RoboTactile: Benchmarking the Robustness of Vision-Tactile Manipulation under Tactile Observation Failures

Anonymous authors
Paper under double-blind review

Abstract

Tactile perception plays a vital role in precise robotic manipulation. Despite their strong performance, existing vision-tactile models may suffer from tactile observation failures caused by contact-induced wear, sensor noise, and signal delays, raising concerns about the reliability of robotic systems. However, existing benchmarks primarily focus on performance under ideal tactile conditions, leaving robustness to tactile failures insufficiently explored. To systematically evaluate this robustness, we introduce **RoboTactile**, a benchmark built on UniVTAC and its Isaac Sim environment. Grounded in sensing and deployment failure mechanisms, we design 14 controlled tactile observation faults spanning four dimensions: **Observation Availability**, **Signal Fidelity**, **Temporal Integrity**, and **Contextual Consistency**. The benchmark records contact-phase annotations, with restored-input evaluation defined as an extension. RoboTactile establishes track-specific protocols for offline prediction and closed-loop manipulation across world models, world–action models, policies, and vision–language–action models. Experiments on four models reveal substantial task- and fault-dependent variation. We report absolute success, performance gaps, and descriptive tactile gain retention relative to available visual references, while distinguishing these from verified matched-control estimates. Matched visual controls, phase-conditioned evaluation, and post-restoration evaluation remain future work. We hope RoboTactile will support the development and evaluation of robust vision-tactile models.

1 Introduction

Vision-tactile manipulation combines visual scene understanding with local contact sensing to support precise robotic tasks such as grasping, insertion, and object handling. Vision provides information about object appearance, spatial configuration, and the surrounding environment, whereas touch provides complementary evidence about contact geometry, incipient slip, and interaction forces that can be difficult to infer from vision alone. Recent work exploits this complementarity through diverse modeling approaches, including vision-tactile representation and policy learning in UniVTAC (Chen et al., 2026), world modeling in ContactWorld (Zhang et al., 2026), joint prediction of contact and action in N_0-TWAM (NeoteAI Team & Fudan TEAI Team, 2026a), and tactile-conditioned vision–language–action modeling in DreamTacVLA (Ye et al., 2026b). These advances demonstrate the potential of touch for contact-rich manipulation, but exploiting tactile information also makes its reliability an important concern. Unlike cameras that observe the scene from a distance, tactile sensors operate at the physical interaction interface, where repeated contact places demands on sensor durability and signal quality (Donlon et al., 2018). Evaluating how models respond when tactile observations become unreliable is therefore essential to understanding their suitability for deployment.

Despite strong performance under clean inputs, vision-tactile models may encounter tactile-stream interruptions, degraded signals, delayed frames, or misattributed observations during operation. Research on GelSlim highlights the importance of durability and stable illumination under repeated grasping (Donlon et al., 2018), while Digger Finger demonstrates how particles trapped at the sensing interface can distort observed contact geometry (Patel et al., 2021). Importantly, sensor failures need not produce obviously invalid outputs: even a broken sensor can return valid-looking values,

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.

Table 1: **Positioning of RoboTactile among tactile and vision-tactile manipulation benchmarks.** The table summarizes task coverage and evaluation scope. RoboTactile standardizes on GelSight Mini and focuses on controlled tactile observation failures rather than sensor or task breadth. Check marks (✓) denote benchmark-level capabilities, not isolated ablations; △ = partial support or a specified extension, "N/R" = not reported, "on-pol." = on-policy (no fixed corpus). "# Seq." counts synchronized trajectories, not frames, streams, or corrupted derivatives; – denotes an unregistered count.

Benchmark	Benchmark scope					Tactile-failure robustness			
	Setting	# Tasks	# Seq.	Tactile sensor(s)	Closed loop	Failure axis	Contact phase	Vision-only harm	Recovery
VTDexManip	Sim	6	2,032	Binary pressure	✓	–	–	–	–
ManiFeel	Sim	9	560	RGB + FF (sim)	✓	–	–	–	–
UniVTAC	Sim	8	400	Optical (3)	✓	–	–	–	–
TacO	Real	3	N/R	4 real types	✓	–	–	–	–
ContactWorld	Sim	12	1,872	RGB/Depth/FF	✓	–	–	–	–
Tactile Genesis	Sim	3	on-pol.	7 abstractions	✓	✓	–	–	–
SoftVTBench	Sim	40	2,000	RGB marker	✓	–	–	–	–
RoboTactile	Sim	8	–	GelSight Mini (1)	✓	✓	△	△	△

motivating cross-modal fault detection and compensation (Lee et al., 2020). For tactile manipulation, a stale contact image or a signal assigned to the wrong finger may remain plausible in isolation while being inconsistent with the current interaction. If a model continues to rely on such observations, touch may become a source of misleading evidence, potentially reducing performance below a matched vision-only baseline. Clean tactile gains and robustness to tactile failures must therefore be evaluated separately.

Table 1 summarizes seven representative tactile and vision-tactile benchmarks: VTDexManip (Liu et al., 2025), ManiFeel (Luu et al., 2025), UniVTAC (Chen et al., 2026), TacO (Zorin et al., 2026), ContactWorld (Zhang et al., 2026), Tactile Genesis (Chung et al., 2026), and SoftVTBench (Jing et al., 2026). These benchmarks provide valuable resources for studying tactile representations and manipulation policies. Prior work also investigates selected robustness factors: VTDexManip examines tactile noise and sensing thresholds, while Tactile Genesis studies tactile sensing abstractions and sensor design choices. These efforts motivate a complementary benchmark that jointly evaluates controlled failure types, contact phases, performance relative to vision-only operation, and recovery after valid touch returns. Corruption benchmarks in image recognition and autonomous-driving perception provide a methodological foundation for systematic evaluation (Hendrycks & Dietterich, 2019; Dong et al., 2023). However, tactile manipulation introduces additional dependencies on contact and action. The same delay may have little effect during free motion but disrupt control at contact onset, and a fault-induced action changes the subsequent state and observations. Static input-corruption evaluation alone cannot capture these effects or determine whether a system recovers when the tactile stream is restored. A suitable benchmark must therefore connect controlled observation faults with contact phase, closed-loop execution, and recovery.

In this work, we introduce RoboTactile, a benchmark for systematically evaluating the robustness of vision-tactile models to tactile observation failures. Grounded in failure mechanisms documented in tactile sensing and deployment, we organize **fourteen controlled fault operators** along four dimensions: *observation availability*, *signal fidelity*, *temporal integrity*, and *contextual consistency*. These dimensions ask whether touch is present, informative, current, and consistent with the active sensor and contact event. The taxonomy concerns the observations delivered to the model rather than unobserved hardware root causes, and includes both explicit signal loss and plausible-but-wrong observations. It also distinguishes sensor failure from valid no-contact observations, which may contain little visual structure without being faulty. Built on the UniVTAC platform in Isaac Sim, RoboTactile defines evaluation protocols over eight manipulation tasks, with recorded contact phases and deterministic exposure windows. Event-triggered schedules, fault composition, and independent restored-input evaluation are specified extensions. Faults are applied only at test time, after clean tactile rendering and before model-specific preprocessing, without modifying simulator physics. This design provides reproducible interventions while separating observation failure from changes to the physical task.

RoboTactile defines separate tracks for world models, world–action models, policies, and vision–language–action models, preserving native prediction and control interfaces. The protocols specify offline prediction for world models, closed-loop manipulation for policies and VLAs, and both for world–action models. Offline evaluation assesses prediction of clean future observations from faulty inputs. The paired closed-loop protocol compares task outcomes under matched *clean*, *faulty*, and qualified *vision-only* conditions, sharing initial states and exogenous random seeds. A separate extension specifies post-restoration comparisons. Current experiments report clean and faulty task success, absolute performance gaps, and descriptive tactile gain retention relative to available visual references, not verified matched-control estimates, which remain future work. For world–action models, dual-mode evaluation supports relating offline prediction degradation to closed-loop robustness. These protocols aim to support failure analysis and the development of more reliable vision-tactile systems. Our contributions are threefold:

1. **A tactile failure taxonomy.** We define fourteen controlled fault operators across four observation-contract dimensions.
2. **A unified robustness benchmark.** We introduce RoboTactile for offline and closed-loop evaluation of tactile failures, with phase annotations and a restored-input extension.
3. **A task-level robustness evaluation.** We analyze four models using task–operator success, performance gaps, and visual-reference retention, with explicit coverage and reference qualification.

2 Related Work

2.1 Vision-Tactile Manipulation

Vision-tactile manipulation exploits the complementarity between visual scene information and local contact sensing to improve perception and decision-making in contact-rich tasks. UniVTAC investigates tactile representation and policy learning (Chen et al., 2026), while ContactWorld studies representation and planning in vision–tactile world models (Zhang et al., 2026). Dream-Tac, N_0-TWAM, and ViTacWorld incorporate touch into future observation prediction or joint action modeling (Lou et al., 2026; NeoteAI Team & Fudan TEAI Team, 2026a; Huang et al., 2026a), and DreamTacVLA integrates tactile prediction into language-conditioned manipulation (Ye et al., 2026b). These studies primarily investigate how to use touch effectively. Our work instead examines model behavior when tactile observations fail, measuring task success and performance degradation across tactile fault conditions.

2.2 Robustness Benchmarks

VTDexManip, ManiFeel, and UniVTAC provide benchmarks for vision-tactile representations and manipulation policies (Liu et al., 2025; Luu et al., 2025; Chen et al., 2026), while TacO, SoftVTBench, and HT-Bench focus on sensor comparison, deformable-object manipulation, and full-hand tactile representations, respectively (Zorin et al., 2026; Jing et al., 2026; Huang et al., 2026b). Selected robustness factors have also been studied: VTDexManip evaluates tactile noise, and Tactile Genesis investigates different sensing abstractions and noise settings (Liu et al., 2025; Chung et al., 2026). In other domains, ImageNet-C, 3D detection corruption benchmarks, and RoboBEV evaluate robustness through controlled corruptions (Hendrycks & Dietterich, 2019; Dong et al., 2023; Xie et al., 2023), while MultiBench and MUTE-Bench study modality availability (Liang et al., 2021; Zheng et al., 2026). However, existing evaluations do not jointly examine diverse tactile observation failures, contact phases, performance degradation relative to matched vision-only baselines, and recovery under a unified protocol. RoboTactile addresses this gap through controlled fault injection and protocols for offline prediction and paired closed-loop manipulation. The present results cover aggregate closed-loop success.

2.3 Failure Mechanisms in Tactile Sensing

Tactile sensors are subject to drift, hysteresis, nonlinearity, and cross-talk, with effects that depend on the sensing technology and mounting conditions (Sánchez-Durán et al., 2015). GelSlim studies durability and signal quality under repeated grasping (Donlon et al., 2018), Digger Finger demonstrates

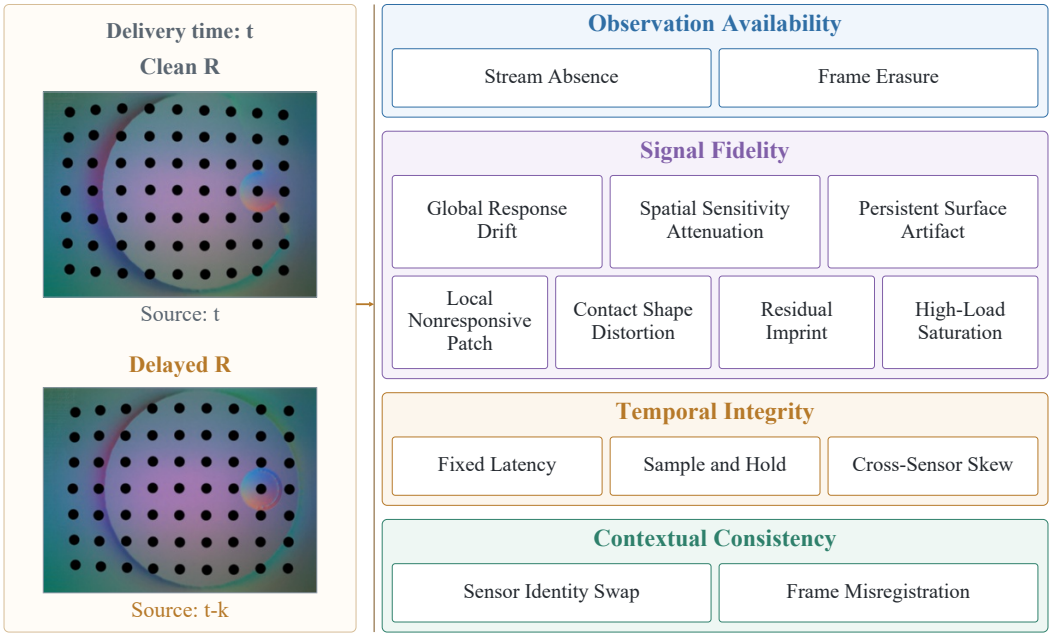

Figure 1: **Fourteen tactile observation faults with a representative example.** Left: clean and delayed observations from the same right sensor at delivery time t. The clean input originates at t, whereas Fixed Latency delivers an earlier frame from $t - k$, with $k = 48$ observations in this example. A single large arrow points to the complete taxonomy rather than to one dimension. Right: the fourteen operators span Observation Availability (2), Signal Fidelity (7), Temporal Integrity (3), and Contextual Consistency (2). Images are from an offline fault replay of a recorded UniVTAC Pull Out Key episode (Chen et al., 2026).

contact-shape distortion caused by trapped particles (Patel et al., 2021), and work on multi-sensor acquisition emphasizes synchronization and low latency (Wang et al., 2024). Moreover, Lee et al. (2020) show that faulty sensors can continue to return valid-looking values, motivating cross-modal detection and compensation. These studies provide a foundation for modeling tactile failures. We translate the relevant mechanisms into controlled, reproducible interventions at the model's observation interface and organize robustness evaluation along four dimensions: observation availability, signal fidelity, temporal integrity, and contextual consistency.

3 Tactile Failures in Vision-Tactile Robot Learning

Real tactile failures can arise from diverse factors at the interaction interface, including wear and contamination from repeated contact (Donlon et al., 2018; Patel et al., 2021), illumination and calibration drift (Donlon et al., 2018; Ren et al., 2025), acquisition latency and temporal desynchronization (Wang et al., 2024), and sensor wiring or registration errors (Lee et al., 2020; Si & Yuan, 2022). We organize these failures according to how they affect the observations delivered to the model into four dimensions: **Observation Availability**, **Signal Fidelity**, **Temporal Integrity**, and **Contextual Consistency**, as illustrated in Figure 1.

Observation Availability. Acquisition or communication failures during deployment may interrupt one or more tactile streams or cause intermittent frame loss (Wang et al., 2024; Lee et al., 2020). Sustained interruptions deprive the model of the corresponding contact feedback over an extended interval, while intermittent losses may omit observations of critical contact transitions. Prior work on cross-modal compensation explores using the remaining sensory information to handle corrupted inputs (Lee et al., 2020). To evaluate whether vision-tactile models can maintain prediction and manipulation performance during tactile-stream interruptions, we define two fault operators.

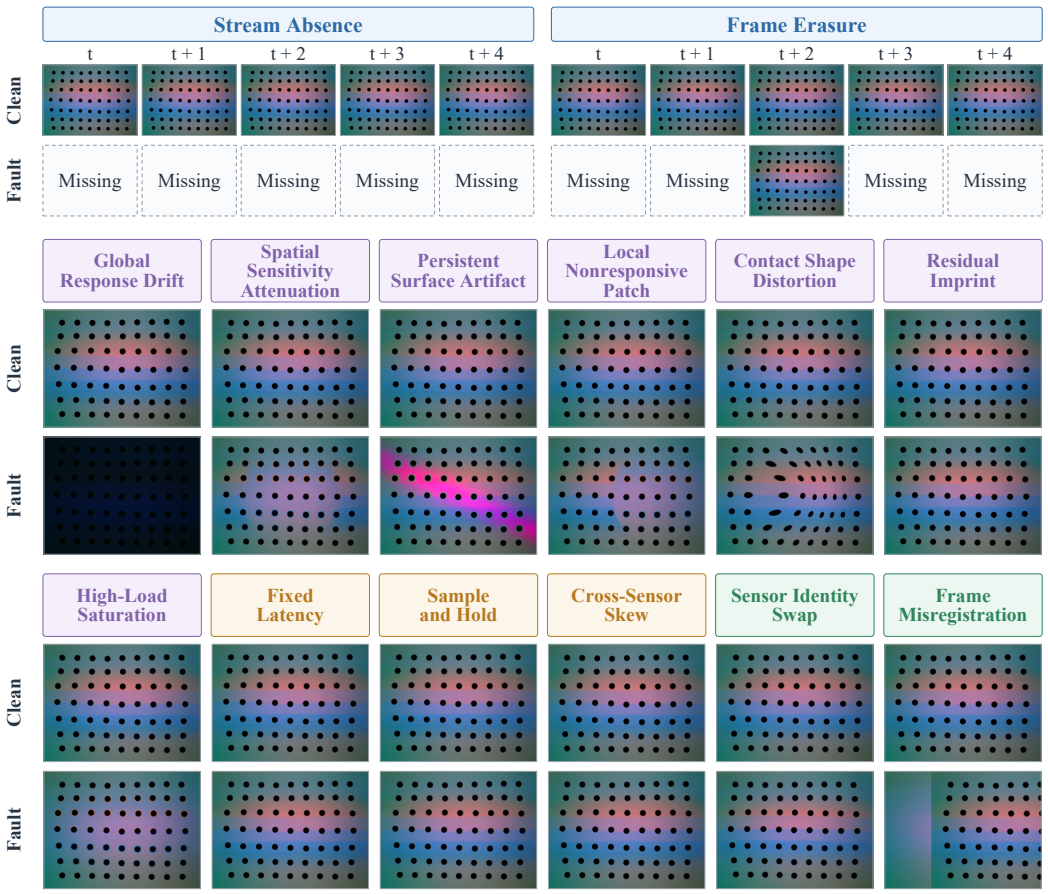

Figure 2: **Qualitative examples of fourteen tactile observation faults.** Stream Absence and Frame Erasure use five consecutive frames; other operators pair clean and faulty observations at matched delivery times. All examples use the right sensor of a recorded UniVTAC Grasp Classify episode. Fault shows diagnostic configurations; interruption markers denote suspended tactile-frame delivery.

- **Stream Absence** interrupts tactile-frame delivery from selected sensor channels over a specified interval, simulating a sustained interruption of the tactile stream.
- **Frame Erasure** removes a predefined set of tactile frames while leaving the remaining frames unchanged, simulating isolated or repeated frame loss.

Both operators can interrupt one or both tactile sensor streams. These transmission faults are distinguished from an all-black image or a valid no-contact observation.

Signal Fidelity. Tactile observations remain available and temporally aligned, but their task-relevant content is degraded. Optical tactile quality depends on elastomer condition, surface contamination, and illumination. Repeated loading can damage the sensing surface and affect contact prediction (Davis & Stuart, 2025; Donlon et al., 2018), trapped particles can distort contact geometry (Patel et al., 2021), and illumination changes can degrade tactile image quality (Ren et al., 2025). Sensor responses may also exhibit drift, hysteresis, and nonlinearity (Sánchez-Durán et al., 2015; Chung et al., 2026). To systematically evaluate robustness to tactile content degradation, we define seven fault operators.

- **Global Response Drift** modifies the overall sensor response.
- **Spatial Sensitivity Attenuation** weakens responses in localized regions.
- **Persistent Surface Artifact** introduces persistent structured artifacts.

- **Local Nonresponsive Patch** fixes the local optical response at its calibrated rest value while preserving current markers.
- **Contact Shape Distortion** warps contact geometry.
- **Residual Imprint** introduces residual responses from previous contacts.
- **High-Load Saturation** approximates saturation of strong optical responses.

These operators cover global, local, and history-dependent signal degradation. Their mathematical definitions and contact-phase activation rules are provided in Appendix.

Temporal Integrity. Contact states evolve with robot actions, making timely tactile feedback essential for tracking the current interaction. In deployment, transmission delays, stalled updates, and asynchronous sensor acquisition may cause models to predict and act on outdated or temporally inconsistent contact information. A prior multi-sensor grasping experiment showed that artificially introduced tactile synchronization errors caused the grasping force to exceed the prescribed threshold (Wang et al., 2024). To systematically evaluate the effects of these temporal errors on prediction and closed-loop manipulation, we define three fault operators.

- **Fixed Latency** replaces the current observation from each selected tactile channel with a frame from a fixed interval earlier, simulating a persistent transmission or processing delay.
- **Sample and Hold** repeatedly delivers the last frame before fault onset for a specified interval, simulating a stalled tactile stream.
- **Cross-Sensor Skew** delays one tactile sensor relative to the other, causing their observations to represent contact states at different times.

We record the source frame used for each input to enable frame-by-frame reproduction of the temporal faults. Mathematical definitions and contact-phase activation rules are provided in Appendix. Random timing jitter is not treated as a separate core fault operator.

Contextual Consistency. Vision-tactile manipulation requires tactile observations to be associated with the correct sensor and contact location. Optical tactile systems rely on calibration to establish the correspondence between images and contact geometry (Donlon et al., 2018; Si & Yuan, 2022). When observations are assigned to the wrong sensor or spatially misaligned with the established calibration, even clear and timely tactile images may cause models to associate contact information with the wrong finger or location. To evaluate the effects of these sensor-assignment and spatial-alignment errors on prediction and manipulation, we define two fault operators.

- **Sensor Identity Swap** delivers the opposite sensor's current observation through the target channel while retaining the target channel's identity, simulating incorrect sensor assignment.
- **Frame Misregistration** applies a deterministic pixel translation to the tactile image relative to the established calibration frame, simulating an error in the registered image-to-sensor mapping.

The two operators alter the sensor assignment and spatial correspondence of current tactile observations, respectively. Their mathematical definitions and implementation details are provided in Appendix.

Figure 2 illustrates the fourteen operators through paired clean and faulty tactile observations under controlled diagnostic configurations. Consecutive-frame views distinguish sustained stream absence from intermittent frame erasure.

4 THE ROBOTACTILE BENCHMARK

We build RoboTactile on UniVTAC running in Isaac Sim (Chen et al., 2026) to evaluate the robustness of vision-tactile models under controlled tactile observation failures. The benchmark instantiates the fourteen fault operators defined in Section 3 and defines matched clean tactile, faulty tactile, and qualified vision-only comparisons. Restored-input evaluation is a separately specified extension.

4.1 BENCHMARK CONSTRUCTION

RoboTactile evaluates the eight vision–tactile manipulation tasks defined by UniVTAC (Chen et al., 2026). The task set spans pose reasoning, shape perception, and contact-rich interaction. Detailed task settings and operations are provided in Appendix.

Table 2 summarizes clean-condition and fault-averaged task success for ACT (Zhao et al., 2023), FTP-1 (Yuan et al., 2026a), N_0-TWAM (NeoteAI Team & Fudan TEAI Team, 2026a), and N_0-VTLA (NeoteAI Team & Fudan TEAI Team, 2026b). FTP-1 covers six reported tasks; the other models cover all eight. The clean N_0-TWAM and N_0-VTLA values are taken from their respective papers and serve as external references rather than matched RoboTactile re-evaluations. These success rates do not by themselves isolate the contribution of touch, which requires qualified matched vision-only controls.

Published no-tactile baselines include ACT, $\pi_{0.5}$ (Physical Intelligence et al., 2025), StarVLA-α (Ye et al., 2026c), InternVLA-A1 (Cai et al., 2026a), Xiaomi-Robotics-0 (Cai et al., 2026b), GigaWorld-Policy (Ye et al., 2026a), LingBot-VA (Li et al., 2026), and FastWAM (Yuan et al., 2026b). We use ACT, VLA, and GigaWorld-Policy results from NeoteAI Team & Fudan TEAI Team (2026b) (Table 1), and LingBot-VA and FastWAM results from NeoteAI Team & Fudan TEAI Team (2026a) (Table 2), rounding task averages to integers.

We use the GelSight Mini tactile representation and inject faults after clean tactile rendering but before model-specific preprocessing, without changing simulator physics. The evaluator records four contact phases: free motion, contact onset, sustained contact, and release. Blind and declared settings specify whether the model receives explicit fault information. The current implementation supports deterministic exposure windows and operator-specific phase gates. Generic event-triggered schedules, fault composition, and independent restored-input campaigns require separate implementation and validation. Unsupported operator–phase–sensor combinations are excluded from scoring.

4.2 EVALUATION PROTOCOLS AND METRICS

Evaluation scope. RoboTactile preserves each model family's native interface and compares results only within compatible tasks and protocols. The protocols cover offline prediction of clean future observations for world models, offline and closed-loop evaluation for world–action models, and closed-loop manipulation for policies and vision–language–action models. Valid matched vision-only counterparts serve as controls when available but are not required for clean and faulty evaluation. The current results report closed-loop success and descriptive comparisons with supplied visual references; verified matched-control, offline, and restoration evaluations are not included. Adapter requirements and paired execution procedures are detailed in Appendix.

Task success and degradation. We report closed-loop task success as percentages. Let $S_{m,u,c}$ denote the reported success of model m on task u under operator c. For an operator set $\mathcal{A}$, define $\overline{S}_{m,u}(\mathcal{A}) = |\mathcal{A}|^{-1} \sum_{c \in \mathcal{A}} S_{m,u,c}$. With nearest-integer rounding $R(x) = \lfloor x + \frac{1}{2} \rfloor$, fault success and the Clean-to-fault gap are

$$\begin{aligned} Q^{\text{fault}}_{m,u}(\mathcal{A}) &= R\big(\overline{S}_{m,u}(\mathcal{A})\big)\,, \\ \Delta Q_{m,u}(\mathcal{A}) &= R(S^{\text{clean}}_{m,u}) - Q^{\text{fault}}_{m,u}(\mathcal{A}). \end{aligned} \tag{1}$$

Table 2 weights all fourteen operators equally. Positive ΔQ indicates lower success under faults, measured in percentage points (pp). The Avg. gap subtracts separately rounded Clean and fault averages; gaps based on external Clean scores are reference gaps rather than matched estimates of degradation.

Tactile gain retention. For a model-specific matched vision-only success rate $S^{\text{VO}}_{m,u}$, tactile gain retention (TGR) measures the fraction of the clean tactile advantage retained under faults:

$$\text{TGR}_{m,u}(\mathcal{A}) = \frac{\overline{S}_{m,u}(\mathcal{A}) - S^{\text{VO}}_{m,u}}{S^{\text{clean}}_{m,u} - S^{\text{VO}}_{m,u}}. \tag{2}$$

TGR equals 1 when the clean advantage is retained, 0 at vision-only performance, and becomes negative below that reference; values are not clipped. The operator set $\mathcal{A}$ can represent an individual

Table 2: **Task success, degradation, and reference tactile gain retention.** Success and 14-fault means are integer percentages; ΔQ (pp) follows Eq. (1). $\mathrm{TGR}^{\mathrm{ref}}$ applies Eq. (2) to the supplied vision-only references, using operator cells before summary rounding. It is a descriptive approximation, not a verified matched-control estimate. Its Avg. covers six eligible tasks per policy; other averages cover eight tasks except FTP-1 (six). N/A denotes an absent visual reference or clean gain ≤ 5 pp; – denotes an unreported score. The N_0 Clean scores and published baselines are external references. Full calculations appear in Appendix.

Method	Measure	Task-level results								Avg.
		Insert HDMI	Insert Hole	Insert Tube	Grasp Classify	Lift Can	Lift Bottle	Pull Out Key	Put Bottle in Shelf	
Published baselines without tactile input										
$\pi_{0.5}$	Clean	8	25	74	49	6	100	35	34	**41**
StarVLA-α	Clean	24	52	69	68	65	32	51	88	**56**
InternVLA-A1	Clean	12	93	95	86	90	58	66	37	**67**
Xiaomi-Robotics-0	Clean	69	96	98	45	13	21	80	12	**54**
GigaWorld-Policy	Clean	0	12	9	20	0	38	32	21	**17**
LingBot-VA	Clean	38	42	96	17	0	0	58	0	**31**
FastWAM	Clean	19	66	98	72	0	21	73	35	**48**
Policies evaluated under tactile faults										
ACT	*Vision-only*	15	19	45	50	20	42	28	28	**31**
	Clean	28	24	56	99	29	71	46	31	**48**
	14-fault mean	25	16	47	80	27	49	43	23	**39**
	ΔQ (pp) $\downarrow$	3	8	9	19	2	22	3	8	**9**
	$\mathrm{TGR}^{\mathrm{ref}}$ $\uparrow$	0.77	N/A	0.16	0.62	0.75	0.23	0.85	N/A	**0.56**
FTP-1	*Vision-only*	–	47	72	–	26	77	30	19	**45**
	Clean	–	64	79	–	65	97	48	47	**67**
	14-fault mean	–	42	60	–	57	75	41	41	**53**
	ΔQ (pp) $\downarrow$	–	22	19	–	8	22	7	6	**14**
	$\mathrm{TGR}^{\mathrm{ref}}$ $\uparrow$	–	-0.32	-1.78	–	0.81	-0.10	0.62	0.79	**0.00**
N_0**-TWAM**	Clean	58	93	79	87	98	68	99	94	**85**
	14-fault mean	52	83	77	84	86	52	95	83	**76**
	ΔQ (pp) $\downarrow$	6	10	2	3	12	16	4	11	**9**
N_0**-VTLA**	Clean	25	95	99	100	88	99	99	60	**83**
	14-fault mean	21	78	97	98	72	82	84	41	**71**
	ΔQ (pp) $\downarrow$	4	17	2	2	16	17	15	19	**12**

operator, a fault dimension, or all fourteen operators. We compute per-task ratios before macro-averaging, requiring complete fault coverage and a clean advantage strictly above five pp.

The supplied ACT and FTP-1 visual references do not establish matched training and evaluation. Their ratios are therefore labeled $\mathrm{TGR}^{\mathrm{ref}}$, computed from reported integer operator scores before summary rounding, and interpreted descriptively rather than as evidence of causal tactile benefit or harm. Neither N_0 model has a model-specific visual reference, so its retention is N/A. Aggregation rules and worked examples are provided in Appendix.

5 Benchmarking Results

Table 2 summarizes success, degradation, and reference retention, while Table 3 provides the N_0-TWAM task–operator breakdown. All comparisons are descriptive, given unequal task coverage, external references, and the absence of trial-level outcomes or verified matched controls. Supplementary results appear in Appendix.

Aggregate performance. Table 2 separately reports model-level means: success falls from 48% to 39% for ACT, 67% to 53% for FTP-1, 85% to 76% for N_0-TWAM, and 83% to 71% for N_0-VTLA. The corresponding gaps are 9, 14, 9, and 12 percentage points. A small gap need not imply high task success: N_0-VTLA scores 21% on Insert HDMI despite a 4-pp gap, compared with 97% on Insert Tube and a 2-pp gap. Unequal task coverage and external N_0 Clean scores preclude a controlled cross-model ranking.

Reference tactile gain retention. ACT and FTP-1 have mean $\mathrm{TGR}^{\mathrm{ref}}$ values of 0.56 and 0.00 (rounded), respectively, over their different sets of six eligible tasks. FTP-1's near-zero mean

Table 3: **Per-operator success rates (%) of N_0-TWAM on RoboTactile.** Scores are integers. Avg. covers eight tasks; the shaded 14-fault mean excludes Clean and matches Table 2. Clean results are external references, not matched RoboTactile re-evaluations.

Fault operator	Task-level results								Avg.
	Insert HDMI	Insert Hole	Insert Tube	Grasp Classify	Lift Can	Lift Bottle	Pull Out Key	Put Bottle in Shelf	
Clean	58	93	79	87	98	68	99	94	**85**
Observation Availability									
Stream Absence	46	93	77	86	97	52	87	66	**76**
Frame Erasure	44	88	77	83	73	58	94	64	**73**
Signal Fidelity									
Global Response Drift	49	90	75	84	64	9	94	65	**66**
Spatial Sensitivity Attenuation	52	70	79	84	92	67	97	79	**78**
Persistent Surface Artifact	46	90	77	82	61	8	95	89	**69**
Local Nonresponsive Patch	53	83	76	83	63	8	85	65	**65**
Contact Shape Distortion	52	88	77	86	95	64	98	89	**81**
Residual Imprint	54	82	76	83	92	65	96	92	**80**
High-Load Saturation	54	63	78	87	98	65	99	92	**80**
Temporal Integrity									
Fixed Latency	56	70	76	85	93	66	94	87	**78**
Sample and Hold	56	89	76	84	93	63	96	91	**81**
Cross-Sensor Skew	53	81	78	85	95	64	95	94	**81**
Contextual Consistency									
Sensor Identity Swap	52	91	77	84	96	67	97	94	**82**
Frame Misregistration	55	81	75	86	94	68	98	91	**81**
14-fault mean	52	83	77	84	86	52	95	83	**76**

combines negative retention on Insert Hole (-0.32), Insert Tube (-1.78), and Lift Bottle (-0.10) with positive retention elsewhere. Thus, aggregate retention can conceal below-reference performance on individual tasks. These ratios describe performance relative to the supplied visual references, not causal tactile benefit or harm.

Operator- and task-specific sensitivity. For N_0-TWAM, Lift Bottle success is 8% under Local Nonresponsive Patch versus 52% under Stream Absence, illustrating that available but corrupted touch can coincide with poorer performance than absent touch. Stream Absence remains a test-time intervention, not a separately trained vision-only control. Temporal operators also differ: Fixed Latency and Sample and Hold yield 70% and 89% on Insert Hole, respectively. Task dependence is evident under High-Load Saturation, whose Clean-reference gaps are 30 pp on Insert Hole but only 1 pp on Insert Tube. These configuration-specific contrasts do not imply equal perturbation magnitudes and motivate task–operator reporting alongside aggregate scores. Phase-conditioned sensitivity and post-restoration recovery require additional experiments.

6 Conclusion

We introduce RoboTactile, a benchmark of fourteen tactile observation faults across availability, fidelity, timing, and contextual consistency. Results from four models show that aggregate success conceals substantial task–operator variation: degraded but available touch can accompany lower success than interrupted touch, and different temporal faults produce distinct outcomes. The task–operator breakdown further reveals that small performance gaps need not imply high absolute success under faults. We combine absolute performance gaps with tactile gain retention, distinguishing descriptive visual-reference comparisons from verified matched-control estimates. By identifying where success is preserved or falls below the supplied references, RoboTactile provides concrete targets for fault detection and adaptive multimodal control.

Limitations and Future Work. The current evidence is limited to aggregate simulation scores, unequal task coverage, external Clean references, and unverified visual-control matching. Future work will extend evaluation to physical robots and diverse sensors, using matched controls and repeated trials to study contact-phase sensitivity, autonomous recovery, and reliable long-term manipulation.

Reproducibility Statement

The benchmark is designed around immutable task, model, and fault manifests. The final release will record dataset versions, simulator and container versions, upstream model commits, checkpoint hashes, preprocessing and action mappings, random seeds, per-episode traces, and evaluation commands. The adapter contract and reference statistical protocol are specified in Sections 3–4 and Appendix. The aggregate records do not document a frozen experiment manifest or an exact compute inventory, which remain necessary for replication.

Ethics Statement

RoboTactile is a simulator-based reliability benchmark and does not involve human subjects or claim hardware safety certification. Controlled sensor faults can help expose unsafe reliance on tactile input, but performance in Isaac Sim does not establish robustness on a physical robot. We therefore report the simulator-only scope prominently, avoid real-world safety claims, document licenses for datasets and evaluated models, and plan to release failure traces that support auditing rather than only aggregate rankings.

Statement on the Use of AI

Generative AI tools were used to help organize the benchmark methodology, draft and edit portions of this manuscript, and format references. They were not used to generate empirical observations or fill result fields in this draft. All AI-assisted text, citations, code, and experimental artifacts will be reviewed and verified by the authors, who take responsibility for the final content. This statement will be updated to reflect the complete research workflow before submission.

References

Junhao Cai, Zetao Cai, Jiafei Cao, Yilun Chen, Zeyu He, Lei Jiang, Hang Li, Hengjie Li, Yang Li, Yufei Liu, Yanan Lu, Qi Lv, Haoxiang Ma, Jiangmiao Pang, Yu Qiao, Zherui Qiu, Yanqing Shen, Xu Shi, Yang Tian, Bolun Wang, Hanqing Wang, Jiaheng Wang, Tai Wang, Xueyuan Wei, Chao Wu, Yiman Xie, Boyang Xing, Yuqiang Yang, Yuyin Yang, Qiaojun Yu, Feng Yuan, Jia Zeng, Jingjing Zhang, Shenghan Zhang, Shi Zhang, Zhuoma Zhaxi, Bowen Zhou, Yuanzhen Zhou, Yunsong Zhou, Hongrui Zhu, Yangkun Zhu, and Yuchen Zhu. InternVLA-A1: Unifying understanding, generation and action for robotic manipulation, 2026a. URL https://arxiv.org/abs/2601.02456.

Rui Cai, Jun Guo, Xinze He, Piaopiao Jin, Jie Li, Bingxuan Lin, Futeng Liu, Wei Liu, Fei Ma, Kun Ma, Feng Qiu, Heng Qu, Yifei Su, Qiao Sun, Dong Wang, Donghao Wang, Yunhong Wang, Rujie Wu, Diyun Xiang, Yu Yang, Hangjun Ye, Yuan Zhang, and Quanyun Zhou. Xiaomi-Robotics-0: An open-sourced vision-language-action model with real-time execution, 2026b. URL https://arxiv.org/abs/2602.12684.

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

Trinity Chung, Kashu Yamazaki, Dhruv Patel, Alexis Duburcq, Yiling Qiao, Katerina Fragkiadaki, and Aran Nayebi. Tactile genesis: Exploring tactile sensors at scale for learning dexterous tasks, 2026. URL https://arxiv.org/abs/2606.22332.

Benjamin Davis and Hannah Stuart. Benchmarking resilience and sensitivity of polyurethane-based vision-based tactile sensors, 2025. URL https://arxiv.org/abs/2511.07797.

Yinpeng Dong, Caixin Kang, Jinlai Zhang, Zijian Zhu, Yikai Wang, Xiao Yang, Hang Su, Xingxing Wei, and Jun Zhu. Benchmarking robustness of 3d object detection to common corruptions in autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1022–1032, 2023. doi: 10.1109/CVPR52729.2023.00105. URL https://arxiv.org/abs/2303.11040.

Elliott Donlon, Siyuan Dong, Melody Liu, Jianhua Li, Edward Adelson, and Alberto Rodriguez. Gelslim: A high-resolution, compact, robust, and calibrated tactile-sensing finger, 2018. URL https://arxiv.org/abs/1803.00628.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations, 2019. URL https://arxiv.org/abs/1903.12261.

Yunao Huang, Shiyu Sang, Haotao Lu, Suting Ni, Shijie Wu, Ziyang Guo, Ye Shi, and Jingya Wang. Vitacworld: Scaling visuo-tactile world models for contact-rich robot manipulation, 2026a. URL https://arxiv.org/abs/2607.22530.

Yuzhe Huang, Jiaping Wu, Jiaming Jiang, Hezhe Lin, Aikebaier Aierken, Yunlong Wang, Kun Cheng, Ziyuan Jiao, and Yuanxin Zhong. Ht-bench: Benchmarking and learning dexterous full-hand tactile representations with egocentric vision, 2026b. URL https://arxiv.org/abs/2606.19161.

Bowen Jing, Mingxin Wang, Ruiyang Hao, Chenchen Ge, Hanwen Shen, Junjie He, Yang Cui, Yiming Hou, Weitao Zhou, Jiawei Wang, Minglei Li, Dandan Zhang, Ding Zhao, Houde Liu, Xiaofan Li, Si Liu, Ping Luo, and Haibao Yu. SoftVTBench: A safety-aware visuo-tactile benchmark for physically constrained robotic manipulation of deformable objects, 2026. URL https://arxiv.org/abs/2607.04234.

Michelle A. Lee, Matthew Tan, Yuke Zhu, and Jeannette Bohg. Detect, reject, correct: Crossmodal compensation of corrupted sensors, 2020. URL https://arxiv.org/abs/2012.00201.

Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Fei Han, Mingrui Yu, Zelin Gao, Nan Xue, Xing Zhu, Yujun Shen, and Yinghao Xu. Causal world modeling for robot control. *arXiv preprint arXiv:2601.21998*, 2026. doi: 10.48550/arXiv.2601.21998. URL https://arxiv.org/abs/2601.21998.

Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie Chen, Peter Wu, Michelle A. Lee, Yuke Zhu, Ruslan Salakhutdinov, and Louis-Philippe Morency. Multibench: Multiscale benchmarks for multimodal representation learning, 2021. URL https://arxiv.org/abs/2107.07502.

Qingtao Liu, Yu Cui, Zhengnan Sun, Gaofeng Li, Jiming Chen, and Qi Ye. VTDexManip: A dataset and benchmark for visual-tactile pretraining and dexterous manipulation with reinforcement learning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=jf7C7EGw21.

Yunfan Lou, Yifan Ye, Yankai Fu, Jun Cen, Xiaowei Chi, Yaoxu Lyu, Peidong Jia, Sirui Han, Zhihe Lu, and Shanghang Zhang. Dream-tac: A unified tactile world action model for contact-rich robot manipulation, 2026. URL https://arxiv.org/abs/2606.08737.

Quan Khanh Luu, Pokuang Zhou, Zhengtong Xu, Zhiyuan Zhang, Qiang Qiu, and Yu She. Manifeel: Benchmarking and understanding visuotactile manipulation policy learning, 2025. URL https://arxiv.org/abs/2505.18472.

NeoteAI Team and Fudan TEAI Team. N_0-TWAM: Scaling Tactile-Native World-Action Model for Contact-Rich Manipulation, 2026a. URL https://arxiv.org/abs/2607.23783.

NeoteAI Team and Fudan TEAI Team. N_0-VTLA: Scaling Vision–Tactile–Language–Action Model with Latent Tactile Tokens, 2026b. URL https://arxiv.org/abs/2607.23782.

Radhen Patel, Rui Ouyang, Branden Romero, and Edward Adelson. Digger finger: Gelsight tactile sensor for object identification inside granular media, 2021. URL https://arxiv.org/abs/2102.10230.

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, 2025. URL https://arxiv.org/abs/2504.16054.

Jieji Ren, Wenxin Du, Yueshi Dong, Ningbin Zhang, Heng Guo, Boxin Shi, Jiang Zou, and Guoying Gu. GelLight: Illumination design, modeling, and optimization for camera-based tactile sensor. *Advanced Intelligent Systems*, 2025. doi: 10.1002/aisy.202400596. URL https://doi.org/10.1002/aisy.202400596.

José Sánchez-Durán, José Hidalgo-López, Julián Castellanos-Ramos, Óscar Oballe-Peinado, and Fernando Vidal-Verdú. Influence of errors in tactile sensors on some high level parameters used for manipulation with robotic hands. *Sensors*, 15(8):20409–20435, 2015. doi: 10.3390/s150820409. URL https://doi.org/10.3390/s150820409.

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.2022.3144949. URL https://arxiv.org/abs/2109.04027.

Meng Wang, Wanlin Li, Hao Liang, Boren Li, Kaspar Althoefer, Yao Su, and Hangxin Liu. Large-scale deployment of vision-based tactile sensors on multi-fingered grippers. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 13946–13952, 2024. doi: 10.1109/IROS58592.2024.10801566. URL https://arxiv.org/abs/2408.02206.

Shaoyuan Xie, Lingdong Kong, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. Robobev: Towards robust bird's eye view perception under corruptions, 2023. URL https://arxiv.org/abs/2304.06719.

Angen Ye, Boyuan Wang, Chaojun Ni, Guan Huang, Guosheng Zhao, Hao Li, Hengtao Li, Jie Li, Jindi Lv, Jingyu Liu, Min Cao, Peng Li, Qiuping Deng, Wenjun Mei, Xiaofeng Wang, Xinze Chen, Xinyu Zhou, Yang Wang, Yifan Chang, Yifan Li, Yukun Zhou, Yun Ye, Zhichao Liu, and Zheng Zhu. GigaWorld-Policy: An efficient action-centered world–action model, 2026a. URL https://arxiv.org/abs/2603.17240.

Guo Ye, Zexi Zhang, Xu Zhao, Shang Wu, Haoran Lu, Shihan Lu, and Han Liu. Learning to feel the future: Dreamtacvla for contact-rich manipulation, 2026b. URL https://arxiv.org/abs/2512.23864.

Jinhui Ye, Ning Gao, Senqiao Yang, Jinliang Zheng, Zixuan Wang, Yuxin Chen, Pengguang Chen, Yilun Chen, Shu Liu, and Jiaya Jia. StarVLA-α: Reducing complexity in vision-language-action systems, 2026c. URL https://arxiv.org/abs/2604.11757.

Chengbo Yuan, Zicheng Zhang, Mingjie Zhou, Wendi Chen, Yi Wang, Zhuoyang Liu, Dantong Niu, Shuo Wang, Hui Zhang, Wenkang Zhang, Yingdong Hu, Yuanqing Gong, Wanli Xing, Chuan Wen, Cewu Lu, Kaifeng Zhang, and Yang Gao. FTP-1: A generalist foundation tactile policy across tactile sensors for contact-rich manipulation, 2026a. URL https://arxiv.org/abs/2606.13102.

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*, 2026b. doi: 10.48550/arXiv.2603.16666. URL https://arxiv.org/abs/2603.16666.

Zhiyuan Zhang, Pokuang Zhou, Kaidi Zhang, Adeesh Desai, Temitope Amosa, Davood Soleymanzadeh, Jiuzhou Lei, Minghui Zheng, and Yu She. Contactworld: What matters in vision-tactile world models for contact-rich manipulation, 2026. URL https://arxiv.org/abs/2606.13877.

Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *Robotics: Science and Systems*, 2023. URL https://arxiv.org/abs/2304.13705.

Wugeng Zheng, Ziwen Kan, Tianlong Chen, Chen Chen, and Song Wang. Mutebench: Modality unavailability tolerance evaluation for incomplete multimodal fusion, 2026. URL https://arxiv.org/abs/2605.15235.

Anya Zorin, Zilin Si, Myungsun Park, Junsung Park, Alexiy Buynitsky, Sachin Bhadang, Taejun Park, Sohee John Yoon, Yong-Lae Park, Oliver Kroemer, Zeynep Temel, Michael T. Tolley, Sha Yi, and Xiaolong Wang. TacO: Benchmarking tactile sensors for object manipulation, 2026. URL https://arxiv.org/abs/2605.21976.