execution. (b) Tracker-based teleoperation platform. This platform consists of two Franka Emika Panda arms (14 DoFs total) with access to joint torque signals. Visual sensing includes two wrist cameras, a third-person camera, and an overhead camera. The end-effector is modular: the standard parallel jaw grippers can be replaced with either a visuotactile gripper or a visuotactile dexterous hand to capture contact-rich interactions with tactile images. (c) VR teleoperation platform. This platform uses a LEJU upper-body humanoid robot, instrumented with a wrist camera and a first-person head camera, and is controlled through a VR interface to collect demonstrations in scenarios that benefit from immersive first-person manipulation. We recruited 8 data-collection volunteers. Each volunteer underwent standardized training before collecting demonstrations. After collection, the same volunteers perform episode filtering, annotation, and scoring, providing structured quality control signals for downstream use.

b) Processing: After data collection, we process raw logs into episode files with clean timestamps, calibrated sensor parameters, and metadata suitable for multimodal learning. Since different sensors run at different native rates, we keep original timestamps for all streams and provide per-episode indexing so that modalities can be aligned by timestamp during training (e.g., nearest-neighbor matching or interpolation). We remove incomplete episodes and trim leading/trailing idle segments based on task start/end markers and operator annotations.

Calibration and frame unification. For each platform, we compute and store the full transformation graph needed to express all observations in consistent coordinate frames. For the Realman exoskeleton-controlled bimanual platform (Fig. 2-a), we unify the two arms, waist, head camera, and wrist cameras under a shared world frame; end-effector wrench measurements are expressed in a clearly defined sensor frame and optionally transformed to the end-effector or base frame. We also synchronize and align exoskeleton joint trajectories with the corresponding robot joint trajectories, storing both streams with consistent joint ordering and metadata so that paired human–robot kinematics can be directly compared. For the Panda tracker-based platform (Fig. 2-b), we calibrate the wrist cameras, third-person camera, and top-down camera w.r.t. the robot base/world frame, and standardize joint torque conventions. When the end-effector is replaced by a visuotactile gripper or a visuotactile dexterous hand, we additionally calibrate the tactile camera(s) and record their intrinsics/extrinsics relative to the end-effector frame. For the LEJU VR platform (Fig. 2-c), we calibrate the head and wrist cameras and align VR control signals with robot kinematics, yielding a consistent representation of end-effector motion and camera observations.

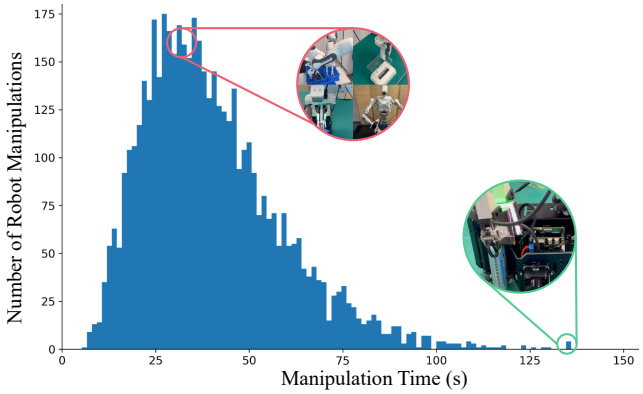


Fig. 3: Statistics on the execution time of different robotic manipulations in our dataset.

Unified episode packaging. Finally, we serialize all processed data into a common schema shared across the three platforms: robot states/actions (including joint angles/torques, end-effector pose, gripper states), end-effector wrench (when available), multi-view RGB-D, visuotactile imagery (when available), calibration parameters, timestamps, platform identifiers, task identifiers, outcome labels, and volunteer-provided ratings. This standardized composition enables training and evaluation across heterogeneous robots, end-effectors, and teleoperation interfaces while maintaining consistent multimodal alignment.

IV. EXPERIMENTS

In this section, we evaluate PRISM as a training and benchmarking resource for contact-rich industrial manipulation. Our experiments are designed to (i) quantify how well state-of-the-art imitation learning policies leverage the dataset's multimodal signals under high-precision constraints, and (ii) examine how demonstration collection choices affect downstream performance, particularly when the same task is collected via different teleoperation interfaces. Concretely, we train and compare three representative behavior-cloning baselines—ACT [40], Diffusion Policy (DP) [6], and π_0 [4] —under consistent training protocols, and report performance using task-relevant metrics such as success rate.

A. Experimental Setup

a) Platform: Experiments are conducted on a bimanual Realman robot with a 3 DoFs waist, matching the embodiment used during data collection. Each arm is equipped with a 3D-printed parallel-jaw gripper, and visual observations are captured using Intel RealSense D515 RGB-D cameras. The workspace layout, background and tabletop settings are kept consistent with the data collection environment to ensure that evaluation reflects the same sensing configuration and physical constraints encountered in the dataset. Fig. 4 illustrates our robot platform.

b) Procedure: We evaluate policies on three representative tasks collected in our industrial workcell: `electronic component plug/unplug`, `packaging a vernier caliper` and `conveyor-based sorting`. These tasks reflect three common categories in industrial manipulation: precise contact-rich operations, product packaging and dynamic object sorting, respectively. For each task, we collect 200 demonstrations, each containing synchronized RGB-D observations, action sequences and 6-axis end-effector force/torque.

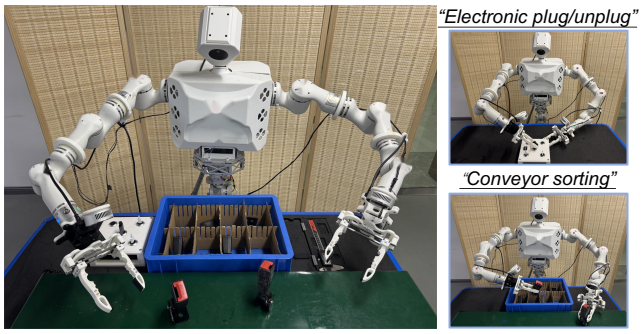
Our training follows a two-stage protocol. We first pretrain each model on the full PRISM dataset (5,000 demonstrations) to learn general multimodal representations and contact-rich control priors. We then fine-tune on the task-specific subset corresponding to the target task. To study data efficiency and the effect of large-scale pretraining, we evaluate: (i) training with 100 vs. 200 task demonstrations, and (ii) with vs. without dataset-level pretraining (e.g., training from scratch using only the task-specific data). All evaluations are performed on the real robot platform and repeated 20 times for each configuration. For the two static tasks (`electronic component plug/unplug` and `caliper packaging`), each trial has a 30-second limit, and it is marked successful if the task-specific completion criteria are satisfied within this budget.

c) Implementation Details: We convert all collected demonstrations into the LeRobot v3.0 [41] dataset format and train policies using the official configuration provided in the LeRobot repository for the three baselines. For each policy, training is split into two stages with a fixed step budget: we allocate 10% of the total optimization steps to pretraining on the full dataset and the remaining 90% to task-specific fine-tuning (e.g., for π_0 we use 30,000 total steps, with 3,000 steps for pretraining and 27,000 steps for fine-tuning). We set the chunk size to 15, which corresponds to 1 second at 15 Hz. The RGB-D images are scaled to 540×960 during training and testing.

B. Experimental Results

We present the model's success rates under different numbers of training demonstrations in Table IV. Increasing the amount of task-specific data from 100 to 200 demonstrations leads to a clear and consistent performance improvement. This trend is most pronounced for the two static, contact-rich tasks—electronic component plug/unplug and vernier caliper packaging—where additional demonstrations provide better coverage over initial condition variations and, critically, a richer distribution of contact transitions and corrective behaviors. With more demonstrations, policies exhibit (i) higher success rates, (ii) fewer early failures caused by misalignment or premature contact, and (iii) more stable execution trajectories with reduced oscillatory corrections. These findings indicate that even within a fixed task definition, contact-rich industrial manipulation remains data-intensive: collecting a larger set of demonstrations substantially improves the learner's ability to handle small geometric deviations and the long tail of interaction outcomes. For conveyor-based sorting, additional demonstrations also improve performance, but the gains are smaller. This is because sorting on a moving conveyor couples perception latency, prediction error, and time-critical actuation. While more data helps the policy see a broader range of object poses and motion states, success still depends on precise temporal coordination and robust tracking, which are harder to recover purely through behavioral cloning under limited coverage of motion patterns.

We also illustrate the model's success rate under different training processes in Table V. Pretraining on the full 5,000-trajectory dataset consistently improves downstream performance after fine-tuning on task-specific data. Beyond higher average success rates, pretrained models tend to be more robust: they fail less catastrophically and more often exhibit partial progress or recovery-like behaviors (e.g., re-approach after a minor misalignment, delayed grasp adjustment, or conservative contact engagement). This suggests that large-scale pretraining provides useful priors for (i) multimodal feature extraction, (ii) common motion motifs in industrial



(a) Experiment Setups (b) Complex Tasks

Fig. 4: **Real-World Experiments** with two Realman manipulators.

	`Electronic plug/unplug`		`Calipers packaging`		`Conveyor sorting`	
Method	100	200	100	200	100	200
ACT	5	10	30	45	15	25
DP	0	5	25	50	10	25
π_0	10	**20**	65	**80**	50	**75**

TABLE IV: Average success rates (%) of three different policies trained with 100 or 200 demonstrations per task and evaluated over 20 episodes.

Method	Pretrain	Electronic plug/unplug	Calipers packaging	Conveyor sorting
ACT	✗	10	45	25
	✓	10	55	30
DP	✗	5	50	25
	✓	10	55	20
π_0	✗	20	80	75
	✓	**25**	**85**	**85**

TABLE V: Average success rates (%) of three different policies w/ and w/o pretraining, trained using 200 demonstrations per task and evaluated over 20 episodes.

Method	Collection	Electronic plug/unplug	Calipers packaging	Conveyor sorting
ACT	VR	5	20	10
	Exoskeleton	10	45	25
DP	VR	0	15	20
	Exoskeleton	5	50	25
π_0	VR	5	35	40
	Exoskeleton	**20**	**80**	**75**

TABLE VI: Average success rates (%) of three different policies trained on 200 demonstrations per task collected via VR vs. Exoskeleton teleoperation, evaluated over 20 episodes.

workcells, and (iii) contact-rich stabilization strategies that transfer across tasks. In contrast, models trained from scratch on only 100–200 demonstrations are more brittle: they often overfit to dominant trajectories, show less tolerance to perturbations in object pose, and degrade noticeably under minor distribution shifts. The benefit of pretraining is particularly visible when the evaluation involves unexpected disturbances or slight changes in scene configuration. Pretrained policies are more likely to maintain stable behavior under small deviations (e.g., slight object pose shifts, transient occlusions, or contact-induced pose drift). In contrast, policies trained without pretraining often collapse once their execution deviates from the trajectory distribution observed during data collection, leading to error accumulation and task failure.

Despite these improvements, overall performance remains far from satisfactory, especially on the two hardest aspects of our benchmark: dynamic manipulation and precise force-aware operations. For conveyor sorting, failures are often caused by imperfect timing and state estimation under late grasps, unstable grasps due to relative motion at contact, or missed intercepts when object speed/pose deviates from training patterns. For electronic component plug/unplug, failures commonly stem from force-control limitations: slight misalignments lead to jamming, excessive contact forces, or repeated micro-corrections that do not converge within the time limit. These observations indicate that current imitation-learning policies—even with multimodal inputs and large-scale pretraining—still struggle to (i) reason about fast-changing task state in dynamic scenes and (ii) regulate contact forces reliably under tight tolerances.

The above results highlight that more demonstrations and larger scale pretraining help improve coverage and robustness, but they do not fully resolve the underlying difficulty of industrial contact-rich manipulation. It still requires stronger interaction modeling (e.g., explicit contact state estimation), better incorporation of force/visuotactile feedback into closed-loop control, and learning objectives or architectures tailored to time-critical dynamic and precision force regulation.

Beyond the training protocol variations described above, we also study how teleoperation modality influences downstream policy learning. Specifically, we collect demonstrations for the same tasks using VR and an exoskeleton-based platform, with 200 demonstrations per task under each modality, and train policies from scratch under identical model configurations and optimization budgets. The results show that policies trained on VR-collected data consistently achieve lower success rates than those trained on exoskeleton-collected data, and they also exhibit reduced execution efficiency (e.g., longer completion times and more corrective motions). We attribute this gap primarily to differences in perceptual feedback during data collection: in our VR setup, operators observe the scene largely through 2D rendered views, which provide weaker depth and spatial cues than natural 3D perception. Moreover, the exoskeleton platform induces operator motions that more closely match the robot's joint-space kinematics, rather than purely tracking end-effector controller positions as in VR. These limitations can induce systematic imprecision during demonstration—such as suboptimal approach trajectories, noisier alignment behaviors, and less stable contact engagement—thereby degrading demonstration quality and lowering data collection efficiency. These findings suggest that, for high-precision contact-rich industrial manipulation, the choice of teleoperation platform is a critical factor that shapes demonstration fidelity and can materially impact the performance of policies.

V. Conclusion

In this paper, we present PRISM, a large-scale multimodal dataset for contact-rich industrial manipulation under high-precision constraints, covering both NIST Assembly Task Board operations and production-oriented workflows such as packaging and conveyor-based sorting, collected across multiple robot embodiments and end-effectors. The collected data include synchronized RGB-D, visuotactile, joint torques, and proprioception, enabling research on multimodal perception and contact-rich control in realistic industrial settings. We further validate PRISM by training and evaluating representative imitation-learning baselines on real robots, showing that existing policies' overall performance remains limited for dynamic object manipulation and contact-rich precision operations, highlighting important opportunities for future work on stronger multimodal fusion, interaction-state estimation, and closed-loop contact regulation for generalizable industrial manipulation.

REFERENCES

[1] M. Yapıcı, A. Tekerek, and N. Topaloglu, "Literature review of deep learning research areas," *Gazi Journal of Engineering Sciences*, vol. 5, pp. 188–215, 12 2019.

[2] G. Lu, T. Yu, H. Deng, S. S. Chen, Y. Tang, and Z. Wang, "Anybimanual: Transferring unimanual policy for general bimanual manipulation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 13 662–13 672.

[3] T. Yu, G. Lu, Z. Yang, H. Deng, S. S. Chen, J. Lu, W. Ding, G. Hu, Y. Tang, and Z. Wang, "Manigaussian++: General robotic bimanual manipulation with hierarchical gaussian world model," in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2025, pp. 12 232–12 239.

[4] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, S. Jakubczak, T. Jones, L. Ke, S. Levine, A. Li-Bell, M. Mothukuri, S. Nair, K. Pertsch, L. X. Shi, J. Tanner, Q. Vuong, A. Walling, H. Wang, and U. Zhilinsky, "π_0: A vision-language-action flow model for general robot control," 2026. [Online]. Available: https://arxiv.org/abs/2410.24164

[5] M. J. Kim, C. Finn, and P. Liang, "Fine-tuning vision-language-action models: Optimizing speed and success," *arXiv preprint arXiv:2502.19645*, 2025.

[6] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," *The International Journal of Robotics Research*, 2024.

[7] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, *et al.*, "Openvla: An open-source vision-language-action model," *arXiv preprint arXiv:2406.09246*, 2024.

[8] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, *et al.*, "Rt-1: Robotics transformer for real-world control at scale," *arXiv preprint arXiv:2212.06817*, 2022.

[9] B. Zitkovich et al., "RT-2: Vision-language-action models transfer web knowledge to robotic control," in *7th Annual Conference on Robot Learning*, 2023.

[10] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman, *et al.*, "Do as i can, not as i say: Grounding language in robotic affordances," *arXiv preprint arXiv:2204.01691*, 2022.

[11] W. Huang, F. Xia, T. Xiao, H. Chan, J. Liang, P. Florence, A. Zeng, J. Tompson, I. Mordatch, Y. Chebotar, *et al.*, "Inner monologue: Embodied reasoning through planning with language models," *arXiv preprint arXiv:2207.05608*, 2022.

[12] K. Rana, J. Haviland, S. Garg, J. Abou-Chakra, I. D. Reid, and N. Suenderhauf, "Sayplan: Grounding large language models using 3d scene graphs for scalable task planning," *CoRR*, 2023.

[13] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, *et al.*, "Palm-e: An embodied multimodal language model," pp. 8469–8488, 2023.

[14] X. B. Peng, E. Coumans, T. Zhang, T.-W. E. Lee, J. Tan, and S. Levine, "Learning agile robotic locomotion skills by imitating animals," in *Robotics: Science and Systems*, 07 2020.

[15] I. Radosavovic, T. Xiao, B. Zhang, T. Darrell, J. Malik, and K. Sreenath, "Real-world humanoid locomotion with reinforcement learning," *Science Robotics*, vol. 9, no. 89, p. eadi9579, 2024.

[16] R. Sinha, A. Elhafsi, C. Agia, M. Foutter, E. Schmerling, and M. Pavone, "Real-Time Anomaly Detection and Reactive Planning with Large Language Models," in *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, 7 2024.

[17] Z. Liu, A. Bahety, and S. Song, "Reflect: Summarizing robot experiences for failure explanation and correction," *arXiv preprint arXiv:2306.15724*, 2023.

[18] A. S. Morgan, B. Wen, J. Liang, A. Boularias, A. M. Dollar, and K. Bekris, "Vision-driven compliant manipulation for reliable, high-precision assembly tasks," *arXiv preprint arXiv:2106.14070*, 2021.

[19] Z. Zhao, Y. Li, W. Li, Z. Qi, L. Ruan, Y. Zhu, and K. Althoefer, "Tac-man: Tactile-informed prior-free manipulation of articulated objects," *IEEE Transactions on Robotics*, vol. 41, pp. 538–557, 2024.

[20] G. J. Garcia, J. A. Corrales, J. Pomares, and F. Torres, "Survey of visual and force/tactile control of robots for physical interaction in spain," *Sensors*, vol. 9, no. 12, pp. 9689–9733, 2009.

[21] A. O'Neill, A. Rehman, A. Maddukuri, A. Gupta, A. Padalkar, A. Lee, A. Pooley, A. Gupta, A. Mandlekar, A. Jain, *et al.*, "Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 6892–6903.

[22] N. Muhammad, A. Rai, H. Etukuru, Y. Liu, I. Misra, S. Chintala, and L. Pinto, "On bringing robots home." *CoRR*, 2023.

[23] A. Khazatsky, K. Pertsch, S. Nair, A. Balakrishna, S. Dasari, S. Karamcheti, S. Nasiriany, M. K. Srirama, L. Y. Chen, K. Ellis, *et al.*, "Droid: A large-scale in-the-wild robot manipulation dataset," *arXiv preprint arXiv:2403.12945*, 2024.

[24] H.-S. Fang, H. Fang, Z. Tang, J. Liu, C. Wang, J. Wang, H. Zhu, and C. Lu, "Rh20t: A comprehensive robotic dataset for learning diverse skills in one-shot," *arXiv preprint arXiv:2307.00595*, 2023.

[25] Y. Wu, Z. Chen, F. Wu, L. Chen, L. Zhang, Z. Bing, A. Swikir, A. Knoll, and S. Haddadin, "Tacdiffusion: Force-domain diffusion policy for precise tactile manipulation," *arXiv preprint arXiv:2409.11047*, 2024.

[26] J. Huang, S. Wang, F. Lin, Y. Hu, C. Wen, and Y. Gao, "Tactile-vla: unlocking vision-language-action model's physical knowledge for tactile generalization," *arXiv preprint arXiv:2507.09160*, 2025.

[27] C. Zhang, P. Hao, X. Cao, X. Hao, S. Cui, and S. Wang, "Vtla: Vision-tactile-language-action model with preference learning for insertion manipulation," *arXiv preprint arXiv:2505.09577*, 2025.

[28] Y. Zhang, Y. Wang, X. Sun, K. Huang, Z. Xu, J. Ji, Z. Che, J. Tang, and J. Sun, "Craft: Adapting vla models to contact-rich manipulation via force-aware curriculum fine-tuning," *arXiv preprint arXiv:2602.12532*, 2026.

[29] D. Sliwowski, S. Jadav, S. Stanovcic, J. Orbik, J. Heidersberger, and D. Lee, "Reassemble: A multimodal dataset for contact-rich robotic assembly and disassembly," *arXiv preprint arXiv:2502.05086*, 2025.

[30] J. Yu, H. Liu, Q. Yu, J. Ren, C. Hao, H. Ding, G. Huang, G. Huang, Y. Song, P. Cai, *et al.*, "Forcevla: Enhancing vla models with a force-aware moe for contact-rich manipulation," *arXiv preprint arXiv:2505.22159*, 2025.

[31] E. Jang, A. Irpan, M. Khansari, D. Kappler, F. Ebert, C. Lynch, S. Levine, and C. Finn, "Bc-z: Zero-shot task generalization with robotic imitation learning," in *corl*. PMLR, 2022, pp. 991–1002.

[32] H. R. Walke, K. Black, T. Z. Zhao, Q. Vuong, C. Zheng, P. Hansen-Estruch, A. W. He, V. Myers, M. J. Kim, M. Du, *et al.*, "Bridgedata v2: A dataset for robot learning at scale," in *Conference on Robot Learning*. PMLR, 2023, pp. 1723–1736.

[33] M. Heo, Y. Lee, D. Lee, and J. J. Lim, "Furniturebench: Reproducible real-world benchmark for long-horizon complex manipulation," *The International Journal of Robotics Research*, vol. 44, no. 10-11, pp. 1863–1891, 2025.

[34] F. Sener, D. Chatterjee, D. Shelepov, K. He, D. Singhania, R. Wang, and A. Yao, "Assembly101: A large-scale multi-view video dataset for understanding procedural activities," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 21 096–21 106.

[35] A. Inceoglu, E. E. Aksoy, A. C. Ak, and S. Sariel, "Fino-net: A deep multimodal sensor fusion framework for manipulation failure detection," in *2021 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2021, pp. 6841–6847.

[36] I. Lenz, H. Lee, and A. Saxena, "Deep learning for detecting robotic grasps," *The International Journal of Robotics Research*, vol. 34, no. 4-5, pp. 705–724, 2015.

[37] J. Mahler, J. Liang, S. Niyaz, M. Laskey, R. Doan, X. Liu, J. A. Ojea, and K. Goldberg, "Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics," *arXiv preprint arXiv:1703.09312*, 2017.

[38] S. Dasari, F. Ebert, S. Tian, S. Nair, B. Bucher, K. Schmeckpeper, S. Singh, S. Levine, and C. Finn, "Robonet: Large-scale multi-robot learning," *arXiv preprint arXiv:1910.11215*, 2019.

[39] K. Kimble, J. Albrecht, M. Zimmerman, and J. Falco, "Performance measures to benchmark the grasping, manipulation, and assembly of deformable objects typical to manufacturing applications," *Frontiers in Robotics and AI*, vol. 9, p. 999348, 2022.

[40] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," *arXiv preprint arXiv:2304.13705*, 2023.

[41] R. Cadene, S. Alibert, A. Soare, Q. Gallouedec, A. Zouitine, S. Palma, P. Kooijmans, M. Aractingi, M. Shukor, D. Aubakirova, M. Russi, F. Capuano, C. Pascal, J. Choghari, J. Moss, and T. Wolf, "Lerobot: State-of-the-art machine learning for real-world robotics in pytorch," https://github.com/huggingface/lerobot, 2024.