SoftVTBench: A Safety-Aware Visuo-Tactile Benchmark for Physically Constrained Robotic Manipulation of Deformable Objects

Bowen Jing[1,*], **Mingxin Wang**[1,2,*], **Ruiyang Hao**[3], **Chenchen Ge**[1,4], **Hanwen Shen**[5], **Junjie He**[6], **Yang Cui**[7], **Yiming Hou**[1,4], **Weitao Zhou**[2,8,‡], **Jiawei Wang**[8], **Minglei Li**[8], **Dandan Zhang**[9], **Ding Zhao**[10], **Houde Liu**[2], **Xiaofan Li**[11], **Si Liu**[12], **Ping Luo**[13], **Haibao Yu**[1,13,‡]

[1]Tuojing Intelligence, [2]Tsinghua University, [3]King's College London, [4]Southeast University, [5]Stevens Institute of Technology, [6]The Hong Kong University of Science and Technology (Guangzhou), [7]University of Manchester, [8]Simple AI, [9]Imperial College London, [10]Carnegie Mellon University, [11]Zhejiang University, [12]Beihang University, [13]The University of Hong Kong

[*]equal contribution, [‡]corresponding author

Deformable object manipulation poses challenges beyond task completion: successful execution must also maintain safe physical interaction, holding the object stably without slip or drop while avoiding excessive deformation. However, existing manipulation benchmarks are predominantly success-oriented and rarely evaluate whether a policy remains physically safe throughout execution. We present SoftVTBench, a safety-aware visuo-tactile benchmark for physically constrained deformable object manipulation. Built in Isaac Sim with finite-element-simulated deformable objects, SoftVTBench provides multi-view RGB observations, RGB tactile sensing with marker motion, proprioception, and language instructions, and defines four matched task suites over object type (deformable vs. rigid) and variation axis (object vs. spatial). It separately reports *Goal Success* and *Safety Success*; the latter additionally requires no drop and peak deformation below a calibrated object-specific threshold, measured from policy-hidden privileged Finite Element Method (FEM) states. We implement $\pi_{0.5}$-based baselines under this protocol. Experiments show that success-only evaluation substantially overstates policy performance, as a large fraction of goal-completing rollouts still violate physical safety. Furthermore, incorporating tactile sensing improves Safety Success (e.g., from 21.4% to 35.6% on object-centric deformable tasks) and reduces object deformation during execution, while maintaining comparable Goal Success. SoftVTBench provides a reproducible benchmark for studying visuo-tactile deformable manipulation under physical interaction constraints.

Code: https://github.com/TuojingAI/SoftVTBench
Website: https://softvtbench.github.io/

1 Introduction

Recent progress in large-scale robot learning and foundation-model-based policies has significantly improved the generality of robotic manipulation across tasks, objects, and environments Walke et al. (2023); Kim et al. (2024); Black et al. (2024); Bjorck et al. (2025); Jang et al. (2025); Ji et al. (2025). However, most existing approaches are primarily evaluated in rigid-object settings Liu et al. (2023); Gu et al. (2023); Nasiriany et al. (2024); Mu et al. (2025), where performance is measured by goal achievement and the physical interaction process is largely abstracted away. Although deformable object manipulation has recently received increasing attention Zhao et al. (2025); Moletta et al. (2026); Moghani et al. (2026), existing evaluation protocols remain largely success-oriented, assessing performance primarily based on task completion Huang et al. (2021); Zhang et al. (2025c). Unlike rigid-object manipulation, deformable object manipulation inherently involves contact-rich dynamics and material-dependent constraints Sun et al. (2025), in which successful execution requires not only accomplishing the task but also maintaining physically appropriate interactions, such as holding the object stably without slip or drop and avoiding excessive deformation or damage caused by improper grasping forces. Therefore, evaluating deformable object manipulation requires going beyond success-oriented metrics to explicitly assess interaction-level physical constraints throughout execution.

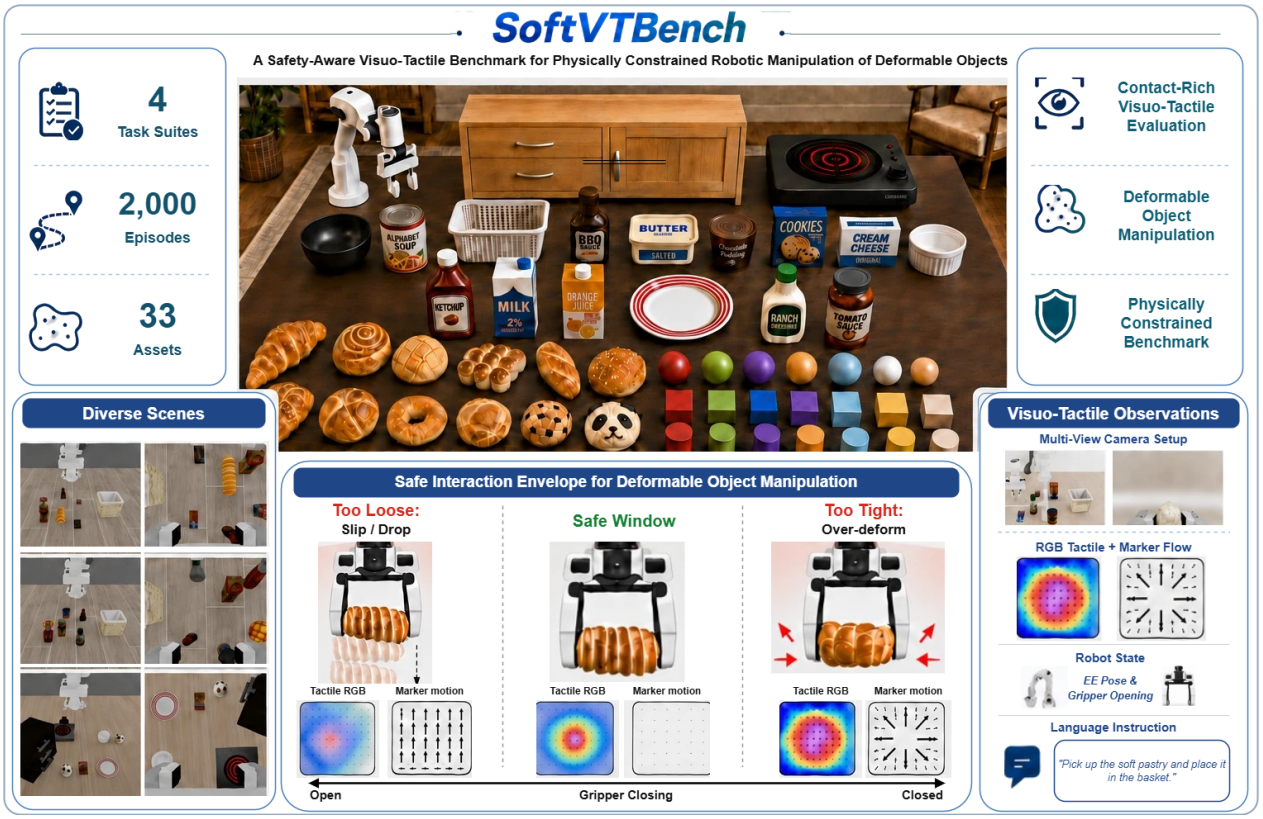

Figure 1 Overview of SoftVTBench. SoftVTBench is a safety-aware visuo-tactile benchmark for physically constrained deformable-object manipulation, comprising four task suites, 2,000 episodes, 33 assets, and diverse tabletop scenes. Each episode provides multi-view RGB observations, dual-finger tactile observations with RGB images and marker-flow signals, robot proprioception, and language instructions. The safe-interaction envelope illustrates the core premise of SoftVTBench: unlike success-only benchmarks that only check whether a task is completed, SoftVTBench evaluates whether completion is achieved under physically safe contact. A grasp that is too loose may cause slip or drop, whereas an overly tight grasp may complete the placement task but over-deform the object; safe manipulation therefore lies in the intermediate regime that preserves grasp stability while limiting deformation.

Evaluating such physical constraints requires robots to perceive fine-grained contact dynamics during manipulation. While vision provides global information about object geometry and pose, it fails to capture critical interaction states at the gripper–object interface, including contact pressure distribution, incipient slip, and local deformation Huang et al. (2024); Zheng et al. (2026). These signals are essential for maintaining grasp stability and physical safety during execution. Hence, vision-based policies often struggle to regulate grasping forces reliably in contact-rich scenarios, especially under varying object compliance and frictional uncertainty He et al. (2025). Tactile sensing complements vision by providing direct feedback on local physical interactions, enabling more accurate perception of contact dynamics and more reliable regulation of manipulation behavior Suresh et al. (2024); Fan et al. (2024); Liu et al. (2026). Consequently, a benchmark for physically constrained deformable object manipulation should support visuo-tactile observations, enabling evaluation beyond vision-only perception.

Despite rapid progress in robotic manipulation, tactile sensing, and deformable object manipulation, existing benchmarks typically address only part of the capabilities required for physically constrained deformable manipulation. Existing robotic manipulation benchmarks such as LIBERO Liu et al. (2023), RoboTwin Mu et al. (2025), RoboCasa Nasiriany et al. (2024), and SIMPLER Li et al. (2024) provide standardized evaluation for general manipulation policies, but primarily focus on rigid-object tasks with success-oriented metrics. In parallel, tactile and visuo-tactile benchmarks such as UniVTAC Chen et al. (2026), TacO Zorin et al. (2026), and Tabero Wu et al. (2026) investigate tactile perception and contact-rich manipulation, but do not specifically target deformation-aware evaluation for deformable objects. Existing deformable object

manipulation benchmarks Lin et al. (2020); Zhang et al. (2025c); Greenland et al. (2024) focus on learning and evaluating manipulation in deformable environments or assessing grasp-induced deformation, rather than systematically evaluating policy behavior under deformation-bounded physical interaction constraints. As a result, the intersection of deformable object manipulation, visuo-tactile perception, and physically constrained evaluation remains underexplored.

To address this gap, we present SoftVTBench, a safety-aware visuo-tactile benchmark for physically constrained deformable object manipulation. Figure 1 provides an overview of SoftVTBench, including its benchmark scale, diverse scenes and assets, multimodal visuo-tactile observation interface, and safe interaction envelope for deformable-object manipulation. SoftVTBench provides a unified closed-loop evaluation environment with realistic deformable assets, multi-view visual observations, dual-finger RGB tactile sensing, proprioception, and language instructions, together with nearly 2,000 collected episodes for visuo-tactile deformable manipulation. Its object-centric and spatial tasks are paired with matched rigid-control suites, separating basic manipulation competence from safe deformable-object interaction. Unlike success-only evaluation, SoftVTBench evaluates physical safety using policy-hidden privileged Finite Element Method (FEM) Zienkiewicz et al. (1977) states to detect drop/slip events and identify excessive deformation beyond calibrated object-specific thresholds. Based on this benchmark, we implement $\pi_{0.5}$-based baseline policies Physical Intelligence et al. (2025) and compare observation modalities under a shared physical and task setup. Experiments show that success-only evaluation substantially overstates policy performance, as a large fraction of goal-completing rollouts violate physical safety, and that adding tactile sensing improves Safety Success while keeping Goal Success comparable and reduces object deformation during execution.

In summary, this work makes three contributions:

- We formulate safety-aware evaluation for deformable object manipulation, in which a policy must complete the task while keeping the object stably grasped and its deformation below a calibrated, object-specific threshold, exposing unsafe completions that are hidden by success-only evaluation.
- We introduce a visuo-tactile Isaac Sim benchmark with FEM-simulated deformable assets, multi-view visual and RGB tactile observations, object-centric and spatial task suites paired with matched rigid-control suites, and privileged-state safety evaluation.
- We implement $\pi_{0.5}$-based baselines and provide systematic evaluation, showing that success-only metrics substantially overstate policy performance and that tactile sensing improves Safety Success and reduces excessive deformation.

2 Related Work

Robotic Manipulation Benchmarks Manipulation benchmarks have established standard protocols for measuring progress in robot learning. Meta-World Yu et al. (2020) and RLBench James et al. (2020) introduced multi-task and language-conditioned manipulation suites, LIBERO Liu et al. (2023) evaluates lifelong knowledge transfer across 130 language-conditioned tasks, and CALVIN Mees et al. (2022) focuses on long-horizon language-conditioned control. Recent benchmarks further scale task, scene, and embodiment diversity: RoboCasa Nasiriany et al. (2024) uses procedural generation for household manipulation, RoboTwin Mu et al. (2025) and RoboTwin 2.0 Chen et al. (2025) target bimanual manipulation with synthetic data generation and domain randomization, THE COLOSSEUM Pumacay et al. (2024) stress-tests robustness under visual and physical perturbations, and SIMPLER Li et al. (2024) provides simulation-based evaluation whose policy rankings are predictive of real-world performance. These benchmarks are central to evaluating spatial competence, generalization, and instruction following, but their success predicates are primarily terminal and kinematic, such as whether an object reaches a target state or a subtask is completed. Recent safety-oriented benchmarks have begun to evaluate safety alongside task success Ji et al. (2023); Zhang et al. (2025a); Fan et al. (2026); Huang et al. (2026), but they mainly target constraint violations, semantic hazards, or temporal-logic properties for rigid-object or general policies. In contrast, SoftVTBench focuses on process-level physical safety induced by contact with deformable objects, where task completion alone does not characterize the quality of interaction during execution. Table 1 summarizes the positioning of SoftVTBench relative to representative manipulation, deformable-object, and visuo-tactile benchmarks.

Table 1 Comparison with representative manipulation, deformable-object, and visuo-tactile benchmarks. ✓: direct support; ❍: partial support; ✗: not supported. *Tactile Observation*: tactile signals in the policy observation; *Deformable Objects*: 3D deformable target objects; *Deformation GT*: deformation measured from physical simulation ground truth; *Safety Metric*: evaluation beyond goal completion; *Rigid–Soft Controls*: matched rigid/deformable task suites.

Benchmark	Tactile Observation	Deformable Objects	Deformation GT	Safety Metric	Rigid–Soft Controls
LIBERO Liu et al. (2023)	✗	✗	✗	✗	✗
SoftGym Lin et al. (2020)	✗	❍	❍	✗	✗
MoDeSuite Zhang et al. (2025c)	✗	✓	❍	✗	✗
SoGraB Greenland et al. (2024)	✗	✓	✓	❍	✗
UniVTAC Chen et al. (2026)	✓	❍	✗	✗	✗
ManiFeel Luu et al. (2025)	✓	✗	✗	✗	✗
Tabero Wu et al. (2026)	✓	❍	✗	❍	✗
SoftVTBench	✓	✓	✓	✓	✓

Deformable-Object Manipulation Deformable-object manipulation has been studied across simulation environments, task suites, and learning methods, where object states are high-dimensional and governed by complex dynamics Zhu et al. (2022). SoftGym Lin et al. (2020) provides reinforcement-learning tasks for cloth, ropes, and other deformable objects, while PlasticineLab Huang et al. (2021) and DaXBench Chen et al. (2022) use differentiable soft-body simulation to support learning and optimization. DEDO Antonova et al. (2021) and GarmentLab Lu et al. (2024) extend this line with dynamic cloth and garment tasks, ManiSkill2 Gu et al. (2023) includes soft-body tasks in a broader manipulation suite, MoDeSuite Zhang et al. (2025c) targets mobile manipulation with deformable objects, and recent foundation-model work such as DeMaVLA Su et al. (2026) reflects growing interest in scalable policies for deformable manipulation. Many of these settings focus on thin-shell or low-dimensional deformables such as cloth, rope, and plasticine, where deformation is often part of the task state or objective. SoftVTBench addresses a complementary regime in which deformation is not the goal, but a physical quantity that must remain bounded while another manipulation objective is completed. This distinction is important for everyday manipulation of volumetric soft objects such as food, packaging, and soft containers, where the robot should grasp and transport the object without changing its physical condition more than necessary Zhu et al. (2022). The closest prior work measures deformation from simulation ground truth: SoGraB Greenland et al. (2024) scores gripper-induced deformation during soft grasping, while DefGraspSim Huang et al. (2022) and DefGraspNets Huang et al. (2023) evaluate FEM stress and deformation of candidate grasps on 3D deformable objects. These methods provide valuable deformation-aware grasp evaluation, whereas SoftVTBench evaluates policy behavior under deformation-bounded interaction constraints throughout a full manipulation task with visuo-tactile observations.

Visuo-Tactile Sensing and Policy Learning Tactile sensing provides local physical information that vision alone cannot observe, including contact geometry, shear, slip, and compression Yuan et al. (2017); Si & Yuan (2022); Suresh et al. (2024). GelSight Yuan et al. (2017) introduced high-resolution optical tactile sensing, and simulation tools such as Taxim Si & Yuan (2022), FOTS Zhao et al. (2024), TacEx Nguyen et al. (2024), TacSL Akinola et al. (2025), and DiffTactile Si et al. (2024) make tactile observations more accessible for learning-based manipulation. Recent policy and benchmark work increasingly treats tactile feedback as a first-class observation stream: VTLA Zhang et al. (2025b), OmniVTLA Cheng et al. (2025), VLA-Touch Bi et al. (2025), and Tactile-VLA Huang et al. (2025) incorporate tactile signals into vision-language-action policies; ManiFeel Luu et al. (2025) benchmarks visuo-tactile manipulation policy learning; UniVTAC Chen et al. (2026) provides a unified simulation platform for visuo-tactile data generation, representation learning, and benchmarking across contact-rich tasks; TacO Zorin et al. (2026) compares tactile sensor modalities under a task-driven imitation-learning protocol; and AT-VLA Li et al. (2026) studies adaptive tactile injection and fast tactile reaction in VLA models. These works show the value of tactile feedback for contact-rich manipulation. However, they primarily study rigid-object or general contact-rich settings, leaving open whether tactile feedback helps policies satisfy deformation-bounded safety constraints when the true object deformation state is hidden. SoftVTBench targets this question directly by evaluating whether visual and tactile feedback enable policies to avoid unsafe over-compression, slip, and drop while still completing manipulation goals.

3 SoftVTBench

3.1 Benchmark Design

SoftVTBench is a safety-aware visuo-tactile benchmark for physically constrained deformable object manipulation. It evaluates closed-loop robot policies under two coupled requirements: completing the manipulation goal and maintaining safe physical interaction throughout execution. A rollout is considered physically safe only if the object remains stably grasped without slip or drop, and its peak deformation stays below a calibrated, object-specific threshold. By reporting *Goal Success* and *Safety Success* separately, SoftVTBench exposes goal-complete but physically unsafe rollouts that are hidden by success-only evaluation.

SoftVTBench instantiates this evaluation protocol in Isaac Sim with simulated deformable objects based on finite element method soft-body dynamics and a Franka Panda arm with a parallel-jaw gripper carrying GelSight Mini tactile sensors on both fingers, as illustrated in Figure 2. The benchmark provides synchronized third-person and wrist RGB observations, tactile RGB images, marker-motion fields, proprioception, and a standardized end-effector and gripper action interface. It contains object-centric and spatial deformable manipulation suites, together with matched rigid-control suites that separate deformable-object safety from basic manipulation competence and robustness to spatial variation.

At each control step, the policy receives only its designated observations and outputs an end-effector command and a gripper command. The simulator advances the robot, object, and contact states, producing the next observations in a closed-loop interaction process. In parallel, SoftVTBench records privileged physical states, including FEM nodal positions, object poses, contact status, and drop events. These states are used only by the evaluator, requiring policies to infer contact conditions, incipient slip, and material compliance from observable visual, tactile, and proprioceptive inputs.

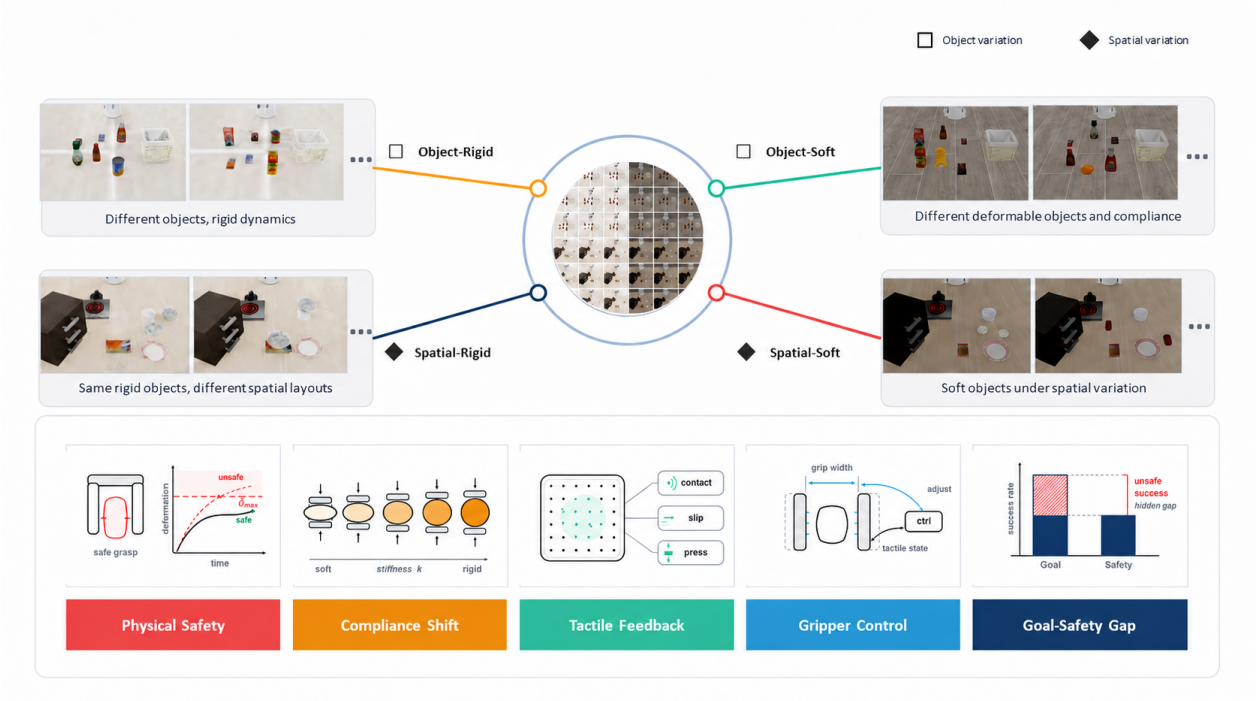

Figure 2 SoftVTBench organizes evaluation into four matched task suites defined by object type and variation type: Object-Soft, Spatial-Soft, Object-Rigid, and Spatial-Rigid. The benchmark studies grasp-and-place manipulation under a hidden safe interaction envelope, where the robot must maintain stable contact without slip or drop while keeping peak FEM deformation below a calibrated object-specific threshold. The bottom panels summarize the key factors considered by SoftVTBench, including physical deformation safety, compliance shift, tactile feedback, gripper control, and the *Goal-Safety Gap*, where goal-completing rollouts can still be physically unsafe under success-only evaluation.

3.2 Observation and Action Space

SoftVTBench provides a standardized multimodal observation-action interface for closed-loop policy evaluation. At each control step, the policy receives visual observations, tactile observations, proprioceptive states, and a per-task language instruction, and outputs an end-effector command together with a gripper command.

The visual observation consists of two RGB views: a third-person camera that captures the tabletop layout, target region, manipulated object, and distractors, and a wrist-mounted camera that provides close-range views during approach, grasping, and transport. The tactile observation is obtained from simulated GelSight Mini sensors Yuan et al. (2017) mounted on both gripper fingers. Each tactile sensor provides an RGB tactile image and a marker-motion field, which capture local contact geometry, elastomer deformation, and shear motion at the gripper–object interface. These tactile signals provide interaction-state information that is often occluded or ambiguous in external visual observations.

The proprioceptive state consists of the end-effector pose, arm joint states, and current gripper width. Each task is additionally specified by a natural-language instruction identifying the target object and goal region. All observation streams are synchronized at 20 Hz. Formally, the policy-visible observation can be written as

$$o_t = \{I_t^{\text{third}}, I_t^{\text{wrist}}, \mathcal{T}_t, p_t, \ell\}, \tag{1}$$

where I denotes the third-person and wrist RGB images, $\mathcal{T}_t$ the tactile observation from both fingers—comprising the tactile RGB images and marker-motion fields of the left and right GelSight sensors—p_t the proprioceptive state, and ℓ the task instruction.

The action space contains an absolute end-effector pose target and a scalar gripper command:

$$a_t = (\mathbf{x}_t^{ee}, \boldsymbol{\theta}_t^{ee}, g_t), \tag{2}$$

where $\mathbf{x}_t^{ee}$ is the 3D end-effector position, $\boldsymbol{\theta}_t^{ee}$ is the axis-angle orientation, and g_t denotes the gripper command. SoftVTBench supports both binary open/close commands and continuous closure targets, with aligned trajectory encodings for controlled ablation of gripper-action granularity. Detailed sensor resolutions, rendering pipelines, synchronization rates, and action parameterization are provided in Appendix A.

3.3 Task Suites

SoftVTBench defines four task suites organized along two axes: object type and task variation, as summarized in Table 2. The object-type axis contrasts deformable objects with matched rigid-control objects, while the task-variation axis contrasts object-centric variation with spatial variation. This design lets us separate deformable-object safety from basic manipulation competence and robustness to spatial variation.

The deformable suites use FEM-simulated objects, including bakery-style objects and geometric primitives. These objects vary in shape, size, and appearance, producing object-specific contact responses, manipulation difficulty, and safety thresholds. Because the safe interaction range is not directly observable from vision alone, policies must infer safe manipulation behavior from visual, tactile, and proprioceptive feedback. Detailed asset statistics and material parameters are provided in Appendix C.

Table 2 Overview of the task suites in SoftVTBench. The benchmark comprises four suites spanning deformable and rigid objects under both object-level and spatial variations, enabling evaluation of manipulation performance and physical safety across different sources of task diversity.

Suite	Object Type	Variation	Purpose
Object-Soft	Deformable	Object	Safe deformable manipulation
Spatial-Soft	Deformable	Spatial	Safe manipulation under layout variation
Object-Rigid	Rigid	Object	Basic manipulation control
Spatial-Rigid	Rigid	Spatial	Spatial-variation control

In Object-Soft, the robot must grasp a deformable object and place it into a target container without dropping or excessively deforming it. The scene layout and target region are fixed, while the manipulated

object varies across tasks. This suite evaluates whether a policy can adapt its interaction to object-level differences in geometry and contact response.

In SPATIAL-SOFT, the robot performs the same grasp-and-place objective under spatial variation. Each scene contains two visually identical instances of the same deformable object, and the language instruction specifies which instance to manipulate; tasks come in mirrored pairs that share the same physical layout. The policy must therefore ground the instruction to the correct instance and complete the transfer safely under changing object and target placements.

The two rigid-control suites, OBJECT-RIGID and SPATIAL-RIGID, follow the same object-centric and spatial task structures using rigid LIBERO-style objects Liu et al. (2023). These suites serve as diagnostic controls: they test whether a policy has basic manipulation and spatial-variation capability before the additional safety constraints introduced by deformable objects are evaluated.

3.4 Evaluation Protocol and Metrics

SoftVTBench evaluates policies through closed-loop rollouts in the simulator. For each task suite, all methods are evaluated on the same evaluation episodes with fixed initial states and fixed seeds. During execution, the policy receives only its designated observations and outputs robot actions. Privileged simulator states, including object pose, contact status, drop events, and FEM deformation, are recorded only for evaluation and are never exposed to the policy.

We report two main metrics: *Goal Success* and *Safety Success*. Goal Success measures whether the task objective is completed. In our grasp-and-place tasks, a rollout is counted as goal-successful if the target object is placed inside the target region or container and remains there for a short terminal horizon.

Safety Success further requires the task to be completed without unsafe physical interaction. For deformable-object suites, the evaluator computes object deformation at each timestep from the hidden FEM state and records the peak deformation over the full rollout:

$$D_{\text{peak}} = \max_t D(t), \tag{3}$$

where $D(t)$ is the object-size-normalized FEM-RMS deformation after removing global rigid-body motion, reported as a percentage of the object bounding-box diagonal.

A deformable-object rollout is counted as safety-successful if

$$\text{Safety Success} = \text{Goal Success} \ \wedge \ \text{NoDrop}_{\text{episode}} \ \wedge \ (D_{\text{peak}} \leq \tau_o)\,, \tag{4}$$

where τ_o is the calibrated, object-specific safety threshold. $\text{NoDrop}_{\text{episode}}$ requires the object to remain under stable manipulation throughout the episode; that is, it is violated by any transient drop, workspace escape, or loss of stable containment. For the rigid-control suites, the deformation term is inactive, and Safety Success reduces to Goal Success under the NoDrop condition.

For a set of N evaluation episodes, each metric is reported as its rate over the set:

$$\text{Goal Success Rate} = \frac{1}{N}\sum_{i=1}^{N} \text{Goal Success}^{(i)}, \qquad \text{Safety Success Rate} = \frac{1}{N}\sum_{i=1}^{N} \text{Safety Success}^{(i)}. \tag{5}$$

Safety Success is therefore stricter than Goal Success. A rollout may complete the task while still violating physical safety through transient drop, unstable contact, or excessive deformation. The gap between Goal Success and Safety Success captures unsafe task completions that are hidden by success-only evaluation.

4 Experiments

We evaluate SoftVTBench on physically constrained deformable object manipulation tasks. Our goal is not only to measure whether a policy completes the task, but also whether it completes the task while maintaining safe physical interaction with the object. We compare policies across object-centric and spatial task settings, and report both task-level success and safety-aware success.

4.1 Baselines

We compare two $\pi_{0.5}$-based policy settings.

$\pi_{0.5}$-Vision Physical Intelligence et al. (2025). The vision-only policy takes as input a third-person RGB image, a wrist RGB image, robot proprioceptive state, and the language instruction. The gripper command is represented as a binary open-close action. This design avoids providing the vision-only policy with continuous gripper-width bounds obtained from contact calibration, which would otherwise introduce implicit interaction information beyond visual observations.

$\pi_{0.5}$-Visuo-Tactile. The visuo-tactile policy uses the same visual and proprioceptive inputs as $\pi_{0.5}$-Vision, and additionally receives tactile observations from both gripper fingers. For tactile RGB observations, we use an 8-frame history from each finger and concatenate the left and right tactile histories into a single 4×4 grid image. This tactile image is encoded by the same visual encoder used for RGB observations. We also include tactile marker motion as a low-dimensional contact signal: marker motion from both fingers is concatenated and provided as a history sequence, allowing the policy to observe recent local contact deformation and shear. The gripper command is represented as a continuous gripper-width action.

Both policies output a 7D action consisting of 3D end-effector position, 3D orientation, and 1D gripper command. We train all policies by LoRA fine-tuning $\pi_{0.5}$ with an action horizon of 50, using 8 NVIDIA A100 GPUs, a global batch size of 256, and 7k training steps by default. During evaluation, the policy predicts an action chunk and executes 10 steps before replanning.

4.2 Main Results

Table 3 Main results on SoftVTBench. VO denotes the vision-only $\pi_{0.5}$ policy, and VT denotes the visuo-tactile $\pi_{0.5}$ policy. Goal Success measures task completion, while Safety Success further requires the rollout to satisfy physical safety constraints.

Suite	Method	Goal Success	Safety Success
Object-rigid	VO	38.8%	–
Object-rigid	VT	32.4%	–
Spatial-rigid	VO	56.4%	–
Spatial-rigid	VT	63.4%	–
Object-soft	VO	70.4%	21.4%
Object-soft	VT	71.8%	35.6%
Spatial-soft	VO	74.2%	32.6%
Spatial-soft	VT	84.2%	44.6%

From Table 3, tactile information does not consistently improve performance on rigid-object tasks. On Object-rigid, VT achieves 32.4% Goal Success, lower than VO at 38.8%; on Spatial-rigid, VT achieves 63.4%, higher than VO at 56.4%. This suggests that, for rigid-object manipulation, visual observations and proprioception already provide the primary information needed for task completion. When deformation-related safety constraints are absent, adding tactile RGB and marker motion provides limited marginal benefit and may introduce additional multimodal noise or over-reliance on local contact signals. Therefore, the rigid-task results do not indicate that tactile sensing is ineffective; rather, they suggest that its main value is not in improving goal success for ordinary rigid manipulation.

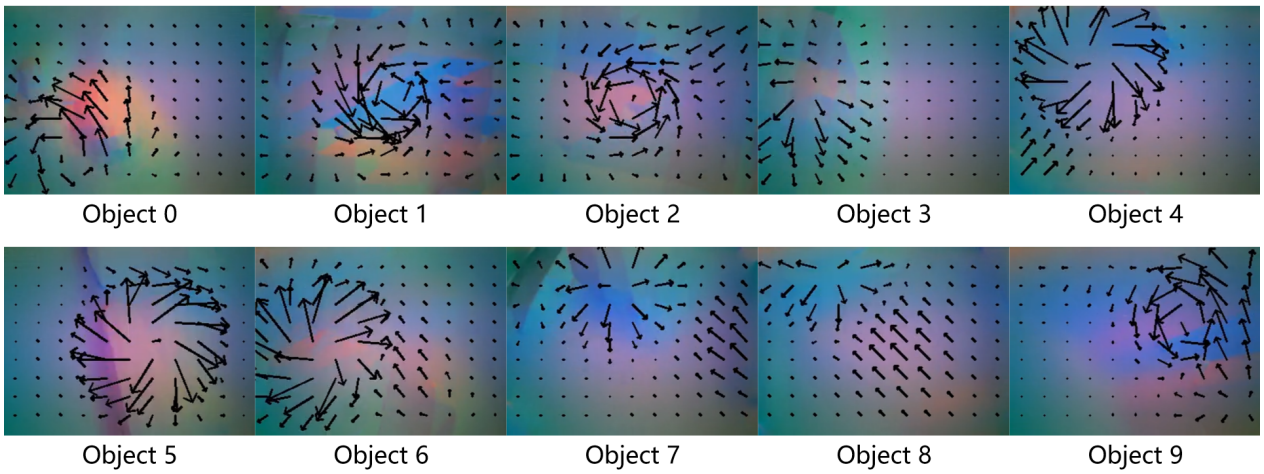

Figure 3 Representative tactile RGB observations from the ten Object-Soft assets in SoftVTBench. Each panel shows a tactile observation with marker-motion overlay during grasping. Different deformable objects induce diverse local contact and shear patterns, reflecting their distinct geometry, compliance, and contact response.

In contrast, VT shows clearer advantages on deformable-object tasks. On Object-Soft, VO and VT achieve similar Goal Success, 70.4% and 71.8%, while VT improves Safety Success from 21.4% to 35.6%. On Spatial-soft, VT improves both Goal Success and Safety Success, with Safety Success increasing from 32.6% to 44.6%. These results indicate that tactile sensing is not merely helping the object reach the target region, but is instead improving contact regulation during execution. By using tactile RGB and marker motion, VT reduces unsafe goal-completing behaviors such as slippage, over-compression, and unstable grasping. Thus, for deformable-object manipulation, Safety Success provides a more faithful measure of policy quality than Goal Success alone.

Overall, these results show that SoftVTBench distinguishes between merely completing a task and completing it safely. Tactile sensing does not universally improve Goal Success across all tasks; its benefit is most pronounced in deformable-object manipulation, where safe physical interaction depends on regulating local contact during execution.

Figure 3 provides a qualitative view of why tactile observations are informative for deformable-object manipulation. Different Object-Soft assets produce distinct tactile RGB and marker-motion patterns during grasping, reflecting object-specific geometry, compliance, and local contact response. These interaction cues are difficult to infer from external RGB observations alone, but are directly captured by tactile sensing at the gripper-object interface.

4.3 Deformation Analysis

Table 4 FEM-RMS deformation distribution over all deformable-object rollouts in SoftVTBench, reported as percentages of the object bounding-box diagonal. We report the mean, the 5th and 95th percentiles (P5, P95), and the median; lower values indicate safer physical interaction. Across both suites, the visuo-tactile policy (VT) shifts the entire distribution downward relative to the vision-only policy (VO)—including the P95 tail that corresponds to severe over-compression—indicating fewer and milder unsafe contacts rather than a change confined to the mean.

Suite	Method	Mean	P5	Median	P95
Object-soft	VO	16.10%	4.30%	10.65%	44.70%
Object-soft	VT	15.12%	3.90%	8.67%	38.81%
Spatial-soft	VO	13.16%	5.16%	10.75%	28.96%
Spatial-soft	VT	11.58%	4.75%	9.67%	26.56%

The deformation statistics in Table 4 further explain why VT achieves higher Safety Success on deformable-

object tasks. On Object-Soft, VT reduces the mean deformation from 16.10% to 15.12%, the median from 10.65% to 8.67%, and P95 from 44.70% to 38.81%. On Spatial-Soft, VT similarly lowers the mean deformation from 13.16% to 11.58%, the median from 10.75% to 9.67%, and P95 from 28.96% to 26.56%.

These results show that tactile sensing improves safety not only by increasing the threshold-based Safety Success metric, but also by shifting the continuous deformation distribution toward safer interactions. The reduction in median deformation indicates that VT improves typical rollout quality, while the reduction in P95 shows that tactile feedback also suppresses high-deformation tail cases, such as severe compression or unstable contact.

Together with the main results in Table 3, this deformation analysis supports the central conclusion of SoftVTBench: tactile sensing is most valuable when policies must regulate local physical interaction with deformable objects. VT may not always produce large gains in Goal Success, but it leads to safer manipulation by reducing deformation and avoiding unsafe yet goal-completing behaviors.

4.4 Goal Success vs. Safety Success

Figure 4 compares Goal Success and Safety Success across task suites and policy settings. The gap between the two metrics captures unsafe goal completions: rollouts where the object reaches the target but the interaction violates safety constraints.

Figure 4 Goal Success and Safety Success across rigid and deformable task suites. Goal Success measures task completion, while Safety Success further requires satisfying physical safety constraints. For rigid suites, deformation-based safety constraints are not applicable and are marked as N/A. For soft suites, the gap between Goal Success and Safety Success indicates unsafe goal completions caused by excessive deformation, dropping, or unstable contact.

This comparison highlights why success-only evaluation is insufficient for deformable object manipulation. A policy must not only achieve the target state, but also maintain appropriate contact throughout the rollout. Safety Success therefore provides a stricter and more physically meaningful measure of policy performance.

5 Conclusion

We presented SoftVTBench, a safety-aware visuo-tactile benchmark for deformable object manipulation. By separating Goal Success from Safety Success and measuring FEM-RMS deformation, SoftVTBench reveals unsafe goal-completing behaviors that success-only evaluation misses. Our experiments show that tactile sensing provides limited and inconsistent gains on rigid tasks, but substantially improves safety on deformable tasks by reducing unsafe contact and object deformation.

SoftVTBench still has several limitations. The current benchmark covers a limited set of assets and task variations, and its deformable-object simulation cannot fully capture the complexity of real-world soft-body dynamics. We also evaluate a focused set of $\pi_{0.5}$-based baselines, leaving broader policy comparisons, such as world-action models, for future work. Future extensions will include more diverse deformable assets, more realistic deformation and contact modeling, and additional visuo-tactile and vision-only baselines.

References

Iretiayo Akinola, Jie Xu, Jan Carius, Dieter Fox, and Yashraj Narang. Tacsl: A library for visuotactile sensor simulation and learning. *IEEE Transactions on Robotics*, 41:2645–2661, 2025. doi: 10.1109/TRO.2025.3547267.

Rika Antonova, Peiyang Shi, Hang Yin, Zehang Weng, and Danica Kragic Jensfelt. Dynamic environments with deformable objects. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

Jianxin Bi, Kevin Yuchen Ma, Ce Hao, Mike Zheng Shou, and Harold Soh. Vla-touch: Enhancing vision-language-action models with dual-level tactile feedback. *arXiv preprint arXiv:2507.17294*, 2025.

Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

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

Baijun Chen, Weijie Wan, Tianxing Chen, Xianda Guo, Congsheng Xu, Yuanyang Qi, Haojie Zhang, Longyan Wu, Tianling Xu, Zixuan Li, et al. Univtac: A unified simulation platform for visuo-tactile manipulation data generation, learning, and benchmarking. *arXiv preprint arXiv:2602.10093*, 2026.

Siwei Chen, Yiqing Xu, Cunjun Yu, Linfeng Li, Xiao Ma, Zhongwen Xu, and David Hsu. Daxbench: Benchmarking deformable object manipulation with differentiable physics. *arXiv preprint arXiv:2210.13066*, 2022.

Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, et al. Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088*, 2025.

Zhengxue Cheng, Yiqian Zhang, Anni Tang, Keyu Wang, Wenkang Zhang, Haoyu Li, Hengdi Zhang, and Li Song. Omnivtla: Vision-tactile-language-action models with semantic-aligned tactile sensing. *arXiv preprint arXiv:2508.08706*, 2025.

Jialiang Fan, Weizhe Xu, Oleg Sokolsky, Insup Lee, and Fanxin Kong. Safevla-bench: A benchmark for the success-safety gap in vision-language-action models. *arXiv preprint arXiv:2606.00773*, 2026.

Wen Fan, Haoran Li, Weiyong Si, Shan Luo, Nathan Lepora, and Dandan Zhang. Vitactip: Design and verification of a novel biomimetic physical vision-tactile fusion sensor. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1056–1062. IEEE, 2024.

Benjamin G. Greenland, Josh Pinskier, Xing Wang, Daniel Nguyen, Ge Shi, Tirthankar Bandyopadhyay, Jen Jen Chung, and David Howard. Sograb: A visual method for soft grasping benchmarking and evaluation. *arXiv preprint arXiv:2411.19408*, 2024.

Jiayuan Gu, Fanbo Xiang, Xuanlin Li, Zhan Ling, Xiqiang Liu, Tongzhou Mu, Yihe Tang, Stone Tao, Xinyue Wei, Yunchao Yao, et al. Maniskill2: A unified benchmark for generalizable manipulation skills. *arXiv preprint arXiv:2302.04659*, 2023.

Zihao He, Hongjie Fang, Jingjing Chen, Hao-Shu Fang, and Cewu Lu. Foar: Force-aware reactive policy for contact-rich robotic manipulation. *IEEE Robotics and Automation Letters*, 2025.

Binghao Huang, Yixuan Wang, Xinyi Yang, Yiyue Luo, and Yunzhu Li. 3d-vitac: Learning fine-grained manipulation with visuo-tactile sensing. *arXiv preprint arXiv:2410.24091*, 2024.

Chengyue Huang, Khang Vo Huynh, Sebastian Elbaum, Zsolt Kira, and Lu Feng. Safemanip: A property-driven benchmark for temporal safety evaluation in robotic manipulation. *arXiv preprint arXiv:2605.12386*, 2026.

Isabella Huang, Yashraj Narang, Clemens Eppner, Balakumar Sundaralingam, Miles Macklin, Ruzena Bajcsy, Tucker Hermans, and Dieter Fox. Defgraspsim: Physics-based simulation of grasp outcomes for 3d deformable objects. *IEEE Robotics and Automation Letters*, 7(3):6274–6281, 2022.

Isabella Huang, Yashraj Narang, Ruzena Bajcsy, Fabio Ramos, Tucker Hermans, and Dieter Fox. Defgraspnets: Grasp planning on 3d fields with graph neural nets. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5894–5901. IEEE, 2023.

Jialei Huang, Shuo Wang, Fanqi Lin, Yihang Hu, Chuan Wen, and Yang Gao. Tactile-vla: Unlocking vision-language-action model's physical knowledge for tactile generalization. *arXiv preprint arXiv:2507.09160*, 2025.

Zhiao Huang, Yuanming Hu, Tao Du, Siyuan Zhou, Hao Su, Joshua B. Tenenbaum, and Chuang Gan. Plasticinelab: A soft-body manipulation benchmark with differentiable physics. *arXiv preprint arXiv:2104.03311*, 2021.

Stephen James, Zicong Ma, David Rovick Arrojo, and Andrew J Davison. Rlbench: The robot learning benchmark & learning environment. *IEEE Robotics and Automation Letters*, 5(2):3019–3026, 2020.

Joel Jang, Seonghyeon Ye, Zongyu Lin, Jiannan Xiang, Johan Bjorck, Yu Fang, Fengyuan Hu, Spencer Huang, Kaushil Kundalia, Yen-Chen Lin, et al. Dreamgen: Unlocking generalization in robot learning through video world models. *arXiv preprint arXiv:2505.12705*, 2025.

Jiaming Ji, Borong Zhang, Jiayi Zhou, Xuehai Pan, Weidong Huang, Ruiyang Sun, Yiran Geng, Yifan Zhong, Josef Dai, and Yaodong Yang. Safety gymnasium: A unified safe reinforcement learning benchmark. *Advances in Neural Information Processing Systems*, 36:18964–18993, 2023.

Yuheng Ji, Huajie Tan, Jiayu Shi, Xiaoshuai Hao, Yuan Zhang, Hengyuan Zhang, Pengwei Wang, Mengdi Zhao, Yao Mu, Pengju An, et al. Robobrain: A unified brain model for robotic manipulation from abstract to concrete. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1724–1734, 2025.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Pannag Sanketi, Quan Vuong, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

Xiaoqi Li, Muhe Cai, Jiadong Xu, Juan Zhu, Hongwei Fan, Yan Shen, Guangrui Ren, and Hao Dong. At-vla: Adaptive tactile injection for enhanced feedback reaction in vision-language-action models. *arXiv preprint arXiv:2605.07308*, 2026.

Xuanlin Li, Kyle Hsu, Jiayuan Gu, Karl Pertsch, Oier Mees, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, et al. Evaluating real-world robot manipulation policies in simulation. *arXiv preprint arXiv:2405.05941*, 2024.

Xingyu Lin, Yufei Wang, Jake Olkin, and David Held. Softgym: Benchmarking deep reinforcement learning for deformable object manipulation. *arXiv preprint arXiv:2011.07215*, 2020.

Aohua Liu, Kun Qian, Zhaokun Yue, Zhuoxuan Wang, Boyi Duan, and Shan Luo. Learning physics-aware sensorimotor model with visual–tactile sensing for dlo manipulation. *IEEE/ASME Transactions on Mechatronics*, 2026.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *arXiv preprint arXiv:2306.03310*, 2023.

Haoran Lu, Ruihai Wu, Yitong Li, Sijie Li, Ziyu Zhu, Chuanruo Ning, Yan Shen, Longzan Luo, Yuanpei Chen, and Hao Dong. Garmentlab: A unified simulation and benchmark for garment manipulation. *Advances in Neural Information Processing Systems*, 37:11866–11903, 2024.

Quan Khanh Luu, Pokuang Zhou, Zhengtong Xu, Zhiyuan Zhang, Qiang Qiu, and Yu She. Manifeel: Benchmarking and understanding visuotactile manipulation policy learning. *arXiv preprint arXiv:2505.18472*, 2025.

Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters*, 7(3):7327–7334, 2022. doi: 10.1109/LRA.2022.3180108.

Masoud Moghani, Mahdi Azizian, Animesh Garg, Yuke Zhu, Sean Huver, and Ajay Mandlekar. Softmimicgen: A data generation system for scalable robot learning in deformable object manipulation. *arXiv preprint arXiv:2603.25725*, 2026.

Marco Moletta, Michael C Welle, and Danica Kragic. Preference aligned visuomotor diffusion policies for deformable object manipulation. *IEEE Robotics and Automation Letters*, 2026.

Yao Mu, Tianxing Chen, Zanxin Chen, Shijia Peng, Zhiqian Lan, Zeyu Gao, Zhixuan Liang, Qiaojun Yu, Yude Zou, Mingkun Xu, et al. Robotwin: Dual-arm robot benchmark with generative digital twins. *arXiv preprint arXiv:2504.13059*, 2025.

Soroush Nasiriany, Abhiram Maddukuri, Lance Zhang, Adeet Parikh, Aaron Lo, Abhishek Joshi, Ajay Mandlekar, and Yuke Zhu. Robocasa: Large-scale simulation of everyday tasks for generalist robots. *arXiv preprint arXiv:2406.02523*, 2024.

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.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

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

Zilin Si and Wenzhen Yuan. Taxim: An example-based simulation model for gelsight tactile sensors. *IEEE Robotics and Automation Letters*, 7(2):2361–2368, 2022. doi: 10.1109/LRA.2022.3142412.

Zilin Si, Gu Zhang, Qingwei Ben, Branden Romero, Zhou Xian, Chao Liu, and Chuang Gan. Difftactile: A physics-based differentiable tactile simulator for contact-rich robotic manipulation. *arXiv preprint arXiv:2403.08716*, 2024.

Taiyi Su, Jian Zhu, Tianjian Wang, Youzhang He, Zitai Huang, Jianjun Zhang, Chong Ma, Hanyang Wang, Tianjiao Zhang, Munan Yin, et al. Demavla: A vision-language-action foundation model for generalizable deformable manipulation. *arXiv preprint arXiv:2605.31286*, 2026.

Yuhao Sun, Shixin Zhang, Zixi Chen, Zirong Shen, Fuchun Sun, Cesare Stefanini, Di Guo, Shan Luo, Jianwei Zhang, Jianhua Shan, et al. Soft contact simulation and manipulation learning of deformable objects with vision-based tactile sensor. *IEEE Transactions on Automation Science and Engineering*, 2025.

Sudharshan Suresh, Haozhi Qi, Tingfan Wu, Taosha Fan, Luis Pineda, Mike Lambeta, Jitendra Malik, Mrinal Kalakrishnan, Roberto Calandra, Michael Kaess, et al. Neuralfeels with neural fields: Visuotactile perception for in-hand manipulation. *Science Robotics*, 9(96):eadl0628, 2024.

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. In *Conference on Robot Learning*, pp. 1723–1736. PMLR, 2023.

Qiwei Wu, Rui Zhang, Xin Xiang, Tao Li, Weihua Zhang, Junjie Lai, and Renjing Xu. Tabero: Learning gentle manipulation with closed-loop force feedback from vision, touch, and language. *arXiv preprint arXiv:2605.27886*, 2026.

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

Wenzhen Yuan, Siyuan Dong, and Edward H. Adelson. Gelsight: High-resolution robot tactile sensors for estimating geometry and force. *Sensors*, 17(12):2762, 2017. doi: 10.3390/s17122762.

Borong Zhang, Yuhao Zhang, Jiaming Ji, Yingshan Lei, Juntao Dai, Yuanpei Chen, and Yaodong Yang. Safevla: Towards safety alignment of vision-language-action model via constrained learning. *Advances in Neural Information Processing Systems*, 38:153335–153373, 2025a.

Chaofan Zhang, Peng Hao, Xiaoge Cao, Xiaoshuai Hao, Shaowei Cui, and Shuo Wang. Vtla: Vision-tactile-language-action model with preference learning for insertion manipulation. *arXiv preprint arXiv:2505.09577*, 2025b.

Yuying Zhang, Kevin Sebastian Luck, Francesco Verdoja, Ville Kyrki, and Joni Pajarinen. Modesuite: Robot learning task suite for benchmarking mobile manipulation with deformable objects. *arXiv preprint arXiv:2507.21796*, 2025c.

Chao Zhao, Chunli Jiang, Lifan Luo, Shuai Yuan, Qifeng Chen, and Hongyu Yu. Learning thin deformable object manipulation with a multi-sensory integrated soft hand. *IEEE Transactions on Robotics*, 2025.

Yongqiang Zhao, Kun Qian, Boyi Duan, and Shan Luo. Fots: A fast optical tactile simulator for sim2real learning of tactile-motor robot manipulation skills. *IEEE Robotics and Automation Letters*, 9(6):5647–5654, 2024. doi: 10.1109/LRA.2024.3396665.

Yuhang Zheng, Songen Gu, Weize Li, Yupeng Zheng, Yujie Zang, Shuai Tian, Xiang Li, Ce Hao, Chen Gao, Si Liu, et al. Omnivta: Visuo-tactile world modeling for contact-rich robotic manipulation. *arXiv preprint arXiv:2603.19201*, 2026.

Jihong Zhu, Andrea Cherubini, Claire Dune, David Navarro-Alarcon, Farshid Alambeigi, Dmitry Berenson, Fanny Ficuciello, Kensuke Harada, Jens Kober, Xiang Li, et al. Challenges and outlook in robotic manipulation of deformable objects. *IEEE Robotics & Automation Magazine*, 29(3):67–77, 2022.

Olgierd Cecil Zienkiewicz, Robert Leroy Taylor, Perumal Nithiarasu, and JZ Zhu. *The finite element method*, volume 3. Elsevier, 1977.

Anya Zorin, Zilin Si, Myungsun Park, Junsung Park, Alexiy Buynitsky, Sachin Bhadang, Taejun Park, Sohee John Yoon, Yong-Lae Park, Oliver Kroemer, et al. Taco: Benchmarking tactile sensors for object manipulation. *arXiv preprint arXiv:2605.21976*, 2026.

A Implementation Details

SoftVTBench is implemented in Isaac Sim 4.5.0 with Isaac Lab 0.41.3 and the GPU-accelerated PhysX 5 pipeline. The simulator runs physics at 60 Hz with a control decimation of 3, resulting in a 20 Hz control and logging rate. All visual, tactile, proprioceptive, action, and privileged-state streams are synchronized at this rate. Table 5 summarizes the robot, control, sensing, simulation, and hardware configuration.

Table 5 Simulation and sensing configuration.

Item	Value
Simulator	Isaac Sim 4.5.0 / Isaac Lab 0.41.3, PhysX 5 GPU pipeline
Physics / control rate	60 Hz physics, decimation 3, 20 Hz control
Robot	Franka arm with Panda parallel-jaw gripper
Controller	Task-space differential inverse kinematics
End-effector action	Absolute pose target: 3D position and 3D axis-angle orientation
Gripper action	Normalized closure command; continuous and binary encodings
Finger friction	Static $\mu_s = 1.5$, dynamic $\mu_d = 1.2$, max combine mode
Camera views	Third-person 1024×1024 and wrist 512×512, resized to 224×224
Tactile sensor	GelSight Mini via TacEx; Taxim optics and FOTS markers
Tactile streams	Tactile RGB and 11×9 marker-motion field, 320×240
FEM model	PhysX soft body with corotational linear elasticity
FEM solver	Hex resolution 6, 64 position iterations, damping 2.5
Collection / evaluation	1× NVIDIA L20 GPU per worker
Training	8× NVIDIA A100-80GB / A800 GPUs

B Task Suite Details

B.1 Task Suite Summary

SoftVTBench contains four task suites arranged as a matched 2×2 design over object type and variation axis. The object type is either rigid or deformable, and the variation axis is either object identity or spatial layout. We refer to the four suites as object-rigid, object-soft, spatial-rigid, and spatial-soft.

The deformable suites contain 10 tasks with 50 expert demonstrations per task. The rigid-control suites are built by re-executing and filtering LIBERO demonstrations under the same SoftVTBench sensing and recording stack. All suites share the same robot embodiment, camera views, tactile sensing interfaces, rollout API, and evaluation protocol.

Table 6 Statistics of the four task suites in SoftVTBench, including object type, variation axis, number of tasks, demonstration trajectories, and validation episodes.

Suite	Object Type	Variation Axis	#Tasks	#Demos	#Val Episodes
Object-Rigid	Rigid	Object identity	10	500	50
Spatial-Rigid	Rigid	Spatial layout	10	500	50
Object-Soft	Deformable	Object identity	10	500	50
Spatial-Soft	Deformable	Spatial layout	10	500	50
Total	–	–	40	2000	200

B.2 Object-Soft Suite

The object-soft suite evaluates deformable-object manipulation under object identity variation. In contrast to the spatial-soft suite, where two visually similar objects appear in the same scene and the target is specified by a spatial-language cue, the object-soft suite keeps the task structure fixed and varies the manipulated deformable object. Each task uses a different soft object category or asset, and the policy must adapt its grasping strategy to the object's geometry, compliance, contact surface, and visual appearance.

Each task contains 50 expert demonstrations. The robot is instructed to pick up the specified deformable object and place it into the target receptacle. The suite therefore tests object-level generalization across soft bodies while holding the high-level manipulation goal fixed.

Table 7 Object-soft task definitions.

Task	Language Prompt	#Demos
0	pick up the pale cream round steamed-bun pastry and place it in the basket	50
1	pick up the soft pastry and place it in the basket	50
2	pick up the soft pastry and place it in the basket	50
3	pick up the soft pastry and place it in the basket	50
4	pick up the orange round bun pastry with sesame speckles and place it in the basket	50
5	pick up the soft pastry and place it in the basket	50
6	pick up the soft cube and place it in the basket	50
7	pick up the soft cylinder and place it in the basket	50
8	pick up the soft wedge and place it in the basket	50
9	pick up the soft capsule and place it in the basket	50

B.3 Spatial-Soft Suite

The spatial-soft suite evaluates deformable-object manipulation under paired spatial layouts. Each scene contains two visually identical pastry instances: one is the target object and the other is a distractor. For each layout, we define two tasks by swapping which instance is referred to in the language instruction. Thus, the physical scene layout is shared within each pair, while the target object is changed through the language prompt. This design tests whether policies can ground language to the correct deformable object under spatial ambiguity.

All spatial-soft demonstrations use the same robot, sensing stack, visual setting, and replay-format recording pipeline. Each task contains 50 expert demonstrations. The final replay-format dataset stores one HDF5 file per task with 50 demonstrations, together with four synchronized single-view videos per demonstration: third-person RGB, wrist RGB, left tactile marker video, and right tactile marker video. The 2×2 preview videos are used only for inspection and are not included in the final replay-format dataset.

Table 8 Spatial-soft task definitions. Each pair shares the same physical layout and pastry type, but swaps the target instance through the language prompt.

Task	Language Prompt	#Demos
0	pick up the right white spiral pastry with ridged frosting and place it on the plate	50
1	pick up the left white spiral pastry with ridged frosting and place it on the plate	50
2	pick up the right pale cream round steamed-bun pastry with black eyes and place it on the plate	50
3	pick up the left pale cream round steamed-bun pastry with black eyes and place it on the plate	50
4	pick up the left small dark brown oval pastry and place it on the plate	50
5	pick up the right small dark brown oval pastry and place it on the plate	50
6	pick up the left long golden braided pastry with orange stripes and place it on the plate	50
7	pick up the right long golden braided pastry with orange stripes and place it on the plate	50
8	pick up the right orange round bun pastry with sesame speckles and place it on the plate	50
9	pick up the left orange round bun pastry with sesame speckles and place it on the plate	50

Table 9 Asset statistics in SoftVTBench. The benchmark combines Tabero/LIBERO rigid scene assets with deformable assets adapted or procedurally generated for safety-aware visuo-tactile manipulation.

Category	Number	Assets
Scene surfaces / tables	2	floor, table
Articulated scene objects	2	wooden cabinet, flat stove
Rigid manipulation objects	15	black bowl, alphabet soup can, basket, BBQ sauce bottle, butter box, chocolate pudding cup, cookies box, cream cheese box, glazed porcelain ramekin, ketchup bottle, milk carton, orange juice carton, plate, salad dressing bottle, tomato sauce bottle
Deformable bakery-style objects	11	round sesame bun, elongated braided roll, striped croissant-like pastry, small ridged bread roll, yellow rolled pastry, curved layered pastry, twisted bread stick, spiral pastry roll, oval loaf with ridges, long soft bread roll, compact golden bun
Deformable procedural objects	3	sphere, cube, cylinder
Total	33	19 Tabero/LIBERO scene and rigid assets + 14 deformable assets

C Deformable Assets and Interaction Calibration

C.1 Deformable Object Simulation

SoftVTBench uses 3D deformable assets represented as volumetric meshes and simulated as PhysX GPU FEM soft bodies with corotational linear elasticity. The assets include naturalistic bakery-style objects and simple geometric primitives. Per-asset physical parameters, such as density, friction, elasticity, and damping, are authored in the simulation assets and are not provided to the policy. Before each episode, the object is settled on the support surface to obtain a stable initial state. The expert policy and the evaluation module use this settled state, rather than the nominal spawn pose, to avoid errors caused by pre-grasp drift.

C.2 Asset List

Table 9 lists the deformable assets and rigid assets used in SoftVTBench. The bakery-style pastry assets are adapted from publicly available 3D assets from the EXTWIN Synthesis asset library[1], and are converted into simulation-ready deformable meshes for SoftVTBench. The main experiments use the subset of assets that are stable under repeated closed-loop manipulation and have valid calibration results.

C.3 Interaction Calibration Protocol

For each deformable object, we perform an offline interaction calibration to estimate its feasible grasping range and deformation threshold. The lower bound is obtained by scanning gripper closure commands from loose to tight and executing a grasp–lift–hold routine. The smallest closure that reliably lifts and holds the object without slip or drop is recorded as the lower bound.

The upper calibration is obtained from a compression sweep. For each closure command, we record the maximum stable FEM RMS deformation of the object. We define the reference deformation of object o as D_o^{ref}, the maximum stable FEM RMS deformation measured during the sweep. The safety threshold is then defined as

$$\tau_o = \kappa D_o^{\text{ref}}, \tag{6}$$

where we use $\kappa = 0.5$ by default and additionally report $\kappa \in \{0.3, 0.7\}$ for sensitivity analysis. The resulting threshold is used only by the evaluator and is not exposed to the policy.

[1] https://synthesis.extwin.com/