SOFTVTBENCH: A DEFORMATION-AWARE VISUO-TACTILE DATASET AND BENCHMARK FOR DEFORMABLE-OBJECT MANIPULATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Robot learning typically evaluates manipulation by its outcome, yet task success cannot distinguish a stable grasp from one that completes the task by over-compressing a deformable object. Measuring both requires a record that existing datasets lack: tactile sensing available to the policy and physical state available only to the evaluator. We introduce SoftVTBench, a simulated visuo-tactile dataset and benchmark for deformable-object manipulation, instantiated through complete pick-and-place tasks. The dataset provides 4,000 demonstrations over matched deformable objects and rigid twins, pairing policy-visible vision, touch, proprioception, language, and actions with evaluator-only finite-element states. The benchmark defines the Deformation-aware Success Rate (DSR), which credits a rollout only if it completes the task without exceeding an object-specific deformation tolerance calibrated before policy training. Across three policy families, this evaluation exposes behavior hidden by task success alone: every in-distribution configuration contains successful rollouts that exceed the tolerance, configurations with nearly identical task success can differ substantially in deformation compliance, and, in a controlled $\pi_{0.5}$ ablation, finer gripper control can reproduce an apparent modality gain. Visuo-tactile variants achieve higher absolute out-of-distribution task success in all six evaluated comparisons, but their in-distribution effects are mixed. Within this simulated setting, these results indicate that task completion and deformation compliance capture distinct aspects of policy behavior: higher task success does not necessarily imply better deformation compliance.

1 INTRODUCTION

Many everyday objects, such as bread, fruit, and packaged food, deform when a robot grasps them. For such objects, how the robot holds an object matters as much as whether it picks it up: too little grip lets the object slip, too much compresses it beyond what the task allows, and the acceptable range differs from object to object. Vision locates the object and the target, but the gripper–object interface is often occluded once contact begins. Vision-based tactile sensors such as GelSight (Yuan et al., 2017) observe this interface directly, recording contact geometry, shear, slip, and local compression (Suresh et al., 2024). Whether and when this local signal improves the manipulation of deformable objects is therefore a natural question for multimodal policy learning.

Studying this question requires shared resources that remain scarce: synchronized visual and tactile demonstrations on deformable objects, tasks on which methods can be compared under matched conditions, and a measure of what a rollout did to the object, not only whether it reached the goal. Task success addresses the last need only in part. A loose grasp that drops the object fails the task and is detected; an overly tight grasp may complete the same task while exceeding an acceptable deformation, and task success does not distinguish it from a grasp that stays within tolerance. Existing resources cover parts of this need (Table 1): general suites evaluate task success mostly on rigid objects (Liu et al., 2023; Nasiriany et al., 2024), deformable-object benchmarks target shape change or isolated grasps (Lin et al., 2020; Greenland et al., 2024), and visuo-tactile benchmarks score task success or contact force rather than soft-body deformation (Luu et al., 2025; Wu et al., 2026).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

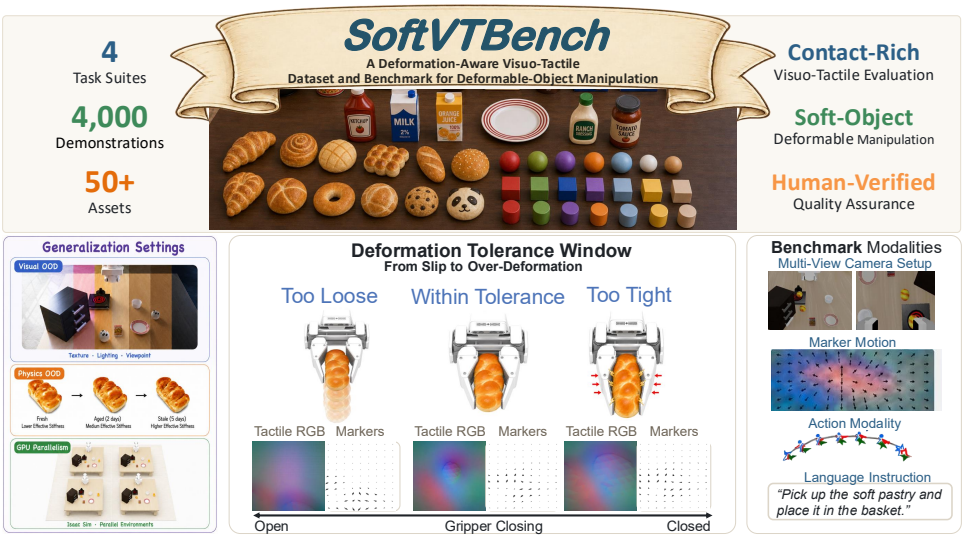

Figure 1: **Overview of SoftVTBench**, a simulated multimodal dataset and benchmark for deformable-object manipulation. It contains four task suites and 4,000 expert demonstrations over ten volumetric deformable objects and their rigid twins, in scenes built from more than 50 assets, with synchronized visual, tactile, proprioceptive, language, and action data. Gripper panels are illustrative.

We introduce **SoftVTBench**, a simulated multimodal dataset and benchmark for deformable-object manipulation with vision and vision-based tactile sensing (Figure 1), instantiated as pick-and-place across four task suites. It contains 4,000 scripted demonstrations over ten volumetric deformable objects and matched rigid twins, with multi-view RGB, dual-finger tactile observations, proprioception, language, and actions recorded at 20 Hz. SoftVTBench supports policy learning, contact analysis, and controlled comparisons through binary and continuous gripper commands, finite-element states used only for scoring, matched rigid twins, and single-factor shifts in lighting, mass, and stiffness. In addition to Task Success Rate (TSR), we report Deformation-aware Success Rate (DSR), which requires both task completion and compliance with an object-specific deformation tolerance calibrated before policy training. This tolerance is an operational benchmark criterion rather than a material-damage threshold. Demonstrations are collected within each object's calibrated grasp range, and tactile observations are rendered using simulators previously validated against physical GelSight sensors (Si & Yuan, 2022; Zhao et al., 2024).

We benchmark Diffusion Policy (Chi et al., 2023), $\pi_{0.5}$ (Physical Intelligence et al., 2025), and FastWAM (Yuan et al., 2026) with vision-only and visuo-tactile inputs in ID and OOD settings. The results highlight two effects. First, task success can conceal deformation violations: all 12 ID configurations contain successful rollouts that exceed the tolerance, accounting for 0.7–24% of their successes. Second, apparent tactile gains can depend on the action interface: on OBJECT-SOFT, switching $\pi_{0.5}$ from binary to continuous gripper control raises TSR by 11.4 points without touch, comparable to the 10.8-point gain from adding touch. Visuo-tactile variants achieve higher OOD task success in all six policy–suite comparisons, while their ID effects are mixed. SoftVTBench thus separates gains due to touch from those due to other design choices.

Our contributions are:

- **Visuo-tactile dataset:** 4,000 demonstrations on deformable objects and rigid twins, with the simulator states needed to recompute deformation;
- **Deformation-aware benchmark:** closed-loop evaluation with pre-calibrated per-object tolerances, TSR and DSR, and matched rigid, action, and OOD comparisons;
- **Diagnostic study:** experiments on three policy families showing that task success can hide tolerance violations and that apparent tactile gains can be confounded with gripper control.

Table 1: **Benchmark positioning.** *Complete Task*: multi-stage manipulation rather than an isolated grasp. *3D Deformable*: volumetric soft bodies (❍: cloth or rope). *Policy-Visible Touch*: tactile observations are policy inputs. *Evaluator-Only Scoring*: deformation from hidden physical state enters the success criterion (❍: scored via contact force, not deformation).

Benchmark	Complete Task	3D Deformable	Policy-Visible Touch	Evaluator-Only Scoring
LIBERO (Liu et al., 2023)	✓	✗	✗	✗
ManiSkill2 (Gu et al., 2023)	✓	✓	✗	✗
SoftGym (Lin et al., 2020)	✓	❍	✗	✗
DefGraspSim (Huang et al., 2022)	✗	✓	✗	✓
SoGraB (Greenland et al., 2024)	✗	✓	✗	✓
ManiFeel (Luu et al., 2025)	✓	✗	✓	✗
Tabero (Wu et al., 2026)	✓	✗	✓	❍
SoftVTBench (ours)	✓	✓	✓	✓

2 Related Work

Robotic Manipulation Benchmarks. General-purpose suites standardize complete-task evaluation across multi-task, language-conditioned, and generalization settings, including Meta-World (Yu et al., 2020), RLBench (James et al., 2020), LIBERO (Liu et al., 2023), CALVIN (Mees et al., 2022), RoboCasa (Nasiriany et al., 2024), RoboTwin (Mu et al., 2025; Chen et al., 2025), THE COLOSSEUM (Pumacay et al., 2024), and SIMPLER (Li et al., 2024). Safety-oriented benchmarks (Zhang et al., 2025a; Fan et al., 2026; Huang et al., 2026) additionally score constraint violations and semantic hazards. These benchmarks primarily focus on rigid-object manipulation and do not directly score soft-body deformation resulting from contact.

Deformable-Object Manipulation. SoftGym (Lin et al., 2020), DEDO (Antonova et al., 2021), and GarmentLab (Lu et al., 2024) cover cloth, rope, and garments; PlasticineLab (Huang et al., 2021) and DaXBench (Chen et al., 2022) provide differentiable soft-body simulation; ManiSkill2 (Gu et al., 2023) and MoDeSuite (Zhang et al., 2025c) include deformable objects in broader suites; and DeMaVLA (Su et al., 2026) studies generalization with foundation-model policies. In many of these tasks, changing the object's shape is itself the objective. SoGraB (Greenland et al., 2024), DefGraspSim (Huang et al., 2022), and DefGraspNets (Huang et al., 2023) study grasp quality under deformable-object physics, including deformation- or stress-aware evaluation and planning, but focus on isolated grasps rather than complete manipulation trajectories. SoftVTBench instead targets deformation-aware handling over complete pick-and-place tasks.

Visuo-Tactile Sensing and Policy Learning. Tactile sensing captures contact geometry, shear, slip, and compression that are difficult to infer from vision (Yuan et al., 2017; Suresh et al., 2024). Tactile simulators (Si & Yuan, 2022; Zhao et al., 2024; Nguyen et al., 2024; Akinola et al., 2025; Si et al., 2024) support policy learning; Taxim and FOTS report agreement with physical GelSight sensors in their original calibration and transfer settings (Si & Yuan, 2022; Zhao et al., 2024), which motivates their use for rendering tactile observations here. Recent vision-language-action models also incorporate touch (Zhang et al., 2025b; Cheng et al., 2025; Bi et al., 2025; Huang et al., 2025; Li et al., 2026). Visuo-tactile benchmarks such as ManiFeel (Luu et al., 2025), TacO (Zorin et al., 2026), and VTDexManip (Liu et al., 2025) primarily evaluate tactile sensing through task-level manipulation performance, mostly on rigid objects, while visuo-tactile methods for deformable objects (Sunil et al., 2023; Wi et al., 2022) study specific manipulation or dynamics-estimation systems rather than complete-task benchmarking with evaluator-only deformation scoring. Closest to our setting, Tabero (Wu et al., 2026) evaluates gentle manipulation using closed-loop force feedback, but scores interaction through contact force on rigid objects rather than deformation of a soft body.

3 THE SOFTVTBENCH DATASET

SoftVTBench pairs policy-visible visuo-tactile observations with evaluator-only physical states over complete pick-and-place tasks. Figure 2 summarizes its construction. It is implemented in Isaac Sim (NVIDIA Corporation, 2025) and Isaac Lab (Mittal et al., 2025).

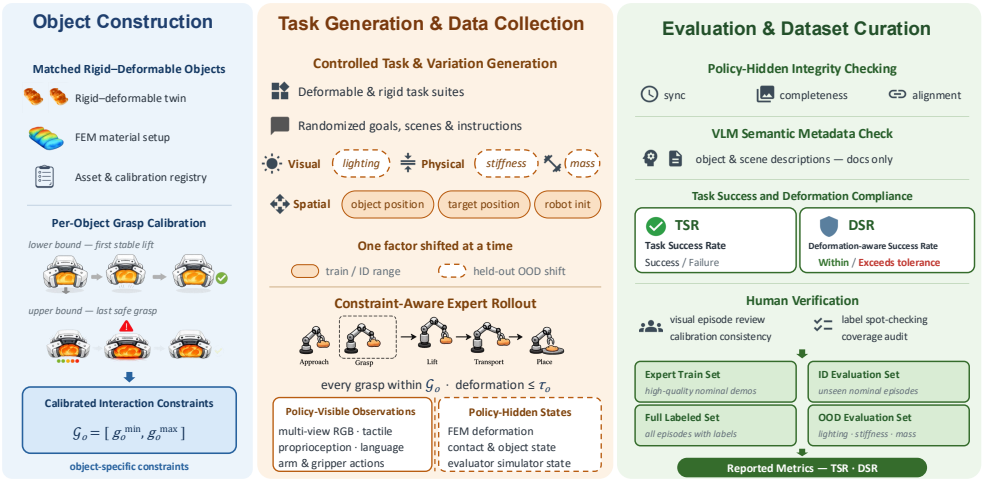

Figure 2: **SoftVTBench construction pipeline.** (1) Matched rigid–deformable objects and object-specific calibration; (2) controlled task generation with separate policy-visible and evaluator-only recording; (3) automatic and human quality control.

3.1 ASSETS AND TASK SUITES

Objects and Scenes. The benchmark manipulates ten volumetric deformable objects, six bakery-style meshes and four procedural primitives that vary in geometry, compliance, contact area, and appearance. Each has a rigid twin with matched mesh, texture, and mass but raised stiffness, so that grasp-induced deformation is negligible. Physical parameters are authored in the simulation assets and never provided to the policy, and each object is settled on its support surface before an episode begins, so pre-grasp drift affects neither collection nor scoring. With tables, floors, cabinets, containers, and distractor objects, the scenes contain more than 50 assets; placement regions follow LIBERO (Liu et al., 2023), and asset details and the task list are given in Appendices C and D.

Task Suites. The 4,000 demonstrations are evenly divided among four suites that form a 2×2 design over object type and variation axis. OBJECT suites fix the layout and vary the manipulated object, probing adaptation to geometry and contact response; SPATIAL suites place two identical instances in one scene and specify the target only through language, probing target grounding. Each suite has ten tasks of 100 demonstrations; OBJECT tasks place the object in a basket and SPATIAL tasks on a plate. Rigid-twin suites reuse the layouts and instructions of their deformable counterparts with the same pick-and-place skill, so suite differences reflect object type or variation axis.

3.2 DEFORMATION MEASURE AND PER-OBJECT CALIBRATION

Deformation from Physical State. Each deformable object is simulated as a volumetric FEM mesh. We define D_t as the RMS nodal displacement at time t after removing rigid-body motion, normalized by the initial bounding-box diagonal; removing rigid motion separates being squeezed from being carried, and normalizing by size makes objects of different sizes comparable. D_t is computed only from evaluator-only states and is negligible for the rigid twins.

Per-Object Calibration. Before any policy is trained, a scripted grasp–lift–hold protocol sweeps the gripper closure from loose to tight. The loosest closure that repeatedly holds the object without

slip defines $g_o^{\min}$; the tolerance τ_o is the 90th percentile of peak D_t over stable grasps; and $g_o^{\max}$ is the tightest closure within τ_o. The envelope $\mathcal{G}_o = [g_o^{\min}, g_o^{\max}]$ constrains expert collection, whereas τ_o defines the benchmark criterion through

$$R_{\max} = \max_t D_t/\tau_o, \tag{1}$$

with $R_{\max} > 1$ indicating that the tolerance was exceeded. Because calibration precedes policy training, no evaluated policy can influence its own criterion. The 90th percentile is a single global constant fixed for all assets, so it cannot be tuned per object or method. Across the ten objects, the envelopes span 2.6 to 26.2 mm of jaw aperture with no closure common to all, and τ_o ranges from 7.1% to 11.2% of the bounding-box diagonal, so both must be set per object. The tolerance is an operational criterion in simulation, not a material-damage threshold. Per-asset parameters are in Appendix C; the released records allow rescoring at other percentiles without new rollouts.

3.3 Observations, Actions, and Physical States

Observations, actions, and simulator states share one 20 Hz timeline per episode. Observations and actions support policy learning; simulator states serve only quality control and evaluation.

Policy-Visible Observations and Actions. Third-person and wrist RGB describe the scene, while each finger records tactile RGB and a marker-motion field capturing contact appearance and local shear, rendered by TacEx (Nguyen et al., 2024) for simulated GelSight Mini sensors (Yuan et al., 2017) with Taxim (Si & Yuan, 2022) optics and FOTS (Zhao et al., 2024) markers. Each finger yields a 320×240 tactile image and an 11×9 marker field. Objects differing in geometry and compliance produce distinct contact patches and shear fields absent from external views (Figure 5, appendix). The observation is $o_t = \{I_t^{\text{third}}, I_t^{\text{wrist}}, \mathcal{T}_t, p_t, \ell\}$, where $\mathcal{T}_t$ denotes the dual-finger tactile images and marker fields, p_t the proprioceptive state, and ℓ the instruction; the action $a_t = (\mathbf{x}_t^{ee}, \boldsymbol{\theta}_t^{ee}, g_t)$ is an absolute end-effector pose and a gripper command g_t, stored as both a binary signal and a continuous closure target so modality and control granularity can be crossed without new data.

Evaluator-Only Physical States. FEM nodal positions, object poses, and contact and drop events are logged but hidden from the policy; only quality control and the evaluator (Section 4) read them. Releasing them lets the criterion be recomputed, audited, or redefined without new data.

3.4 Collection, Quality Control, and Splits

Collection. A scripted expert executes approach, grasp, lift, transport, and place, drawing its closure from $\mathcal{G}_o$ and randomizing object pose, robot configuration, and target location. Each trajectory is simulated once and rendered from its stored replay, so all streams are aligned. Across the 2,000 deformable-object demonstrations, $R_{\max}$ has median 0.433 and 95th percentile 0.713, and none exceeds its tolerance, so policy violations in Section 5 are not inherited from supervision.

Quality Control. Of 4,867 collected episodes, automatic checks of stream synchronization, completeness, and alignment, a vision-language model check of object and scene descriptions, and human review of episodes, outcome labels, and calibration consistency rejected 867 (17.8%), yielding the 4,000 released demonstrations.

Variation Factors and Splits. Training episodes vary the spatial factors (object placement, target placement, and robot initialization) over per-task supports, while the visual (dome-light intensity) and physical (object mass and Young's modulus) factors are held at their nominal values (Table 10); together these define the in-distribution support. In-distribution evaluation uses 500 held-out initial states per suite, 50 per task. Out-of-distribution evaluation applies nine single-factor conditions to the two deformable suites: dome-light intensity changes from 135 to 67.5, 180, and 270; mass is scaled by 1.25, 1.75, and 2.5; and Young's modulus by 0.5, 0.8, and 2.0. All held-out values differ from the nominal training values. Each condition uses 100 of the in-distribution initial states (10 per task) with their seeds and changes only the shifted factor.

4 THE SOFTVTBENCH BENCHMARK

SoftVTBench evaluates whether a policy completes an instructed task while keeping object deformation within its calibrated tolerance. Policies are executed in closed loop on the fixed evaluation splits of Section 3.4, using the same physics and observation interface as data collection.

4.1 CLOSED-LOOP PROTOCOL

Interface. At each policy query, the policy receives the policy-visible observation of Section 3.3: third-person and wrist RGB, proprioception, the instruction for language-conditioned policies, and, for visuo-tactile variants, dual-finger tactile images and marker fields. It predicts a chunk of absolute end-effector pose targets and gripper commands, in the binary or continuous encoding on which it was trained, which the simulator executes at 20 Hz for a model-specific number of steps before the policy replans. Each executed action advances the robot, object, and contact states under physics, so perception and control are coupled in closed loop rather than through a replayed trajectory. An episode runs for at most 300 control steps and terminates early on success or on a drop. Evaluator-only states are read throughout the rollout but never returned to the policy, which must infer contact conditions, incipient slip, and object compliance from its own inputs. All methods are evaluated on the same task instances, initial states, and random seeds.

ID and OOD Evaluation. In-distribution evaluation uses 500 episodes per suite and configuration. Out-of-distribution evaluation applies the nine single-factor conditions of Section 3.4 to the two deformable suites, 100 episodes each, and reports scores pooled over the resulting 900 episodes per suite and configuration. Each OOD condition reuses task instances, initial states, and seeds from the in-distribution evaluation. Reported Δ values are computed relative to the corresponding full in-distribution score in Table 2.

4.2 METRICS

Task Success. Task success requires the instructed object to come to rest inside the designated target region; when multiple identical instances are present, the transported object must be the instructed instance. The predicate is purely kinematic: deformation does not enter it, which is what allows us to measure how much a success-only criterion misses.

Deformation-aware Success Rate. Let $T_i \in \{0, 1\}$ indicate task success in episode i, and let $R^{(i)}_{\max}$ be its normalized peak deformation (Eq. 1), computed from the trace $R^{(i)}_t = D_{i,t}/\tau_{o_i}$. Here, $R^{(i)}_{\max} > 1$ means that the rollout exceeded its calibrated deformation tolerance at some point, whether or not it recovered afterwards. We report

$$\mathrm{TSR} = \frac{1}{N}\sum_{i=1}^{N} T_i, \qquad \mathrm{DSR} = \frac{1}{N}\sum_{i=1}^{N} T_i \cdot \mathbf{1}\left[R^{(i)}_{\max} \le 1\right]. \tag{2}$$

DSR is the primary metric and TSR a diagnostic reference. Because task success is part of DSR, a policy that avoids contact and never completes the task receives no credit. The difference $\mathrm{TSR} - \mathrm{DSR}$ is exactly the fraction of episodes that succeed while exceeding the tolerance, and $(\mathrm{TSR} - \mathrm{DSR})/\mathrm{TSR}$ is their share among a policy's own successes. For the rigid suites, deformation is negligible by construction, so we report TSR only.

5 EXPERIMENTS

Setup. We benchmark Diffusion Policy (Chi et al., 2023), trained from scratch without language conditioning; $\pi_{0.5}$ (Physical Intelligence et al., 2025), LoRA-fine-tuned from its pretrained checkpoint; and the world–action model FastWAM (Yuan et al., 2026). Each has a vision-only (VO) and a visuo-tactile (VT) variant, the latter adding tactile RGB and marker motion through family-specific encoders; unless stated otherwise, both use continuous gripper control (-C). Each configuration is a single training run evaluated at its final checkpoint (Appendix E). VT variants have more trainable parameters and, for Diffusion Policy and FastWAM, a smaller batch size, so VO–VT differences reflect the full input interface rather than touch in isolation.

5.1 Does Task Success Hide Deformation Violations?

Table 2: **In-distribution TSR and DSR on the deformable suites** (%). DSR is lower than TSR in every configuration; the gap is the fraction of rollouts that succeed while exceeding the tolerance.

		Object-Soft		Spatial-Soft	
Model	**Input**	**TSR**	**DSR**	**TSR**	**DSR**
Diffusion Policy	VO-C	37.4	33.6	15.6	13.4
	VT-C	40.0	30.4	33.0	25.0
$\pi_{0.5}$	VO-C	41.6	38.4	26.0	22.6
	VT-C	41.4	35.0	27.6	22.0
FastWAM	VO-C	**62.0**	**58.0**	37.0	36.6
	VT-C	57.6	54.4	**56.4**	**56.0**

Finding 1: Task success conceals tolerance violations in every configuration. The TSR–DSR gap is non-zero in all twelve configurations of Table 2. For Diffusion Policy VT-C it reaches 9.6 points on Object-Soft and 8.0 on Spatial-Soft, 24% of its successes on each suite. Because TSR and DSR are computed on the same rollouts and the tolerance is fixed per object before training, the gap is an exact count under a criterion no policy can adjust.

Violations are policy dependent. The share of successes that exceed the tolerance is 10–24% for Diffusion Policy and 8–20% for $\pi_{0.5}$, but only 0.7–6.5% for FastWAM, which keeps TSR and DSR within 0.4 points on both spatial configurations while attaining the highest spatial TSR. TSR alone therefore cannot separate a policy that succeeds within tolerance from one that succeeds by exceeding it. In Figure 4 (appendix), rollouts within tolerance peak at $R_{\max}$ of 0.28–0.74 and settle, whereas equally successful rollouts rise above 1 during the grasp; their external views look similar, but their marker fields and deformation traces do not.

5.2 What Does Deformability Cost Relative to a Rigid Twin?

Table 3: **TSR on deformable objects and their rigid twins** (%). Twins match geometry, appearance, and mass under the same layouts and instructions, so a rigid–deformable gap isolates deformability.

		Object variation		Spatial variation	
Model	**Input**	**Rigid**	**Soft**	**Rigid**	**Soft**
Diffusion Policy	VO-C	40.0	37.4	14.0	15.6
	VT-C	35.0	40.0	11.0	33.0
$\pi_{0.5}$	VO-C	60.0	41.6	**50.4**	26.0
	VT-C	59.6	41.4	**54.0**	27.6
FastWAM	VO-C	**64.0**	**62.0**	25.0	**37.0**
	VT-C	**61.6**	**57.6**	30.0	**56.4**

Finding 2: The measured cost of deformability depends on policy and task structure. On object variation, the matched VO-C pairs of Table 3 show that $\pi_{0.5}$ loses 18.4 points from rigid to soft, whereas Diffusion Policy and FastWAM lose 2.6 and 2.0, within evaluation noise. On spatial variation, FastWAM instead scores 12.0 points *higher* on deformable objects, suggesting that target grounding rather than contact physics dominates difficulty there, a confound that only the rigid-twin control can separate. Two caveats apply: the rigid VT rows decode binary gripper execution from continuous-policy checkpoints and are corroborating rather than primary evidence, and Diffusion Policy, lacking language conditioning, cannot identify the instructed instance on the spatial suites.

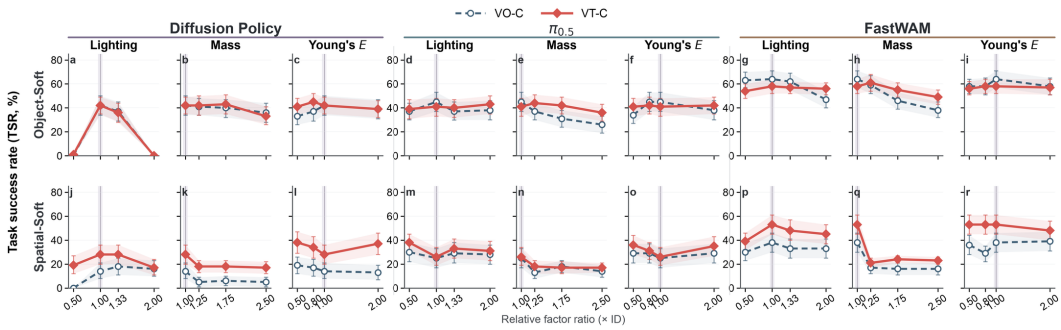

Figure 3: **OOD TSR resolved by shift factor.** Rows: OBJECT-SOFT and SPATIAL-SOFT; columns: illumination, mass, and Young's-modulus shifts for each policy family. The horizontal axis gives the factor relative to its nominal value (grey band); each point averages 100 episodes, with 95% binomial confidence intervals.

5.3 Is an Apparent Tactile Gain Due to Sensing or to Control Granularity?

Comparing binary VO with continuous VT policies confounds sensing with control granularity. Because both gripper encodings are stored for every demonstration (Section 3.3), Table 4 crosses the two factors for $\pi_{0.5}$ without new data.

Table 4: **Matched sensing–control ablation for $\pi_{0.5}$ (%).** VO/VT denote vision-only and visuo-tactile inputs; B/C denote binary and continuous gripper control.

	OBJECT-SOFT		SPATIAL-SOFT	
Configuration	**TSR**	**DSR**	**TSR**	**DSR**
VO-B	30.2	27.2	**34.2**	20.0
VO-C	**41.6**	**38.4**	26.0	**22.6**
VT-B	41.0	28.0	30.0	21.4
VT-C	41.4	35.0	27.6	22.0

Finding 3: An apparent tactile gain can be reproduced by finer gripper control. On OBJECT-SOFT, continuous control alone (VO-C) raises TSR over VO-B by 11.4 points and touch alone (VT-B) by 10.8, while combining both (VT-C) adds nothing beyond continuous control; comparing VO-B with VT-C would thus credit touch with a gain that finer actuation reproduces. DSR sharpens this attribution: VO-C and VT-B differ by 0.6 points in TSR but 10.4 in DSR, with 7.7% versus 31.7% of successes exceeding the tolerance, and continuous control raises DSR in all four binary-to-continuous comparisons. On SPATIAL-SOFT both upgrades lower TSR relative to VO-B, consistent with target grounding being the binding constraint. Action spaces should therefore be matched before a gain is attributed to touch, as in all our main comparisons.

5.4 How Do Visuo-Tactile Variants Behave under Distribution Shift?

Finding 4: Visuo-tactile variants attain higher absolute out-of-distribution success, but not consistently smaller degradation. Pooled over the nine held-out conditions (Table 5), VT-C exceeds VO-C in TSR in all six policy–suite comparisons and in DSR in five, the exception being Diffusion Policy on OBJECT-SOFT; in distribution, the TSR comparison splits four to two. A one-sided sign test gives $p = 0.016$ for six of six and $p = 0.11$ for five of six, which we treat as descriptive because the comparisons share policies and data. The advantage lies in absolute performance rather than robustness: VT-C degrades less than VO-C in only three of six TSR comparisons. Resolved by factor (Figure 3), VT generally lies at or above VO on SPATIAL-SOFT, while on OBJECT-SOFT the curves overlap more closely, separating mainly under mass shifts for $\pi_{0.5}$ and FastWAM; Diffusion Policy fails at both illumination extremes on OBJECT-SOFT under either modality. One plausible reading is that tactile cues, available only after contact, reflect mass-altered contact dynamics, whereas illumination perturbs vision throughout the rollout; the benchmark does not isolate this mechanism.

Table 5: **Out-of-distribution TSR and DSR on the deformable suites** (%), pooled over the nine held-out conditions of Section 3.4.

Model	Input	OBJECT-SOFT		SPATIAL-SOFT	
		TSR ↑ (Δ vs. ID)	DSR ↑ (Δ vs. ID)	TSR ↑ (Δ vs. ID)	DSR ↑ (Δ vs. ID)
Diffusion Policy	VO-C	29.2 (−8.2)	**26.6 (−7.0)**	11.0 (−4.6)	8.8 (−4.6)
	VT-C	**31.2 (−8.8)**	25.0 (−5.4)	**25.2 (−7.8)**	**17.8 (−7.2)**
$\pi_{0.5}$	VO-C	35.8 (−5.8)	33.2 (−5.2)	24.4 (−1.6)	19.4 (−3.2)
	VT-C	**41.0 (−0.4)**	**34.2 (−0.8)**	**28.4 (+0.8)**	**23.2 (+1.2)**
FastWAM	VO-C	54.4 (−7.6)	53.8 (−4.2)	27.8 (−9.2)	27.2 (−9.4)
	VT-C	**55.8 (−1.8)**	**55.8 (+1.4)**	**39.4 (−17.0)**	**38.8 (−17.2)**

Completion and compliance gains need not coincide. Under shift, VT shows a larger TSR–DSR gap than VO for Diffusion Policy on both suites (6.2 vs. 2.6 and 7.4 vs. 2.2 points) and for $\pi_{0.5}$ on OBJECT-SOFT (6.8 vs. 2.6), whereas FastWAM keeps TSR and DSR within one point in every pooled comparison, so its OOD gain from touch adds no violations. A completion-only protocol would record both patterns as uniform progress.

5.5 DOES THE DEFORMATION CRITERION ADD INFORMATION?

The criterion is discriminative. Beyond the VO-C/VT-B contrast of Section 5.3, roughly 3.5 binomial standard errors at 500 episodes, DSR reverses the ranking that TSR gives Diffusion Policy on OBJECT-SOFT: VT-C leads in TSR (40.0% vs. 37.4%, within noise) but trails in DSR (30.4% vs. 33.6%). DSR thus separates configurations that completion alone does not.

The criterion is attainable. No accepted expert demonstration exceeds its tolerance (Section 3.4), and FastWAM attains the highest TSR while keeping TSR and DSR within one point under shift, so the gap is not inherent to deformable objects.

6 CONCLUSION

We presented SoftVTBench, a visuo-tactile dataset and closed-loop benchmark for deformation-aware pick-and-place. The dataset provides 4,000 expert demonstrations over 40 tasks, 2,000 on ten volumetric deformable objects and 2,000 on their visually matched rigid twins, pairing multi-view RGB with dual-finger tactile RGB and marker motion, synchronized with proprioception, language, dual gripper-action encodings, and evaluator-only FEM states. The benchmark uses DSR as its primary metric, crediting a rollout only when it completes the task within an object-specific deformation tolerance calibrated before policy training, under paired ID and OOD evaluation.

Benchmarking three policy families demonstrates the diagnostic value of this design. All twelve in-distribution configurations contain successful rollouts that exceed the tolerance, accounting for 0.7–24% of each configuration's successes. The matched controls further show that tolerance compliance depends on policy family, that the cost of deformability depends on task structure, that finer gripper control can reproduce an apparent tactile gain, and that visuo-tactile variants attain higher absolute OOD success in all six comparisons despite mixed in-distribution effects. These results suggest that making touch available does not by itself yield effective multimodal fusion, while FastWAM's small TSR–DSR gap shows that task completion and tolerance compliance can coincide.

Limitations. SoftVTBench is simulation-only. Its tactile observations come from component simulators validated in their original settings, but no real-sensor comparison is performed on our assets, and the linear-elastic FEM model captures recoverable deformation rather than material damage; the tolerance is therefore an operational criterion, not a measured damage threshold. Characterizing the simulation-to-real tactile gap and grounding the tolerances in measurements of real food items are left to future work; because the released records store full deformation traces, any recalibrated criterion can be applied without new rollouts.

AI USE STATEMENT

In this work, we used generative AI tools for limited assistance with manuscript editing, language refinement, and figure polishing. We did not use generative AI tools to generate experimental results or replace human judgment in the scientific analysis. We have reviewed all AI-assisted work and take full responsibility for the final content of this work, including all text, claims, results, and artifacts.

REPRODUCIBILITY STATEMENT

We describe the dataset construction, simulation environment, tactile sensing pipeline, and per-object deformation calibration in Section 3 and Appendices B–C, and the task configurations, evaluation protocol, and metric definitions in Section 4 and Appendix D. Baseline implementations, including training hyperparameters, inference settings, and hardware configurations, are given in Appendix E. Additional analyses and the scope of tactile simulation are provided in Appendices F–G.

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.

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.

Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 2023.

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.

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.

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.

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.

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.

Qingtao Liu, Yu Cui, Zheng Sun, Gaofeng Li, Jiming Chen, and Qi Ye. Vtdexmanip: A dataset and benchmark for visual-tactile pretraining and dexterous manipulation with reinforcement learning. In *ICLR*, 2025.

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.

Mayank Mittal et al. Isaac lab: A gpu-accelerated simulation framework for multi-modal robot learning. *arXiv preprint arXiv:2511.04831*, 2025. URL https://arxiv.org/abs/2511.04831.

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.

NVIDIA Corporation. NVIDIA Isaac Sim. https://github.com/isaac-sim/IsaacSim, 2025. Version 4.5.0.

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.

Neha Sunil, Shaoxiong Wang, Yu She, Edward Adelson, and Alberto Rodriguez Garcia. Visuotactile affordances for cloth manipulation with local control. In *Conference on Robot Learning*, pp. 1596–1606. PMLR, 2023.

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.

Youngsun Wi, Andy Zeng, Pete Florence, and Nima Fazeli. Virdo++: Real-world, visuo-tactile dynamics and perception of deformable objects. *arXiv preprint arXiv:2210.03701*, 2022.

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.

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

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.

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.

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 Overview

This appendix provides implementation details, per-object calibration results, task-suite specifications, baseline training settings, additional analyses, and the scope of tactile simulation. Table 6 summarizes the contents in section order.

Table 6: Contents of the appendix in document order.

Section	Contents
Sec. B: Simulation and implementation details	Simulator, robot, sensing, FEM, and compute configuration.
Sec. C: Assets and deformation tolerances	Asset inventory, rigid twins, per-object calibration, and the calibration percentile.
Sec. D: Task suite details	Suite design, task mapping, and variation factors.
Sec. E: Baseline training details	Training and inference settings for all evaluated policy families.
Sec. F: Additional analyses	Qualitative cases and stiffness-conditioned deformation statistics.
Sec. G: Tactile simulation pipeline and scope	Tactile rendering stages, recorded outputs, and validity boundary.

B Simulation and Implementation Details

SoftVTBench is implemented in Isaac Sim 4.5.0 with Isaac Lab 0.41.3 and the GPU-accelerated PhysX 5 pipeline. Physics runs at 60 Hz with a control decimation of 3, giving a 20 Hz control and logging rate; all visual, tactile, proprioceptive, action, and evaluator-only physical-state streams are synchronized at this rate. Table 7 summarizes the robot, control, sensing, simulation, and hardware configuration.

Table 7: 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	$4\times$ NVIDIA L20 GPU
Training	NVIDIA A100-80GB; per-family counts in Table 11

C Assets and Per-Object Deformation Tolerances

Deformable assets. The dataset contains ten volumetric deformable assets: six bakery-style meshes and four procedurally generated geometric primitives. The surface meshes are converted into simulation-ready volumetric meshes and the deformable instances are simulated as PhysX GPU FEM soft bodies with corotational linear elasticity. Per-asset density, friction, elasticity, and damping are authored in the simulation assets and are never provided to the policy. Before each episode the object is settled on the support surface; the expert and the evaluator both use this settled state rather than the nominal spawn pose, so that pre-grasp drift does not contaminate either data collection or scoring.

Rigid twins. Each rigid twin shares the mesh, texture, and mass of its deformable counterpart, with stiffness raised so that deformation is negligible under any achievable grasp. Twins populate the OBJECT-RIGID and SPATIAL-RIGID suites under the same layouts, instructions, and recording stack as their soft counterparts.

Per-asset calibration. Table 8 reports, for each of the ten deformable assets, the calibrated gripper envelope $\mathcal{G}_o = [g_o^{\min}, g_o^{\max}]$, its physical aperture span, and the normalized deformation tolerance τ_o. The envelope endpoints are the per-object compression ratios recorded in the locked collection cards (larger values denote tighter grasps). The millimeter span is computed directly as the difference between the card's loose and tight measured jaw apertures; it is not inferred from a universal linear command-to-width conversion.

Table 8: Calibrated gripper envelopes and deformation tolerances of the ten deformable assets. τ_o is the object-specific 90th-percentile threshold used by the released evaluator, expressed as a percentage of the reference bounding-box diagonal.

Asset	$g_o^{\min}$	$g_o^{\max}$	**Span (mm)**	τ_o/diag. (%)
`soft_pastry001`	0.4000	0.8000	26.21	7.6
`soft_pastry002`	0.6547	0.6900	2.56	9.3
`soft_pastry003`	0.4400	0.8000	13.26	8.6
`soft_pastry005`	0.4700	0.5800	5.87	10.7
`soft_pastry010`	0.4200	0.6600	24.82	8.3
`soft_pastry011`	0.4400	0.5800	10.63	7.1
`soft_stw_cube_hq`	0.3200	0.6200	19.43	11.2
`soft_stw_cuboid_hq`	0.3800	0.4400	10.95	9.8
`soft_stw_cylinder_hq`	0.2900	0.6600	21.95	9.8
`soft_stw_sphere_hq`	0.3700	0.6300	20.77	9.7

Calibration percentile. All thresholds in Table 8 use the 90th percentile of peak displacement over the stable scripted grasps of the calibration sweep. The percentile is a single global constant: it is fixed once, before any policy is trained, and the same value is applied to every asset, so it cannot be tuned per object or per method. A lower percentile tightens every threshold uniformly and a higher one loosens them, which rescales $R_{\max}$ by a per-object constant and can only change a policy's score through episodes whose peak deformation lies between the two thresholds. Because the released evaluation records store the per-episode deformation trace R_t rather than only the binary outcome, scoring at another percentile is a re-aggregation of the released records and requires no new rollouts; the accompanying code package includes the rescoring entry point for this purpose.

D TASK SUITE DETAILS

The four suites form a matched 2×2 design over object type (deformable or rigid twin) and variation axis (object identity or spatial layout). Each suite contains 10 tasks with 100 expert demonstrations per task, for 1,000 demonstrations per suite and 4,000 in total. All suites share the same robot embodiment, camera views, tactile interfaces, rollout API, and evaluation protocol.

OBJECT-SOFT fixes the scene layout and varies the manipulated object across tasks, so that the policy must adapt its grasp to each object's geometry, compliance, and contact surface. SPATIAL-SOFT places two visually identical instances of the same object in each scene and identifies the target through a spatial referring expression in the language instruction, so that the scene alone does not determine the referent. The success predicate checks the identity of the transported instance and requires that the instructed one, not merely some instance, reaches the target. OBJECT-RIGID and SPATIAL-RIGID replicate both structures with the corresponding rigid twins.

Task list. Table 9 lists the tasks of the two deformable suites; the rigid suites replicate them with the corresponding twins under identical layouts and instructions.

Table 9: Task specifications for the deformable suites. Task identifiers follow the released suite configuration and are zero-indexed; instructions are reproduced verbatim from that configuration, and object descriptions are therefore not normalized across suites.

Suite	Task	Instruction	Asset
OBJECT-SOFT	0	Pick up the white swirled pastry and place it in the basket.	`soft_pastry001`
	1	Pick up the panda-face pastry and place it in the basket.	`soft_pastry002`
	2	Pick up the small chocolate rectangular pastry and place it in the basket.	`soft_pastry003`
	3	Pick up the soft cream-colored cube and place it in the basket.	`soft_stw_cube_hq`
	4	Pick up the yellow square layered pastry and place it in the basket.	`soft_pastry005`
	5	Pick up the soft tan cylinder and place it in the basket.	`soft_stw_cylinder_hq`
	6	Pick up the golden pastry and place it in the basket.	`soft_pastry011`
	7	Pick up the golden knotted pastry and place it in the basket.	`soft_stw_cuboid_hq`
	8	Pick up the soft red ball and place it in the basket.	`soft_stw_sphere_hq`
	9	Pick up the orange pumpkin-shaped pastry and place it in the basket.	`soft_pastry010`
SPATIAL-SOFT	0	Pick up the left white swirled pastry and place it on the plate.	`soft_pastry001`
	1	Pick up the right panda-face pastry and place it on the plate.	`soft_pastry002`
	2	Pick up the left small chocolate rectangular pastry and place it on the plate.	`soft_pastry003`
	3	Pick up the left soft cream-colored cube and place it on the plate.	`soft_stw_cube_hq`
	4	Pick up the left yellow square layered pastry and place it on the plate.	`soft_pastry005`
	5	Pick up the left soft tan cylinder and place it on the plate.	`soft_stw_cylinder_hq`
	6	Pick up the right golden pastry and place it on the plate.	`soft_pastry011`
	7	Pick up the left soft cream-colored rectangular block and place it on the plate.	`soft_stw_cuboid_hq`
	8	Pick up the left soft red ball and place it on the plate.	`soft_stw_sphere_hq`
	9	Pick up the left orange pumpkin-shaped pastry and place it on the plate.	`soft_pastry010`

Variation factors. Table 10 lists the training ranges of the variation families and the held-out values used for out-of-distribution evaluation. Only the spatial factors are varied during training; the visual and physical factors are fixed at their nominal values, and every held-out value differs from its nominal value. Each out-of-distribution condition shifts a single factor while holding the others at their nominal values.

Table 10: Variation factors: training ranges and held-out out-of-distribution values. The nine out-of-distribution conditions are the three held-out levels of dome-light intensity, mass, and Young's modulus.

Family	Factor	Training range	Held-out values
Visual	Dome-light intensity	135 (fixed nominal)	67.5, 180, 270
Spatial	Object placement	Per-task recorded support	—
	Target placement	Per-task recorded support	—
	Robot initialization	Per-task recorded support	—
Physical	Mass scale	$\times$1.0 (nominal)	$\times$1.25, $\times$1.75, $\times$2.5
	Young's modulus scale	$\times$1.0 (nominal)	$\times$0.5, $\times$0.8, $\times$2.0

E Baseline Training Details

We evaluate eight policy configurations. Six are primary continuous-control baselines: the vision-only (VO-C) and visuo-tactile (VT-C) variants of Diffusion Policy, $\pi_{0.5}$, and FastWAM. The remaining two, VO-B and VT-B, are binary-control ablations of $\pi_{0.5}$. These ablations differ from their continuous counterparts only in the gripper-action encoding used during training and inference; all other settings remain unchanged.

Unless stated otherwise, each method follows its official backbone, optimizer family, learning-rate schedule, and model-specific inference procedure. Benchmark-specific adaptations are restricted to the observation interfaces—including the tactile encoders and marker-motion adapters described below—the 7D action interface, and micro-batch sizes selected to fit GPU memory. The added tactile streams increase activation memory, so the visuo-tactile variants of Diffusion Policy and FastWAM use a smaller micro-batch than their vision-only counterparts; since no gradient accumulation is applied in either configuration, their effective global batch size differs accordingly (Table 12). The two $\pi_{0.5}$ variants are matched at a global batch size of 256. Comparisons between vision-only and visuo-tactile variants therefore reflect this optimization difference in addition to the observation interface. Each reported configuration corresponds to a single training run and uses the final checkpoint from the fixed schedule listed in Table 11. We apply neither validation-based checkpoint selection nor a shared cross-family hyperparameter sweep. Consequently, differences across policy families may reflect both their architectures and their prescribed training recipes.

E.1 Diffusion Policy (DP)

We retain the image-conditioned diffusion U-Net architecture and DDPM parameterization of Diffusion Policy. The vision-only variant receives two consecutive observations, each comprising a third-person RGB image, a wrist RGB image, and the 7D robot state. Their representations form the global conditioning vector of the U-Net, which denoises a 16-step chunk of 7D actions. DP is not language-conditioned.

The visuo-tactile variant preserves the same policy and action interface while adding left- and right-finger tactile RGB as two image streams. A ResNet-18 image encoder processes each tactile stream, and the resulting representations are fused with the visual features. The two-frame marker-motion representation is additionally concatenated with the robot-state conditioning vector, providing a 792D marker-motion input alongside the 7D robot state. Apart from this observation interface, the VO and VT variants share the same network design and normalization procedure. Their common optimization and inference settings are reported in Table 11.

E.2 $\pi_{0.5}$ POLICY

We initialize both modality variants from the pretrained $\pi_{0.5}$ model. The vision-only (VO) variant receives third-person RGB, wrist RGB, the 7D robot state, and the language instruction. Except for the task-specific observation and action transformations, VO follows the official LoRA fine-tuning architecture and optimization procedure. LoRA adapters are applied to both the PaliGemma backbone and the action expert, while the pretrained backbone parameters remain frozen.

The visuo-tactile (VT) variant retains all VO inputs and additionally receives tactile RGB and marker-motion observations from both fingers. At each policy query, we collect the past eight tactile RGB frames from each finger, yielding 16 frames in total, and tile them temporally into a single 4×4 mosaic. The shared visual encoder processes this mosaic in the same manner as the external RGB observations. This design represents tactile appearance and contact deformation in the visual token space without introducing a separate tactile image backbone. Marker motion is encoded as low-dimensional contact features: the current marker observation is injected into the policy prefix, whereas a two-frame marker-motion history is supplied to the action suffix to condition action generation on recent contact dynamics. The additional tactile projection layers are trained jointly with the LoRA adapters.

Both variants predict a 50-step chunk of 7D actions comprising 3D position, 3D axis-angle orientation, and a 1D gripper command. During closed-loop execution, the policy executes 10 predicted actions before replanning. Unless otherwise specified, each policy is LoRA-fine-tuned for 7k steps on eight NVIDIA A100-80GB GPUs with a global batch size of 256. The complete training and inference configuration is reported in Table 11.

E.3 FASTWAM

We retain the official FastWAM world–action model architecture and flow-matching formulation. In the vision-only setting, the RGB video expert receives the third-person and wrist RGB observations, while the action expert predicts the 7D action sequence. Cached tokens from the frozen T5 encoder represent the language instruction, and a learnable projection of the 7D robot state is appended to the language context through cross-attention.

The visuo-tactile variant extends the original two-expert mixture-of-transformers (MoT) with a third tactile DiT expert. At each observation time, tactile input is constructed as a 2×2 mosaic containing the left and right tactile images from the current and preceding frames. The frozen Wan VAE encodes this mosaic, which is then processed by the tactile expert. The tactile expert reuses the ActionDiT-style pretrained backbone, while its tactile input and output layers are learned for the tactile latent space. For marker motion, we use the current frame together with the preceding two frames. After normalization, a lightweight MLP maps the three marker-motion vectors to temporally indexed context tokens. These tokens are appended only to the text–proprioceptive context of the tactile and action experts, leaving the RGB expert unchanged.

We further use an anchor-only tri-branch attention mask: the action expert can attend to the observed first-frame RGB and tactile tokens but never to future tactile tokens. Consequently, inference-time action prediction depends only on observations available at the current control step. The RGB, tactile, and action branches are jointly optimized with flow matching using loss weights $(1.0, 0.2, 1.0)$, respectively. All remaining FastWAM training and inference settings follow the official configuration and are reported in Table 11.

Table 12 summarizes the differences between the vision-only (VO) and visuo-tactile (VT) variants of Diffusion Policy, $\pi_{0.5}$, and FastWAM in terms of trainable parameters, batch sizes, and input modalities.

Table 11: Training and inference configurations for the three baseline policy families.

Setting	Diffusion Policy	$\pi_{0.5}$	FastWAM
Model and compute			
Policy backbone	Image-conditioned diffusion U-Net	—	—
GPUs	1× NVIDIA A100-80GB	8× NVIDIA A100-80GB	8× NVIDIA A100-80GB
Precision	—	bfloat16 mixed precision	bfloat16
Distributed training	—	FSDP	ZeRO-1
Fine-tuning method	—	LoRA	—
LoRA rank / alpha (backbone)	—	16/16	—
LoRA rank / alpha (action expert)	—	32/32	—
Training and optimization			
Training duration	20 epochs	7k steps	10 epochs
Batch size (micro / global)	—	32/256	—
Gradient accumulation	1	1	1
Optimizer	AdamW	AdamW	AdamW
Learning rate	1×10^{-4}	2.5×10^{-5} (peak)	1×10^{-4}
AdamW betas	$(0.95, 0.999)$	$(0.9, 0.95)$	$(0.9, 0.95)$
Weight decay	1×10^{-6}	1×10^{-10}	1×10^{-2}
LR schedule	Cosine decay; 100-step warmup	Cosine decay; 1k-step warmup	Cosine decay; 5% warmup
Gradient clipping	—	1.0	1.0
EMA	Enabled	Disabled	—
Closed-loop inference			
Denoising timesteps (train / inference)	DDPM: 100/100	—	Flow matching: 1000/10
Action horizon	16	50	32
Actions executed per replanning	8	10	10

Table 12: Differences between the vision-only and visuo-tactile policy variants. Trainable parameter counts include all optimized modules and exclude frozen model weights and preprocessing components.

Policy	Setting	Vision-only (VO)	Visuo-tactile (VT)
Diffusion Policy	Trainable parameters	329.2M	455.7M
	Batch size (micro / global)	256/256	128/128
	Input modalities	RGB; robot state	RGB; robot state; tactile RGB; marker motion
$\pi_{0.5}$	Trainable parameters	467.0M	538.0M
	Batch size (micro / global)	32/256	32/256
	Input modalities	RGB; robot state; language	RGB; robot state; language; tactile RGB; marker motion
FastWAM	Trainable parameters	6.02B	7.05B
	Batch size (micro / global)	24/192	16/128
	Input modalities	RGB; robot state; language	RGB; robot state; language; tactile RGB; marker motion

F Additional Analyses

Qualitative comparison of scored rollouts. Figure 4 illustrates the distinction exposed by process-level evaluation: all displayed policies ultimately place the object successfully, but two trajectories cross the calibrated object-specific deformation limit during interaction. The curve and the three visual observations in each row come from the same synchronized rollout, and the peak-interaction frame is selected at $\arg\max_t R_t$. The tactile marker fields make the contact evolution visible, while the DSR label itself is computed from the FEM-based deformation trace. Because the rows cover different tasks and policy configurations, they provide qualitative coverage rather than a controlled comparison of policy performance. The cases were selected to span three object geometries and both DSR outcomes; they are not used to estimate outcome prevalence.

Deformation measurement under stiffness shift. Table 13 reports TSR and DSR alongside the median and interquartile range of $R_{\max}$ over the complete 100-episode paired subset for each policy, input modality, and stiffness level; the $\times 1.0$ rows use the 100 in-distribution initial states reused by

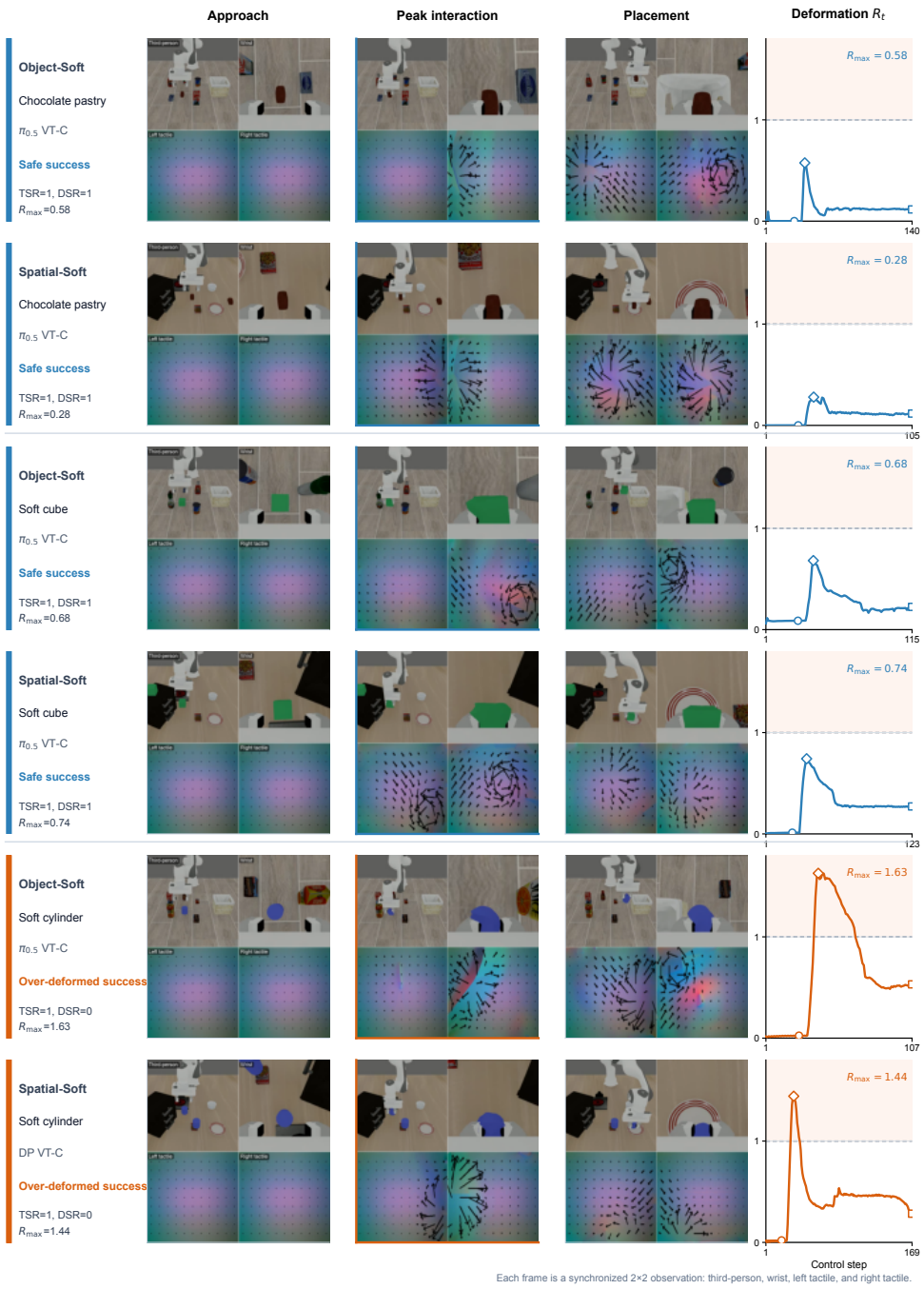

Figure 4: Illustrative rollouts on three deformable objects across OBJECT-SOFT and SPATIAL-SOFT. Each row shows synchronized third-person, wrist, and tactile observations at approach, peak interaction, and placement, followed by the normalized deformation trace R_t. Circle, diamond, and square indicate the displayed steps. All rollouts achieve task success (TSR$=1$). Blue trajectories satisfy the calibrated deformation tolerance (DSR$=1$), while orange trajectories exceed $R_t = 1$ despite success (DSR$=0$). Rows are selected examples from different rollouts, not matched-initial-state comparisons.

the out-of-distribution conditions and therefore differ from the 500-episode results of Table 2. The two axes move independently, and in both directions. At nominal stiffness, the two Diffusion Policy variants complete equally often (TSR 42 for both) while differing by 9 points in DSR and by 0.06 in median $R_{\max}$, so deformation separates configurations that completion ranks as equal. Conversely, softening the object to $\times 0.5$ lowers the completion rate of vision-only Diffusion Policy from 42 to 33 while its median $R_{\max}$ barely moves (0.63 to 0.66), so completion also registers effects that the deformation median does not. Because these are closed-loop rollouts in which the policy may adapt its actions to the perturbed dynamics, we claim only that the two axes are non-redundant under stiffness shift; we do not claim that either axis tracks the other, nor that $R_{\max}$ varies monotonically with stiffness.

Table 13: Completion and deformation statistics under stiffness shift on OBJECT-SOFT (rates in %).

Policy	E scale	Input	TSR	DSR	$R_{\max}$ median [IQR]	N
DP	$\times 1.0$	VO-C	42	39	0.63 [0.47, 0.75]	100
		VT-C	42	30	0.69 [0.48, 0.97]	100
	$\times 0.5$	VO-C	33	32	0.66 [0.06, 0.76]	100
		VT-C	41	30	0.76 [0.57, 0.98]	100
	$\times 0.8$	VO-C	37	31	0.63 [0.06, 0.79]	100
		VT-C	45	33	0.75 [0.54, 0.96]	100
	$\times 2.0$	VO-C	39	37	0.60 [0.37, 0.71]	100
		VT-C	39	32	0.63 [0.41, 0.92]	100
$\pi_{0.5}$	$\times 1.0$	VO-C	45	42	0.46 [0.15, 0.60]	100
		VT-C	41	35	0.63 [0.51, 0.88]	100
	$\times 0.5$	VO-C	34	31	0.56 [0.35, 0.83]	100
		VT-C	41	31	0.71 [0.55, 0.92]	100
	$\times 0.8$	VO-C	45	40	0.50 [0.31, 0.69]	100
		VT-C	42	33	0.68 [0.54, 0.96]	100
	$\times 2.0$	VO-C	38	32	0.46 [0.27, 0.64]	100
		VT-C	42	37	0.59 [0.47, 0.79]	100
FastWAM	$\times 1.0$	VO-C	64	64	0.49 [0.38, 0.59]	100
		VT-C	58	58	0.50 [0.39, 0.57]	100
	$\times 0.5$	VO-C	58	58	0.55 [0.42, 0.63]	100
		VT-C	56	56	0.54 [0.44, 0.60]	100
	$\times 0.8$	VO-C	58	58	0.49 [0.40, 0.62]	100
		VT-C	58	58	0.50 [0.41, 0.57]	100
	$\times 2.0$	VO-C	58	58	0.47 [0.33, 0.58]	100
		VT-C	57	57	0.48 [0.39, 0.57]	100

G TACTILE SIMULATION PIPELINE AND SCOPE

Pipeline. Tactile observations are produced by TacEx (Nguyen et al., 2024), which integrates GelSight-style sensor simulation into Isaac Sim. At each control step, the gel–object interaction is derived from the simulator contact state; Taxim (Si & Yuan, 2022) renders the optical tactile image from the resulting contact geometry using its example-based calibration approach, and FOTS (Zhao et al., 2024) produces the marker-motion field. Each finger yields a 320×240 tactile RGB image and an 11×9 marker field; both fingers are recorded at the 20 Hz control rate, synchronized with all other streams. SoftVTBench retains the TacEx GelSight Mini rendering profile and its Taxim example-based optical calibration rather than fitting a new real-sensor calibration. Markers are arranged as a uniform 11×9 field (99 markers per finger); the released observations store both the rendered RGB contact image and the 2D displacement of every marker. No benchmark-specific physical GelSight calibration is used.

Figure 5 shows the tactile RGB and marker-motion observations at the moment of grasping for the ten OBJECT-SOFT tasks; differences in geometry, compliance, and contact area produce distinct contact patches and shear fields under the same pick-and-place task structure.

Scope. The component simulators were validated in their original publications under their respective conditions: Taxim reports real-to-sim accuracy against a physical GelSight sensor on its calibration objects, and FOTS evaluates marker-motion realism for sim-to-real skill transfer. These published

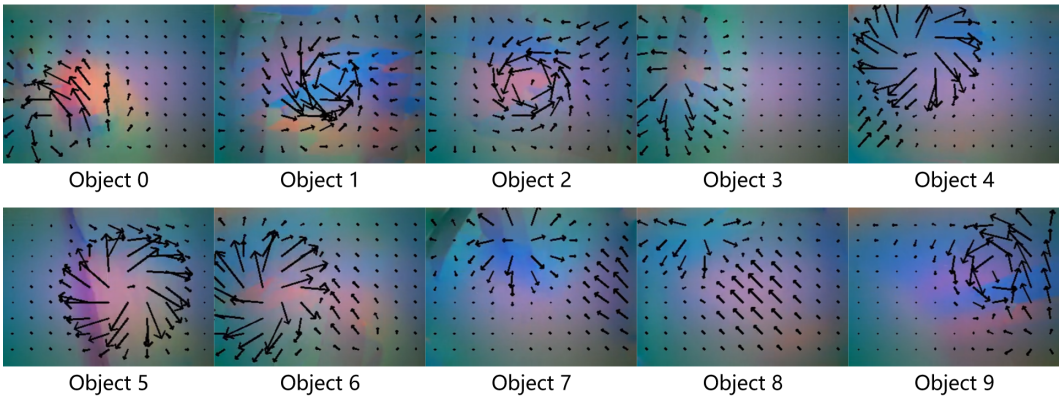

Figure 5: **Tactile observations across deformable objects.** Tactile RGB with marker-motion overlay at the moment of grasping, for the ten OBJECT-SOFT tasks.

validations concern the component methods under their own experimental conditions; they do not constitute a sim-to-real validation of SoftVTBench's specific assets, materials, contact conditions, or rendering parameters. We do not claim validated transfer of SoftVTBench's simulated tactile signals to physical GelSight sensors, and no real-sensor comparison is performed in this work. The benchmark's conclusions are accordingly stated for simulated visuo-tactile policy learning; characterizing the simulation-to-real tactile gap on these assets is left to future work.