Touch-R1: Reinforcing Touch Reasoning in MLLMs

Yingxin Lai[1,5*] **Yafei Zhou**[2,5*] **Fucai Zhu**[5] **Siyu Zhu**[3] **Weihao Yuan**[4,5†]
[1]Xiamen University [2]Great Bay University [3]Fudan University [4]Nanjing University
[5]Daimon Robotics

Abstract

While rule-based reinforcement learning has recently catalyzed explicit reasoning in multimodal models, tactile reasoning remains largely underexplored. Existing tactile-language models primarily rely on supervised or contrastive objectives, which limits their capacity to ground predictions in physical evidence or rectify misleading visual priors. Tactile reasoning introduces two modality-specific challenges: the ordinal nature of physical attributes (e.g., hardness, roughness) and the cross-sensor distribution shifts inherent in optical tactile hardware. In this work, we introduce TouchReason-1M, a large-scale multimodal dataset comprising over 1M synchronized tactile pairs across four distinct sensors, and TouchReason-Bench, a rigorous framework for evaluating tactile perception and visual-tactile conflict resolution. Building upon these, we propose Touch-R1, a tactile reasoning MLLM based on Qwen2.5-VL-7B. Touch-R1 is trained via a tactile-grounded GRPO objective that combines ordinal-aware accuracy, cross-sensor physical consistency, structured-format control, and an input-side tactile grounding objective. Specifically, the input-side tactile grounding objective encourages the policy to respond to perturbations of tactile inputs. On TouchReason-Bench, Touch-R1-7B outperforms Octopi-13B and GPT-4o by 18.4 and 24.7 points, respectively, in the aggregate score. Its structured reasoning traces reveal emergent behaviors of probing, comparison, and revision, demonstrating that R1-style reasoning can be effectively grounded in physical contact. Our code and data will be made public on the project page: `https://laiyingxin2.github.io/Touch-R1/`.

1 Introduction

Humans rarely infer physical properties from vision alone. A visual impression is often followed by tactile contact, and the resulting physical evidence may revise the initial judgment when appearance is misleading. This perception-verification-revision loop is central to embodied interaction with the physical world. Yet current multimodal LLMs often fail on precisely this loop: in visual-tactile conflict settings, models such as GPT-4o and Gemini-2.5-Pro tend to follow the visual object prior rather than revise their predictions from tactile evidence. Our experiments confirm this gap on TouchReason-Bench, where general-purpose MLLMs substantially underperform Touch-R1, and qualitative examples show that tactile-grounded reasoning can correct visually plausible but physically inconsistent predictions. This suggests a paradigm gap: current models can associate touch with language, but do not reliably use touch as evidence for physical-property revision. In this work, we study tactile-grounded reasoning. For the tactile modality, we focus on optical tactile sensors because they provide high-resolution contact observations for tactile reasoning.

Recent tactile-language models, including Octopi [1], TVL [2], Touch100k [3], AnyTouch [4], SToLa [5], and VitaTouch [6], have advanced tactile-language modeling through supervised fine-tuning, contrastive alignment, and mixture-of-experts architectures. However, these methods do not

*Equal contribution. Author order is random.
†Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

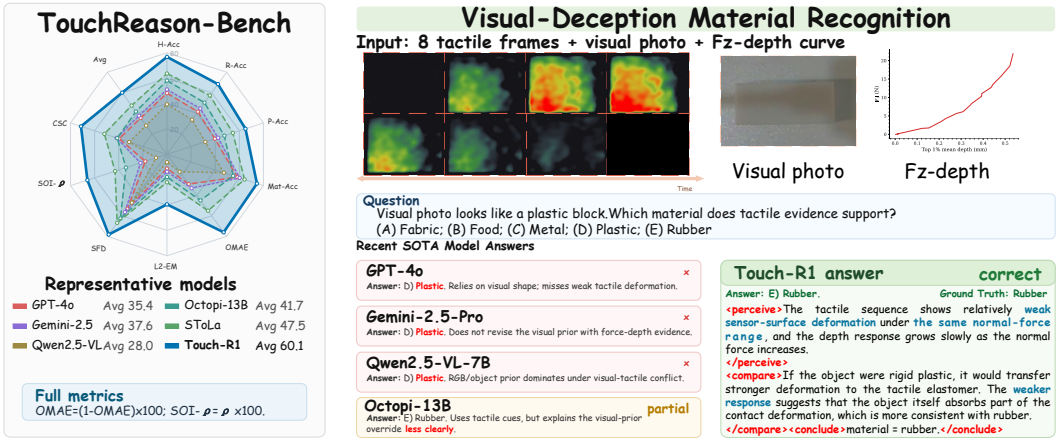


Figure 1: Motivating example and benchmark overview. Left: Touch-R1 improves over representative closed-source MLLMs, open-source VLMs, and tactile-specialist models on TouchReason-Bench. Right: in a visual-deception material-recognition example, the visual photo suggests plastic, while tactile frames and the force-depth curve indicate compliant deformation. Touch-R1 revises the visual prior and predicts rubber using structured tactile-grounded reasoning.

verify whether the final answer is grounded in tactile evidence. As a result, they may learn visually plausible or object-category-driven answers without learning when tactile evidence should override vision. In this work, we combine rule-based RL objectives with an input-side grounding objective to encourage tactile dependence.

The R1 paradigm [7, 8] has recently been extended to vision [9, 10, 11] and video [12, 13], but these extensions largely rely on strong pretrained priors in their input modalities. Optical tactile imagery violates this assumption in three ways. First, marker displacement, elastomer deformation, and contact-induced texture patterns are largely absent from natural-image pretraining corpora. Second, GelSight Mini [14], Xense [15], Tac3D [16], DM-Tac X [17], and other optical tactile sensors differ in resolution, marker layout, illumination, and optical geometry, making pixel-level alignment across sensors unreliable. Third, physical attributes are ordinal rather than categorical: predicting moderate hardness for a soft object is physically closer than predicting hard, but sparse binary rewards treat both mistakes equally. These properties make tactile reasoning a poor fit for standard supervised alignment and vanilla rule-based RL.

To address these problems, we propose Touch-R1, a tactile reasoning MLLM trained through three stages: tactile dynamics pretraining, QA supervised fine-tuning, and rule-based reinforcement learning. First, a ViT-based tactile encoder is trained to predict the latent code of the next tactile frame, encouraging sensitivity to deformation dynamics rather than static appearance. Second, we construct TouchReason-1M, a four-optical-sensor tactile reasoning dataset with over 1M synchronized tactile data pairs, and use verified reasoning-style QA annotations to align tactile observations with structured physical-property answers. Third, we train Touch-R1 with a tactile-grounded GRPO objective that evaluates final conclusions against ground-truth labels and encourages policy sensitivity to noise-perturbed tactile inputs. This design encourages the use of tactile evidence alongside visual and object-category information.

Touch-R1 combines tactile-specific output rewards with an input-side tactile grounding objective. The Ordinal-Aware Accuracy Reward replaces binary correctness with dense credit over ordered attributes such as hardness, roughness, and protrusion, so adjacent-level errors are penalized less than larger ordinal deviations. The Cross-Sensor Physical Consistency Reward encourages paired sensors observing the same object to agree in physical-property space without requiring pixel-level sensor alignment. The Structured-Format Reward stabilizes the `<perceive>`, `<compare>`, and `<conclude>` interface required for reward parsing. In addition, the input-side tactile grounding objective perturbs tactile streams and encourages the policy distribution to depend on tactile evidence rather than visual or object-category shortcuts. In experiments, these rewards reduce ordinal error, improve cross-sensor consistency, and help the model revise misleading visual priors using tactile evidence. Our contributions can then be summarized as follows:

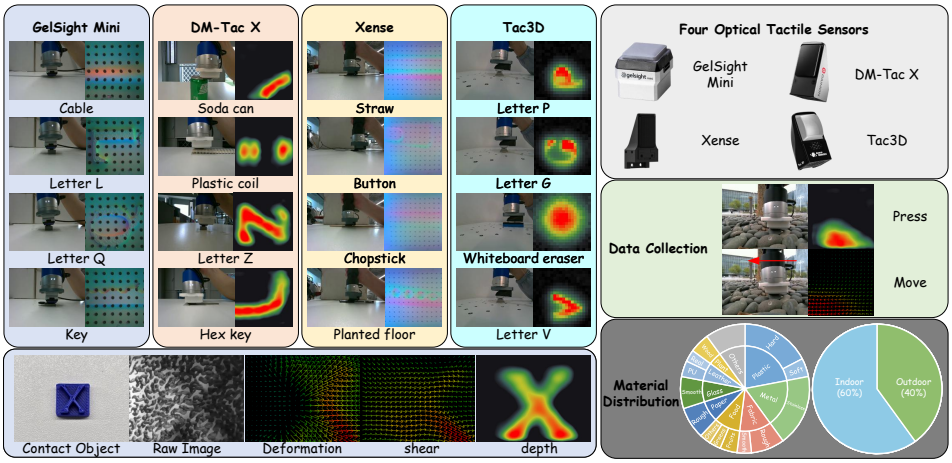


Figure 2: Overview of TouchReason-1M collected with four optical tactile sensors on 1000+ objects across diverse materials and scenes.

- We propose **Touch-R1**, a tactile reasoning MLLM trained with tactile-grounded GRPO, combining ordinal-aware accuracy, cross-sensor physical consistency, structured-format control, and input-side tactile grounding to encourage physically grounded tactile reasoning.

- We construct **TouchReason-1M**, a multimodal tactile reasoning dataset containing over 1M synchronized tactile data pairs collected with four optical tactile sensors, and introduce **TouchReason-Bench** for evaluating tactile perception, ordinal comparison, cross-sensor consistency, and visual-tactile conflict.

- Extensive experiments show that Touch-R1 improves tactile reasoning accuracy, reduces ordinal error, strengthens cross-sensor consistency, and produces structured perception-comparison-revision rationales grounded in physical contact.

2 Related Work

Tactile perception and tactile-language models. Touch is the key modality for understanding and interacting with the physical world, especially for robot manipulation [18, 19, 20, 21]. Optical tactile sensors [14, 22, 16, 17] share an elastomer-plus-camera design, but differ in resolution, marker layout, illumination, and optical geometry. This heterogeneity has motivated cross-sensor tactile representation learning [4, 23, 24, 25]. Recent tactile-language models connect touch to language through supervised fine-tuning, contrastive alignment, or expert routing, including TVL [2], Touch100k [3], Octopi [1], and SToLa [5]. However, these models are mainly trained with imitation or alignment objectives and do not explicitly verify whether final answers are grounded in tactile evidence, especially when visual and tactile cues conflict. In addition, the datasets [26, 2, 3, 27] used by previous methods are usually limited in scale or collection standardization.

Reinforcement learning for multimodal reasoning. DeepSeek-R1 [7] shows that GRPO [8] with rule-based rewards can elicit reasoning, inspiring extensions to visual and video reasoning [9, 10, 11, 12] and more stable variants such as DAPO [28], Dr. GRPO [29], and GSPO [30]. Closely related consistency-based methods include GRPO-CARE [13], Visionary-R1 [31], and Video-R1's T-GRPO [12], which verify agreement across reasoning, perception, answers, or temporal order. Touch-R1 extends this principle to tactile reasoning by perturbing tactile inputs and rewarding policies that remain sensitive to tactile evidence while enforcing ordinal-aware accuracy, cross-sensor physical consistency, and structured reasoning.

3 Method

3.1 TouchReason-1M and TouchReason-Bench

Overview. We introduce TouchReason-1M, a multimodal tactile reasoning dataset containing over 1M synchronized tactile data pairs for R1-style training, together with TouchReason-Bench, a four-tier benchmark for evaluating tactile MLLMs. TouchReason-1M contains over 1,000 household and everyday objects across nine material categories: plastic, metal, fabric, food, paper, glass, leather, wood, and other materials. Each object is captured using four heterogeneous optical tactile sensors: GelSight Mini, Xense, Tac3D, and DM-Tac X. In this work, each tactile data pair refers to a synchronized tactile-force record, including the raw tactile observation, deformation field, shear field, depth field, force measurement, and associated metadata. Unlike prior tactile-language datasets that primarily target single-sensor alignment, tactile captioning, or property classification, TouchReason-1M is designed for tactile reasoning. It integrates cross-sensor tactile observations, force-conditioned contact dynamics, ordinal physical-property labels, and verifiable reasoning-style QA. This design supports both supervised tactile-language alignment and rule-based reinforcement learning, where final answers, structured outputs, and tactile-input dependence can be explicitly evaluated or regularized.

Acquisition. Optical tactile sensors perceive contact by using an internal camera to capture the visual changes of a soft elastomer as it deforms during contact. Depending on the sensor structure and imaging principle, these changes may appear as sparse marker displacements, dense pattern deformation, or variations in brightness, color, and reflectance caused by surface deformation. Through dedicated processing, raw visual observations are converted into surface marker or pattern displacement fields, which are then used to estimate shear fields, depth fields, and contact force fields.

Taking DM-Tac X as an example, Fig. 2 shows representative tactile signals generated during pressing and sliding interactions, including the deformed raw image, deformation field, shear field, and depth field. The deformation field is obtained from the raw image and represents the two-dimensional displacement field of the surface marker pattern under contact. The shear and depth fields are further estimated from the deformation field, with the shear field capturing tangential contact-related responses and the depth field describing local geometry associated with normal indentation. Together, these fields provide complementary cues for subsequent tactile perception and modeling.

As shown in Fig. 2, TouchReason-1M is collected through a force-guided two-stage acquisition protocol inspired by prior tactile collection pipelines. Each object is first recorded with calibrated RGB-D cameras to obtain visual appearance, depth geometry, and a contact-region reference. The object is then explored by four optical tactile sensors under the same standardized interaction protocol. In the first stage, each sensor performs continuous normal pressing from initial contact to 20 N, producing force-indexed deformation trajectories that capture hardness, compliance, and pressure-dependent responses. In the second stage, each sensor performs six exploratory motions at target normal-force levels to capture shear responses, local texture variations, and protrusion-induced deformation. All tactile streams are synchronized with compensated 3D contact force signals and stored as lossless frame stacks rather than compressed videos, preserving marker displacement, fine-grained texture, and deformation patterns critical for tactile reasoning. After filtering failed contacts, unstable force traces, saturated responses, and invalid contact regions, TouchReason-1M retains 14,700 valid tactile sequences and 1,323,000 synchronized tactile data pairs.

Human Labeling. Two trained annotators assigned hardness, roughness, protrusion, and material labels for each object using synchronized RGB-D images, tactile sequences, and compensated 3D contact force signals. The first three attributes were encoded as three-level ordinal attributes and cross-verified against force–deformation curves, shear responses, and contact geometry. Inter-annotator Cohen's κ was 0.82, 0.78, and 0.80 for the three ordinal attributes, with 94.3% exact-match agreement for material labels; disagreements were adjudicated by a senior annotator.

QA Generation and Quality Verification. As shown in Fig. 3, based on the human-labeled tactile data pairs, Gemini 2.5 Pro was used to generate candidate reasoning-style QA pairs and rationales from RGB-D views, tactile frame stacks, force traces, sensor metadata, and contact ROIs. The prompt explicitly required each rationale to ground its conclusion in concrete tactile evidence, such as marker displacement, deformation magnitude, surface texture, or force-dependent response, rather than vague assertions. We randomly sampled 2,000 generated pairs for independent human

Table 1: Comparison of TouchReason-1M with representative tactile and tactile-language datasets. TouchReason-1M is the first to combine multi-sensor optical tactile acquisition, force-guided interaction, ordinal physical labels, and verifiable reasoning-style QA.

Dataset	Objects	Touches / Frames	Source	Sensors	Force guided	Ordinal label	Reason. QA
More Than a Feeling [32]	65	6.5K trials	Robot	GelSight	✗	✗	✗
The Feeling of Success [33]	106	9.3K frames	Robot	GelSight	✗	✗	✗
VisGel [34]	195	12K frames	Robot	GelSight	✗	✗	✗
Touch and Go [26]	3,971	13.9K frames	Human	GelSight	✗	✗	✗
TVL [2]	in-the-wild	44K pairs	Human + VLM	DIGIT	✗	✗	✗
Touch100k [3]	–	∼100K pairs	VLM + filter	GelSight	✗	✗	✗
Octopi/PhysiCLeAR [1]	74	408 videos	Human	GelSight Mini	partial	✓	✗
AnyTouch/TacQuad [4]	–	72K frames	VLM	4 sensors	✗	✗	✗
TouchReason-1M (Ours)	**1000+**	**14K seq. / 1.32M fr.**	**Human + Gemini 2.5 Pro**	**4 sensors**	✓	✓	✓

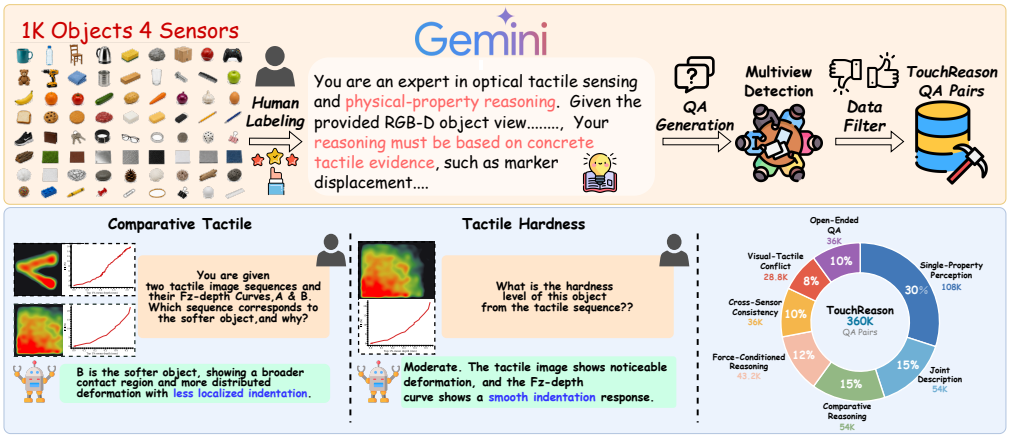


Figure 3: Overview of TouchReason QA pair construction. The upper panel shows the data collection, human labeling, QA generation, multiview detection, and filtering pipeline. The lower panel displays typical QA examples for comparative tactile reasoning and hardness recognition, together with the distribution of QA task types.

evaluation, yielding a 91.6% pass rate across three axes: tactile-evidence grounding, consistency with human ordinal labels, and self-consistency of the reasoning chain. Rule-based filters further removed label description conflicts, hallucinated contact regions, missing answers, and reasoning–answer inconsistencies.

3.2 Touch-R1

This section introduces Touch-R1, a reinforcement-learning framework for tactile-grounded multi-sensor reasoning, as shown in Fig. 4. Our key observation is that answer correctness alone is insufficient for tactile reasoning: a policy may predict the correct label while relying on category priors, ignoring tactile streams, or producing cross-sensor contradictions. Such answer-only success masks poor grounding, where the output is correct but the reasoning may not be physically reliable.

Touch-R1 addresses this issue by augmenting GRPO with three output-side rewards and one input-side grounding objective. The output-side rewards capture ordinal tactile attributes, enforce cross-sensor physical consistency, and stabilize structured reasoning, while the input-side objective perturbs tactile streams to encourage policy sensitivity to tactile evidence. As a result, Touch-R1 optimizes for answers that are not only correct, but also grounded in physically consistent and input-dependent tactile evidence.

Ordinal-aware reward. Tactile attributes such as hardness, roughness, and protrusion are not purely categorical; when elasticity is annotated, it is also treated as an ordered physical attribute. A binary correctness reward ignores this structure and treats adjacent mistakes the same as distant mistakes. For example, if the ground-truth hardness level is high, predicting a neighboring level is

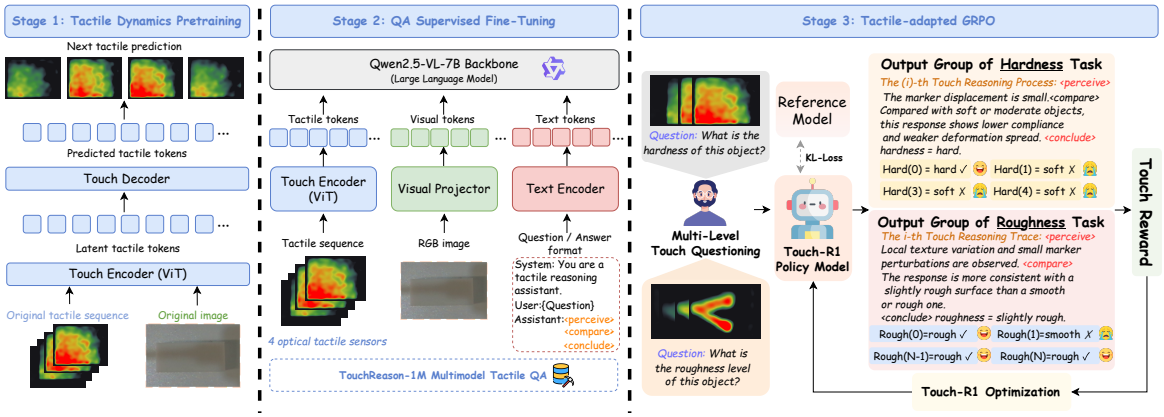


Figure 4: **Touch-R1: a three-stage framework for tactile-grounded reasoning. (1) Tactile Dynamics Pretraining**: a ViT touch encoder is pretrained by predicting future tactile tokens from past ones, capturing deformation dynamics across optical tactile sensors. **(2) QA Supervised Fine-Tuning**: tactile, visual, and text tokens are aligned to a Qwen2.5-VL-7B backbone on TouchReason-1M, supervising the assistant to answer under a structured `<perceive>`–`<compare>`–`<conclude>` template. **(3) Tactile-adapted GRPO**: starting from the SFT checkpoint, rollouts are scored by an ordinal-aware accuracy reward, a cross-sensor physical consistency reward, and a structured-format reward, optimizing the policy toward answers that are correct, cross-sensor consistent, and grounded in tactile evidence.

physically closer to the truth than predicting the opposite end of the scale. Under sparse answer-only feedback, this mismatch can encourage the policy to collapse toward frequent middle levels, leading to ordinal collapse.

To counteract this failure, we replace binary correctness with an ordinal-aware reward. Let $y \in \{0, \ldots, K-1\}$ denote the ground-truth ordinal label, let $\hat{y}$ denote the prediction parsed from the `<conclude>` segment, and let $d = |\hat{y} - y|$ be the ordinal distance. We define

$$R_{\text{acc}}(\hat{y}, y) = \begin{cases} 1.0, & \hat{y} \text{ is parsable and } d = 0, \\ 0.4, & \hat{y} \text{ is parsable and } d = 1, \\ 0.0, & \hat{y} \text{ is parsable and } d \geq 2, \\ -0.1, & \hat{y} \text{ is unparsable.} \end{cases} \tag{1}$$

This reward encodes the prior that predictions closer on the ordinal axis are closer in physical meaning. The adjacent-error reward remains strictly below the exact-match reward, so it does not weaken the incentive for correct prediction; however, it is still preferred over distant errors, which densifies the reward signal along the answer space. For categorical labels such as material, we use the exact-match variant of this reward, assigning 1.0 to correct parsable predictions and 0 otherwise.

Cross-sensor physical consistency reward. Answer-only rewards also fail to penalize cross-sensor contradictions. For the same object, heterogeneous tactile sensors such as GelSight Mini, Xense, Tac3D, and DM-Tac X may produce different low-level signals, but their summarized physical attributes should agree after being mapped into the same ordinal space. If the policy reports conflicting per-sensor evidence but happens to predict the correct final label, an answer-only reward will still reinforce the rollout.

We therefore introduce a cross-sensor physical consistency reward. Let p_a and p_b denote the ordinal property summaries parsed from the `<perceive>` segment for two sensors a and b, and let p_c denote the cross-sensor conclusion parsed from the `<compare>` segment. The normalized ordinal gap is $w(u, v) = \min(|u - v|/(K-1), 1)$. We define

$$R_{\text{cs}} = \begin{cases} 1 - \frac{1}{2}\big(w(p_a, p_c) + w(p_b, p_c)\big), & p_a, p_b, p_c \text{ are parsable,} \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

This reward reaches its maximum when both per-sensor summaries agree with the cross-sensor conclusion. Small ordinal disagreements incur mild penalties, whereas distant contradictions are penalized more strongly. The reward does not require different sensors to produce identical raw

tactile signals; instead, it requires their summarized physical-property evidence to be consistent. This encourages the policy to ground its conclusion in shared tactile evidence rather than sensor-specific artifacts or language shortcuts.

Structured-format reward. Touch-R1 uses a three-segment reasoning template with `<perceive>`, `<compare>`, and `<conclude>` segments. This template is not merely cosmetic: it defines the interface through which the reward functions parse per-sensor evidence, cross-sensor comparison, and the final ordinal prediction. Once RL training causes the output format to drift, R_{acc} and R_{cs} become unreliable because the required fields may be missing, duplicated, disordered, or unparsable.

To stabilize this interface, we use a structured-format reward:

$$R_{\text{fmt}}(o) = \mathbf{1}[o \text{ matches } \langle\text{perceive}\rangle \cdots \langle\text{compare}\rangle \cdots \langle\text{conclude}\rangle]\,. \tag{3}$$

The deterministic parser requires all opening and closing tags to appear exactly once and in order. We keep this reward binary rather than continuous because partially formatted outputs are already penalized by the accuracy and consistency rewards. Assigning extra reward to partial formats can introduce redundant signals and encourage superficially valid but semantically empty templates.

The scalar reward used by GRPO combines the three output-side rewards:

$$R = \lambda_{\text{acc}} R_{\text{acc}} + \lambda_{\text{cs}} R_{\text{cs}} + \lambda_{\text{fmt}} R_{\text{fmt}}, \qquad (\lambda_{\text{acc}}, \lambda_{\text{cs}}, \lambda_{\text{fmt}}) = (0.6, 0.3, 0.1). \tag{4}$$

Here R_{acc} provides the primary supervision for ordinal prediction, R_{cs} regularizes cross-sensor physical consistency, and R_{fmt} preserves a parsable reasoning interface.

Implicit tactile grounding objective. The three rewards above supervise what the policy outputs, but they do not guarantee the policy actually uses tactile input. A policy can exploit object-category priors, visual appearance, or language shortcuts to predict tactile attributes, remaining nearly unchanged even when a tactile stream is corrupted. We refer to this behavior as a tactile-blind shortcut.

To address this issue, we introduce an implicit tactile grounding objective. Let T denote the original tactile observation, and let T_{mask} denote a perturbed observation obtained by replacing one randomly selected tactile stream with shape-matched Gaussian noise. Given question q and prefix $y_{<t}$, we compare the policy distributions under the intact and perturbed tactile inputs:

$$G_{\text{tg}}(\theta) = \mathbb{E}_t[D_{\text{KL}}(\pi_\theta(\cdot \mid q, T, y_{<t}) \,\|\, \pi_\theta(\cdot \mid q, T_{\text{mask}}, y_{<t}))]\,. \tag{5}$$

In practice, we estimate the expectation by averaging over generated token positions. Maximizing G_{tg} encourages output distributions to shift under tactile perturbations, yielding a label-free signal for input dependence without explicit perception targets.

R_{cs} and G_{tg} impose complementary constraints. The consistency reward checks whether the generated per-sensor evidence agrees in physical-property space, while the grounding objective checks whether the policy distribution depends on tactile observations themselves. The former regularizes output-side physical consistency; the latter enforces input-side tactile dependence.

Overall objective. We optimize the policy with standard GRPO using the composite reward R, and add the tactile grounding objective as an input-dependence regularizer. The final Touch-R1 objective is

$$\mathcal{J}_{\text{Touch-R1}}(\theta) = \mathcal{J}_{\text{GRPO}}(\theta; R) + \gamma G_{\text{tg}}(\theta), \tag{6}$$

where γ controls the strength of tactile grounding.

4 Experiments

4.1 Experimental Setup

Training data and protocol. We train Touch-R1 on TouchReason-1M and evaluate it primarily on TouchReason-Bench. Training follows a three-stage protocol. First, we pretrain the tactile backbone on 1M sampled tactile frames from the training split, where 1M denotes frame-level tactile samples rather than interaction sequences, using future tactile-frame prediction to encourage the model to capture force-conditioned contact dynamics. Second, we perform cold-start SFT on a curated 240K subset of generated and verified QA pairs to initialize the structured `<perceive>` `<compare>` `<conclude>` reasoning format. Third, we apply tactile-grounded GRPO on the remaining 120K

Table 2: Main results on TouchReason-Bench. OMAE is lower-better; all other metrics are higher-better. Avg denotes the reported aggregate score.

Model	Property identification				Ordinal reasoning		Comparative reasoning		Cross-sensor	Avg
	H-Acc	R-Acc	P-Acc	Mat-Acc	OMAE↓	L2-EM	SFD	SOI-ρ	CSC	
CLOSED-SOURCE FRONTIER MLLMS										
GPT-4o [35]	48.2	41.5	39.7	58.4	0.71	11.3	53.0	0.18	39.6	35.4
Gemini-2.5-Pro [36]	51.1	44.7	42.3	60.2	0.67	13.8	56.4	0.22	41.8	37.6
Claude-3.5-Sonnet [37]	46.3	39.1	37.5	55.8	0.74	9.8	50.7	0.15	36.9	33.5
OPEN-SOURCE GENERAL VLMS										
LLaVA-OneVision-7B [38]	37.6	32.4	30.1	45.8	0.86	5.7	45.9	0.08	30.2	26.9
Qwen2.5-VL-7B [39]	39.4	33.7	31.8	47.2	0.83	6.2	47.1	0.09	31.4	28.0
InternVL2.5-8B [40]	40.8	35.2	33.0	49.6	0.81	7.5	48.6	0.11	33.1	29.4
TACTILE-SPECIALIST MODELS										
Octopi-7B [1]	53.4	45.9	43.1	51.7	0.62	14.5	58.2	0.27	38.4	38.1
Octopi-13B [1]	58.1	50.3	47.6	54.9	0.55	17.9	62.5	0.34	41.2	41.7
AnyTouch+Qwen [4, 39]	60.2	52.7	50.4	62.1	0.51	19.3	64.1	0.39	49.7	45.0
VTV-LLM-7B [41]	61.5	54.1	51.8	63.0	0.49	20.2	65.4	0.41	50.6	46.3
SToLa [5]	63.7	57.4	54.6	64.3	0.45	22.1	66.8	0.43	51.9	47.5
Touch-R1-7B	**76.8**	**68.4**	**65.1**	**74.6**	**0.24**	**39.5**	**78.2**	**0.66**	**71.3**	**60.1**

Table 3: Cumulative effect of each Touch-R1 component, starting from the Qwen2.5-VL-7B base.

Configuration	OMAE↓	CSC	Avg
Qwen2.5-VL-7B (zero-shot)	0.83	31.4	28.0
+ Cold-start SFT	0.46	52.6	47.8
+ GRPO (0/1)	0.42	54.3	50.1
+ Ordinal Reward	0.31	56.1	54.6
+ Consistency Reward	0.27	69.2	58.3
+ Format Reward	0.26	70.4	59.2
Touch-R1	**0.24**	**71.3**	**60.1**

generated QA samples, where final conclusions are evaluated by rule-based tactile rewards. The three stages use disjoint training subsets to avoid leakage. Unless otherwise specified, the language backbone is Qwen2.5-VL-7B-Instruct. All experiments are conducted on 16 NVIDIA H200 GPUs.

Benchmarks. TouchReason-Bench contains 4,800 QA pairs over 200 held-out objects that are unseen during training. It covers four tactile sensor types, GelSight Mini, Xense, Tac3D, and DM-Tac X, and nine material categories. To evaluate generalization beyond our benchmark, we additionally report tactile video prediction on Touch-and-Go and TouchReason-Video, and compare with the tactile-specialized VTV-LLM-7B on VTV150K. We compare Touch-R1 with three groups of baselines: closed-source frontier MLLMs, including GPT-4o, Gemini-2.5-Pro, and Claude-3.5-Sonnet; open-source general-purpose VLMs, including Qwen2.5-VL-7B, InternVL2.5-8B, and LLaVA-OneVision-7B; and tactile-specialist models, including Octopi, AnyTouch+Qwen, SToLa, and VTV-LLM-7B.

Implementation details. Each tactile sequence is downsampled to 8 frames at 448×448 resolution. GRPO uses group size $G=8$ and KL coefficient $\beta=10^{-3}$. We use AdamW with learning rates 1×10^{-4} for tactile pretraining, 1×10^{-5} for cold-start SFT, and 1×10^{-6} for GRPO. Additional hyperparameters are provided in Appendix B.1.

Metrics. TouchReason-Bench evaluates four aspects of tactile reasoning. Property identification is measured by H-Acc, R-Acc, P-Acc, and Mat-Acc, corresponding to hardness, roughness, protrusion, and material classification. Ordinal and compositional reasoning are measured by OMAE and L2-EM, where OMAE is the mean absolute ordinal error and L2-EM is the exact-match rate for multi-attribute answers. Comparative reasoning is measured by SFD and SOI-ρ, which evaluate pairwise and listwise tactile comparisons. Cross-sensor consistency is measured by CSC. OMAE is lower-better, and all other metrics are higher-better. The task families and metrics are summarized in Appendix B.2.

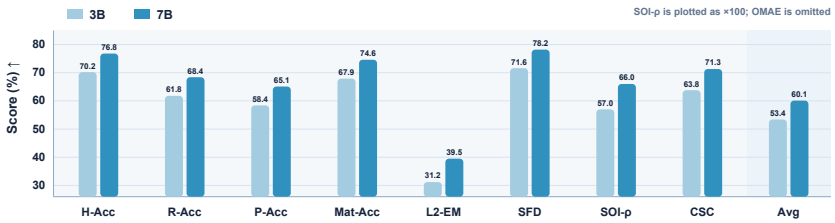


Figure 5: Scaling behavior of Touch-R1 across 3B and 7B backbones on TouchReason-Bench. SOI-ρ is plotted as $\times 100$; OMAE is omitted because it is lower-better.

Table 4: Tactile video prediction results on Touch-and-Go and TouchReason-Video.

Horizon	Method	Modality	Touch-and-Go			TouchReason-Video		
			L1↓	SSIM↑	LPIPS↓	L1↓	SSIM↑	LPIPS↓
Skip 3	SVG [42]	Touch only	0.782	0.572	0.391	0.834	0.529	0.428
	Geng et al. [43]	Touch only	0.628	0.708	0.103	0.681	0.661	0.146
	Ours	Touch only	**0.574**	**0.742**	**0.082**	**0.627**	**0.696**	**0.118**
	SVG [42]	Touch + vision	0.757	0.602	0.368	0.812	0.553	0.409
	Geng et al. [43]	Touch + vision	0.617	0.719	0.091	0.668	0.674	0.132
	Ours	Touch + vision	**0.552**	**0.763**	**0.071**	**0.608**	**0.715**	**0.106**
Skip 5	SVG [42]	Touch only	0.807	0.513	0.412	0.861	0.468	0.454
	Geng et al. [43]	Touch only	0.691	0.698	0.279	0.746	0.648	0.331
	Ours	Touch only	**0.638**	**0.724**	**0.241**	**0.694**	**0.676**	**0.291**
	SVG [42]	Touch + vision	0.762	0.546	0.397	0.819	0.501	0.441
	Geng et al. [43]	Touch + vision	0.663	0.713	0.265	0.721	0.663	0.314
	Ours	Touch + vision	**0.612**	**0.742**	**0.228**	**0.671**	**0.692**	**0.277**

4.2 Main Results on TouchReason-Bench

Table 2 shows that Touch-R1-7B achieves an Avg score of 60.1, outperforming the strongest tactile-specialist baseline, SToLa, by 12.6 points and the strongest closed-source baseline, Gemini-2.5-Pro, by 22.5 points. The gains are largest on metrics that require structured tactile reasoning: Touch-R1 improves CSC by 19.4 percentage points and L2-EM by 17.4 percentage points over SToLa. OMAE also decreases from 0.45 to 0.24, indicating that Touch-R1 reduces the ordinal distance of predictions from the ground truth.

4.3 Ablation and Analysis

Table 3 highlights how each component contributes to Touch-R1. Cold-start SFT provides the strongest initial boost, raising Avg from 28.0 to 47.8 by teaching the model the structured tactile reasoning format. Adding standard GRPO further improves Avg to 50.1. The Touch-R1 components address distinct failure modes. The ordinal reward lowers OMAE from 0.42 to 0.31, indicating that distance-aware feedback mitigates ordinal collapse. The consistency reward raises CSC from 56.1 to 69.2, giving the largest improvement on cross-sensor physical consistency. The format reward further improves Avg from 58.3 to 59.2 while preserving the parsable reasoning interface.

Scale analysis. Figure 5 compares the task-wise performance of Touch-R1 with 3B and 7B backbones. The 7B model improves over the 3B model on all reported metrics, with gains on L2-EM, SOI-ρ, CSC, and Avg. The Touch-R1 objective remains effective with both backbones.

Qualitative Analysis. Figure 6 illustrates cases where visual appearance is insufficient or misleading. Touch-R1 grounds its conclusions in tactile cues such as marker displacement and deformation patterns, leading to physically plausible predictions.

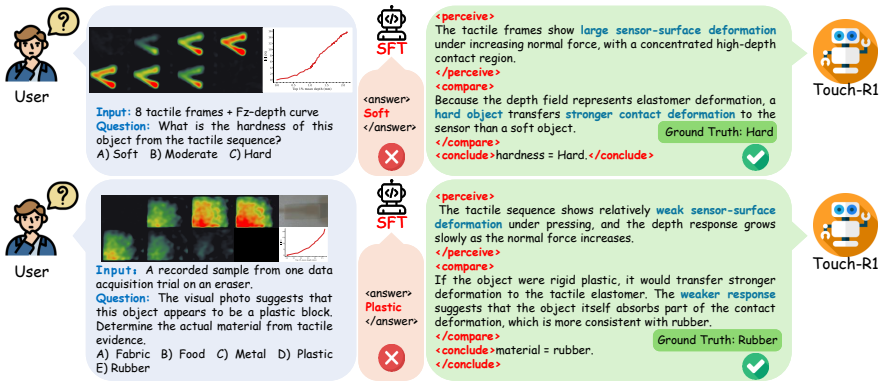


Figure 6: Qualitative examples of Touch-R1 reasoning. The model uses tactile evidence to estimate hardness and resolves visual-tactile conflicts in material recognition.

Table 5: Comparison on VTV150K. Results are reported in percentages, following the official evaluation protocol.

Model	Hardness	Protrusion	Elasticity	Friction	Combined	SFD	SOI	OSC	TSA	Average
GPT-4o [35]	34.7	32.6	32.6	18.7	2.1	40.9	38.4	16.6	36.0	28.0
Gemini-2.5-Pro-Exp [36]	36.2	34.7	39.1	21.0	4.3	42.6	29.4	18.5	40.0	29.5
LLaVA-OneVision-7B [38]	27.5	32.6	26.0	20.2	0.7	40.9	28.2	11.7	30.0	24.2
LLaVA-Video-Qwen2-7B [44]	30.4	29.7	28.9	18.1	2.1	33.6	29.4	17.2	36.0	25.0
InternVL2.5-VL-8B [40]	18.1	23.9	21.0	13.7	0.0	24.5	17.9	11.1	24.0	17.1
VideoLLaMA3-7B [45]	15.2	21.7	14.4	10.8	0.0	11.4	12.8	7.4	20.0	12.6
Qwen2.5-VL-7B [39]	25.3	28.9	17.3	15.9	1.4	22.9	28.2	16.0	30.0	20.6
VTV-LLM-7B [41]	73.9	75.0	67.3	56.5	35.6	71.3	57.6	43.2	64.0	60.4
Ours	**78.6**	**79.3**	**71.8**	**61.2**	**41.9**	**76.5**	**63.4**	**48.7**	**68.0**	**65.5**

5 Conclusion

Touch-R1 introduces the R1 paradigm to heterogeneous optical tactile reasoning. By combining ordinal-aware rewards, cross-sensor physical consistency, structured-format control, and input-side tactile grounding, Touch-R1 improves tactile reasoning accuracy, reduces ordinal error, and strengthens cross-sensor consistency. Experiments on TouchReason-Bench and external tactile-video benchmarks show that Touch-R1 can use physical contact evidence to revise misleading visual priors and produce structured perception-comparison-revision rationales.

References

[1] Samson Yu, Kelvin Lin, Anxing Xiao, Jiafei Duan, and Harold Soh. Octopi: Object property reasoning with large tactile-language models. *arXiv preprint arXiv:2405.02794*, 2024.

[2] Letian Fu, Gaurav Datta, Huang Huang, William Chung-Ho Panitch, Jaimyn Drake, Joseph Ortiz, Mustafa Mukadam, Mike Lambeta, Roberto Calandra, and Ken Goldberg. A touch, vision, and language dataset for multimodal alignment. *arXiv preprint arXiv:2402.13232*, 2024.

[3] Ning Cheng, Jinan Xu, Changhao Guan, Jing Gao, Weihao Wang, You Li, Fandong Meng, Jie Zhou, Bin Fang, and Wenjuan Han. Touch100k: A large-scale touch-language-vision dataset for touch-centric multimodal representation. *Information Fusion*, 124:103305, 2025.

[4] Ruoxuan Feng, Jiangyu Hu, Wenke Xia, Tianci Gao, Ao Shen, Yuhao Sun, Bin Fang, and Di Hu. Anytouch: Learning unified static-dynamic representation across multiple visuo-tactile sensors. *arXiv preprint arXiv:2502.12191*, 2025.

[5] Ning Cheng, Jinan Xu, Jialing Chen, Bin Fang, and Wenjuan Han. Stola: Self-adaptive touch-language framework for tactile commonsense reasoning in open-ended scenarios. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 18198–18206, 2026.

[6] Junyi Zong, Qingxuan Jia, Meixian Shi, Tong Li, Jiayuan Li, Zihang Lv, Gang Chen, and Fang Deng. Vitatouch: Property-aware vision-tactile-language model for robotic quality inspection in manufacturing. *arXiv preprint arXiv:2604.03322*, 2026.

[7] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[8] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[9] Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Xu Tang, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025.

[10] Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025.

[11] Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2034–2044, 2025.

[12] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*, 2025.

[13] Yi Chen, Yuying Ge, Rui Wang, Yixiao Ge, Junhao Cheng, Ying Shan, and Xihui Liu. Grpo-care: Consistency-aware reinforcement learning for multimodal reasoning. *arXiv preprint arXiv:2506.16141*, 2025.

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

[15] Xense Robotics. Photon - Tactile Sensor. `https://www.xenserobotics.com/product/367222/detail/15`, 2025. Accessed: 2026-05-06.

[16] Lunwei Zhang, Yue Wang, and Yao Jiang. Tac3d: A novel vision-based tactile sensor for measuring forces distribution and estimating friction coefficient distribution. *arXiv preprint arXiv:2202.06211*, 2022.

[17] Daimon (Shenzhen) Robotics Technology Co., Ltd. DM-Tac X. [Online]. Available: `https://www.dmrobot.com/en/product/p1/dm-tac-x.html`, 2026. Accessed: 2026-05-01.

[18] Weihao Yuan, Johannes A Stork, Danica Kragic, Michael Y Wang, and Kaiyu Hang. Rearrangement with nonprehensile manipulation using deep reinforcement learning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 270–277. IEEE, 2018.

[19] Weihao Yuan, Kaiyu Hang, Danica Kragic, Michael Y Wang, and Johannes A Stork. End-to-end nonprehensile rearrangement with deep reinforcement learning and simulation-to-reality transfer. *Robotics and Autonomous Systems*, 119:119–134, 2019.

[20] Jiaxin Shi, Xidong Zhang, Fucai Zhu, Zhe Li, Siyu Zhu, and Weihao Yuan. 3dthinkvla: Endowing vision-language-action models with latent 3d priors via 3d-thinking-guided co-training. *arXiv preprint arXiv:2606.04436*, 2026.

[21] Xidong Zhang, Yichi Zhang, Jiaxin Shi, Fucai Zhu, Siyu Zhu, Michael Yu Wang, Xiaojun Wu, and Weihao Yuan. Unitacvla: Unified tactile understanding and prediction in vision language action models. *arXiv preprint arXiv:2606.31723*, 2026.

[22] Mike Lambeta, Po-Wei Chou, Stephen Tian, Brian Yang, Benjamin Maloon, Victoria Rose Most, Dave Stroud, Raymond Santos, Ahmad Byagowi, Gregg Kammerer, et al. Digit: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation. *IEEE Robotics and Automation Letters*, 5(3):3838–3845, 2020.

[23] Fengyu Yang, Chao Feng, Ziyang Chen, Hyoungseob Park, Daniel Wang, Yiming Dou, Ziyao Zeng, Xien Chen, Rit Gangopadhyay, Andrew Owens, and Alex Wong. Binding touch to everything: Learning unified multimodal tactile representations. *arXiv:2401.18084*, 2024.

[24] Jialiang Zhao, Yuxiang Ma, Lirui Wang, and Edward H Adelson. Transferable tactile transformers for representation learning across diverse sensors and tasks. *arXiv preprint arXiv:2406.13640*, 2024.

[25] Carolina Higuera, Akash Sharma, Chaithanya Krishna Bodduluri, Taosha Fan, Patrick Lancaster, Mrinal Kalakrishnan, Michael Kaess, Byron Boots, Mike Lambeta, Tingfan Wu, et al. Sparsh: Self-supervised touch representations for vision-based tactile sensing. *arXiv preprint arXiv:2410.24090*, 2024.

[26] Fengyu Yang, Chenyang Ma, Jiacheng Zhang, Jing Zhu, Wenzhen Yuan, and Andrew Owens. Touch and go: Learning from human-collected vision and touch. *arXiv preprint arXiv:2211.12498*, 2022.

[27] Fucai Zhu, Xidong Zhang, Yichi Zhang, Yipan Wang, Fubo Wang, Zibin Wu, Kaiqi Li, Jianhao Jiao, Shihan Lu, Di Hu, et al. Tacclaw-infinity: A large-scale vision-tactile-language-action dataset for dexterous manipulation. *arXiv preprint arXiv:2610.20320*, 2026.

[28] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

[29] Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025.

[30] Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

[31] Jiaer Xia, Yuhang Zang, Peng Gao, Sharon Li, and Kaiyang Zhou. Visionary-r1: Mitigating shortcuts in visual reasoning with reinforcement learning. *arXiv preprint arXiv:2505.14677*, 2025.

[32] Roberto Calandra, Andrew Owens, Dinesh Jayaraman, Justin Lin, Wenzhen Yuan, Jitendra Malik, Edward H Adelson, and Sergey Levine. More than a feeling: Learning to grasp and regrasp using vision and touch. *IEEE Robotics and Automation Letters*, 3(4):3300–3307, 2018.

[33] Roberto Calandra, Andrew Owens, Manu Upadhyaya, Wenzhen Yuan, Justin Lin, Edward H Adelson, and Sergey Levine. The feeling of success: Does touch sensing help predict grasp outcomes? *arXiv preprint arXiv:1710.05512*, 2017.

[34] Yunzhu Li, Jun-Yan Zhu, Russ Tedrake, and Antonio Torralba. Connecting touch and vision via cross-modal prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10609–10618, 2019.

[35] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[36] Google DeepMind. Gemini 2.5 Pro Preview 05-06. [Online]. Available: `https://ai.google.dev/gemini-api/docs/models#gemini-2.5-pro-preview-05-06`, 2025. Accessed: 2026-05-06.

[37] Anthropic. Claude 3.5 Sonnet. [Online]. Available: `https://www.anthropic.com/news/claude-3-5-sonnet`, 2024. Accessed: 2026-05-06.

[38] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[39] S Bai, K Chen, X Liu, J Wang, W Ge, S Song, K Dang, P Wang, S Wang, J Tang, et al. Qwen2. 5-vl technical report (no. arxiv: 2502.13923). arxiv, 2025.

[40] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

[41] Yifan Xie, Mingyang Li, Shoujie Li, Xingting Li, Guangyu Chen, Fei Ma, Fei Richard Yu, and Wenbo Ding. Universal visuo-tactile video understanding for embodied interaction, 2025. URL `https://arxiv.org/abs/2505.22566`.

[42] Emily Denton and Rob Fergus. Stochastic video generation with a learned prior. In *International conference on machine learning*, pages 1174–1183. PMLR, 2018.

[43] Daniel Geng, Max Hamilton, and Andrew Owens. Comparing correspondences: Video prediction with correspondence-wise losses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3365–3376, 2022.

[44] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024.

[45] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025.

A Details of TouchReason Collection

A.1 Acquisition Device

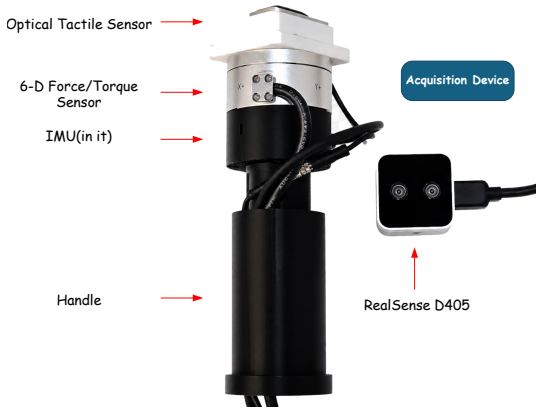


Figure 7: Acquisition setup with the handheld tactile-force device and an RGB-D camera.

As shown in Fig. 7, the handheld acquisition device consists of a handle, an IMU, a six-axis force/torque sensor, and an optical tactile sensor arranged from top to bottom. The handle allows human operators to perform natural pressing, sliding, and twisting motions, while the IMU is used to compensate for pose-dependent effects in the force measurements, yielding compensated 3D contact force signals for each tactile sequence. Although the device uses a six-axis force/torque sensor, we use the compensated three-axis contact force signals as the primary force labels in this work.

A.2 Object Visual Recording

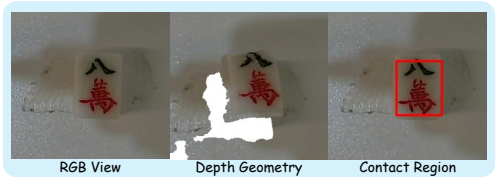


Figure 8: Object-level visual reference. The RealSense camera records RGB, depth, and contact-region information before tactile exploration.

Fig. 8 illustrates the object-level visual recording procedure. Before tactile exploration, each object is fixed in place and recorded with a RealSense RGB-D camera. The camera is positioned approximately along the normal pressing direction of the handheld device, providing a visual reference consistent with the tactile contact direction. We annotate the RGB image with a red bounding box to indicate the projected contact area of the optical tactile sensor. The resulting RGB-D observation provides the object appearance, depth geometry, and a contact-region reference for subsequent tactile acquisition.

A.3 Human-Held Tactile-Force Acquisition

During tactile exploration, a RealSense RGB-D camera is positioned at the side of the object to record the full acquisition process. The side-view RGB-D stream provides a visual reference for the contact state and the motion of the handheld device during interaction.

The tactile-force acquisition procedure consists of two stages. In the first stage, the operator starts from a non-contact state and gradually presses the tactile sensor against the object until the normal force reaches 20 N. This stage captures force-indexed normal deformation trajectories. Since the depth field describes the deformation of the tactile sensor elastomer, objects with different hardness

or compliance produce different depth responses under the same normal-force level. For contacts with comparable geometry, harder objects tend to induce more pronounced elastomer indentation, while softer objects may absorb part of the contact deformation through their own compliance.

In the second stage, the operator again starts from a non-contact state and performs exploratory motions at two target normal-force levels: 5 N and 15 N. At each force level, the operator performs six motions: upward, downward, leftward, and rightward shear-inducing motions, together with clockwise and counterclockwise rotations. The four directional motions are defined in the tactile image plane and are designed to produce tangential deformation on the tactile elastomer while minimizing macroscopic relative sliding between the sensor and the object surface. The rotational motions are used to capture torsional contact responses. This stage provides force-conditioned tactile responses under different loading conditions.

Representative examples of the pressing stage are shown in Fig. 9 and Fig. 10.

A.4 Temporal Synchronization and IMU-Aided Force Compensation

Due to the heterogeneous imaging pipelines of different tactile sensors, the tactile streams are recorded at their native frame rates. Specifically, GelSight Mini, Xense, and Tac3D are recorded at 30 Hz, while DM-Tac X is recorded at 120 Hz. In addition, the IMU is sampled at 400 Hz and the force sensor is sampled at 1 kHz. To preserve the original sensor information, all raw streams are stored before any temporal alignment or offline processing, including raw tactile images, RealSense RGB-D frames, IMU readings, and force-sensor measurements.

All modalities are timestamped using the host clock during acquisition. In post-processing, we use the tactile stream as the reference timeline, since tactile frames provide the primary observations for contact perception. For each tactile frame, the corresponding force measurement is obtained by temporal interpolation from the force-sensor stream. The IMU readings are associated by nearest-neighbor matching or interpolation, while the RealSense RGB-D frames are selected according to the nearest timestamp. The time offsets between modalities are also retained for data-quality analysis.

The force measurements are compensated using the IMU readings and the calibrated spatial relationship between the IMU and the force sensor. The IMU outputs quaternions (q_0, q_1, q_2, q_3), which are used to compute the rotation from the IMU/body frame to the world frame. This rotation is used to express the gravity direction and acceleration measurements in a common coordinate frame, and to transform the predicted non-contact force components into the force-sensor frame. Since the accelerometer readings (a_x, a_y, a_z) contain gravitational acceleration, the gravity component is handled according to the quaternion-derived orientation before estimating the translational acceleration of the handheld device. The gyroscope readings $(\omega_x, \omega_y, \omega_z)$ are further used to account for rotational effects. The angular acceleration is estimated from adjacent gyroscope readings and their time interval.

Based on the calibrated mass properties, sensor bias, and the spatial relationship between the IMU and the force sensor, we estimate the force components caused by the device assembly under the current pose and motion state. These non-contact force components are transformed into the force-sensor frame and subtracted from the raw force readings. This IMU-aided compensation reduces pose- and motion-dependent disturbances introduced by handheld operation and yields compensated 3D contact force signals for each tactile frame.

Both the raw force measurements and the compensated force signals are stored in the dataset. In our benchmark and annotation pipeline, the compensated 3D force is used as the primary force label, while the raw measurements are retained for reproducibility and further analysis. This synchronization and compensation procedure allows each tactile sequence to be paired with temporally aligned visual observations, IMU states, and contact-force information.

A.5 Offline Tactile Processing and Filtering

After acquisition, all raw tactile frames are first stored as lossless image sequences and then processed offline using the corresponding sensor SDKs. For each tactile sensor, the raw tactile images are fed into its dedicated processing pipeline to estimate tactile representations, including the tangential shear field and the normal depth field. The shear field describes contact-induced tangential

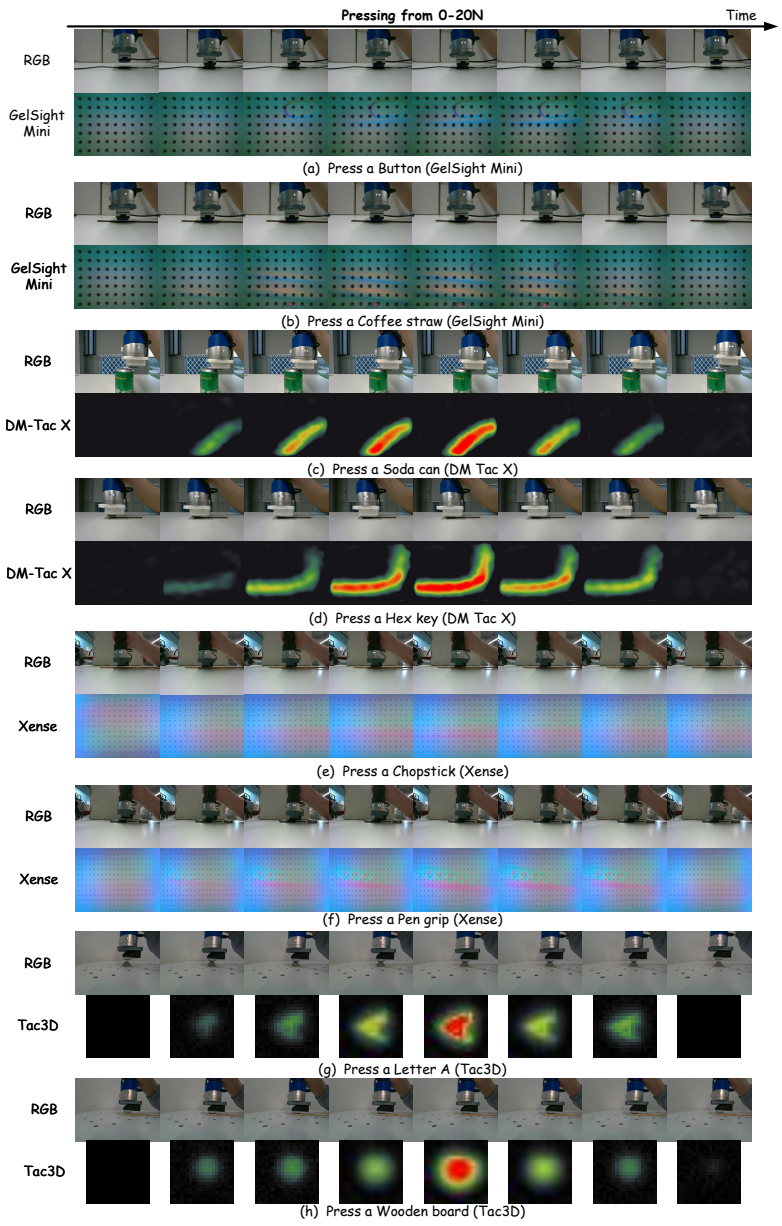


Figure 9: Pressing sequences from 0 to 20 N. Each example shows side-view RGB frames and the corresponding tactile observations during continuous normal pressing with GelSight Mini, Xense, Tac3D, and DM-Tac X. The sequence illustrates force-indexed tactile deformation as the contact force increases over time.

deformation on the tactile elastomer, while the depth field reflects the normal indentation response of the elastomer surface. We retain both the raw tactile images and the processed tactile fields to support different levels of tactile perception and downstream analysis.

For the force signal, we use the first frame of each acquisition sequence as the initial reference. Since each sequence starts from a non-contact state, the compensated 3D force at frame 0 is treated as the sequence-specific baseline. For each subsequent tactile frame, the final force label is computed by subtracting this initial force value from the temporally aligned compensated force measurement:

$$\Delta \mathbf{F}_t = \mathbf{F}_t^{\text{comp}} - \mathbf{F}_0^{\text{comp}},$$

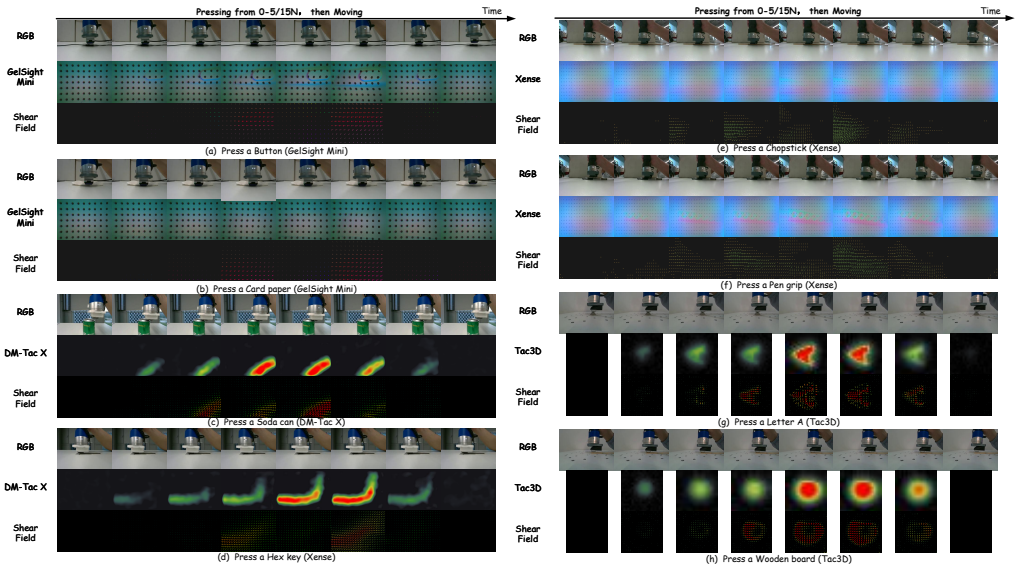


Figure 10: Exploratory motion sequences under target force levels. Each example shows side-view RGB frames, tactile observations, and shear fields during force-conditioned exploratory motions. The operator first presses the sensor to the target normal force and then performs representative shear-inducing motions, producing motion-dependent tactile responses over time.

where $\mathbf{F}_t^{\text{comp}} \in \mathbb{R}^3$ denotes the compensated 3D force aligned with the t-th tactile frame. This zero-offset operation removes the sequence-level bias and represents the contact-induced force change during tactile interaction.

We further filter invalid samples before constructing the final dataset. Sequences with failed contacts, contact regions outside the valid tactile sensing area, saturated tactile images, abnormal SDK outputs, unstable force traces, or invalid RGB-D observations are removed. After filtering, each valid sequence contains synchronized raw tactile images, shear fields, depth fields, side-view RGB-D recordings, front-view RGB-D observations, IMU states, and zero-offset 3D contact force labels.

B Implementation details and benchmark definitions

B.1 Training hyperparameters

Table 6: Training hyperparameters for cold-start supervised fine-tuning and Touch-R1 reinforcement learning.

Hyperparameter	Cold-start SFT	Touch-R1 RL
Optimizer	AdamW β_1=0.9, β_2=0.95	AdamW
Peak LR	1×10^{-5}	1×10^{-6}
Schedule	cosine, 3% warm-up	linear, 20-step warm-up
Global batch	64	256 prompts $\times$ 8 rollouts
Sequence length	4096	4096 + 4096 gen
Frames per sequence	8	8
Image resolution	448×448	448×448
Epochs	1	≈ 1.5
KL coeff β	–	1×10^{-3}
Clip-Higher $\varepsilon_l, \varepsilon_h$	–	$0.2, 0.28$
Group size G	–	8
Dynamic Sampling	–	enabled
Loss aggregation	token	segment-level
Reward weights	–	$0.6, 0.3, 0.1$
Hardware	16×H200 140G	16×H200 140G
Compute	$\approx$ 14 GPU-days	$\approx$ 62 GPU-days

B.2 TouchReason-Bench task families and metrics

TouchReason-Bench is organized to separate direct perception, compositional attribute reasoning, comparative reasoning, and visual-tactile conflict resolution. Table 7 gives the metric definitions used in the main experiments.

Table 7: Task families and metrics in TouchReason-Bench. The examples are abbreviated templates; generated prompts instantiate them with object metadata, sensor names, and tactile-frame sequences.

Tier	Metric	Evaluated ability	Example template
L1 perception	H/R/P/Mat-Acc	Single physical attribute recognition	"What is the hardness/material of this tactile sequence?"
L1 ordinal	OMAE↓	Distance-aware ordinal prediction	"Rate hardness from 1 to 3 and justify with tactile evidence."
L2 joint	L2-EM, L2-MCQ, L2-MS	Joint attribute consistency	"Select the full tuple: hardness, roughness, protrusion, material."
L3 pair/rank	SFD, SOI-ρ	Pairwise and listwise tactile comparison	"Which sequence is softer / has stronger protrusion?"
L3 force	FRC	Force-conditioned contact dynamics	"How does the response change from low to high normal force?"
L4 conflict	CSC, VTC, V-Trick	Cross-sensor and visual-tactile contradiction handling	"Vision suggests metal; tactile evidence suggests compliance. Decide."

C Limitations

Touch-R1 improves tactile-grounded reasoning, but several limitations remain. First, TouchReason focuses on optical tactile sensors with high-resolution elastomer imagery. The current model is evaluated across GelSight-family optical sensors, Tac3D, and DM-Tac X, but it may require additional calibration or adaptation for sensors that output sparse force arrays, capacitive taxels, or event-based tactile streams. Second, while TouchReason contains 1000+ filtered objects and substantially more sequences than prior tactile-reasoning datasets, the long tail of real-world materials, liquids, powders, deformable cloth folds, and fragile objects is still underrepresented. Third, our strongest model uses a 7B-scale multimodal backbone and long reasoning traces, which are useful for analysis but costly for real-time robotic deployment. Fourth, most benchmark prompts focus on short tactile clips and bounded property queries. Long-horizon manipulation, active touch planning, and memory over repeated contacts remain open problems.

D Broader impacts

The positive applications of Touch-R1 include physical-property understanding for assistive manipulation, safer human-robot interaction, prosthetic feedback analysis, industrial inspection, and agricultural sorting. The dataset also makes tactile reasoning more measurable by requiring models to justify decisions with contact evidence rather than visual appearance alone.

Potential risks should be considered. High-resolution tactile sensing can reveal material identity, object condition, and contact patterns that may be sensitive in private or medical contexts. Tactile reasoning models could also be misused for covert physical inspection if deployed without consent. Finally, tactile hardware and data-collection expertise are still expensive, which may concentrate the benefits of this technology in well-funded labs. We therefore recommend transparent data provenance, consent-aware deployment, and explicit exclusion of human-identification or covert measurement use cases.

E Future tactile prediction and Stage-1 diagnostics

Following the qualitative style of Touch-and-Go, we include a direct history-to-future tactile prediction visualization in Fig. 11. The left block shows synchronized RGB and tactile history frames from a held-out TouchReason pressing sequence, while the right block compares the future tactile ground truth with multimodal and touch-only predictions. The prediction model uses four history frames and predicts the next tactile frame from the Stage-1 tactile latent. Both predictions recover the coarse contact geometry, while the absolute-error column highlights remaining high-frequency texture error around the reconstructed deformation field. We therefore use the panel as a qualitative diagnostic: it verifies that the learned future predictor can recover contact geometry from synchronized vision-touch histories, and it also exposes where the latent decoder still adds local speckle.

Stage 1 learns the tactile representation used by later instruction tuning, future prediction, and reinforcement learning. As a complementary diagnostic, we inspect held-out reconstructions from

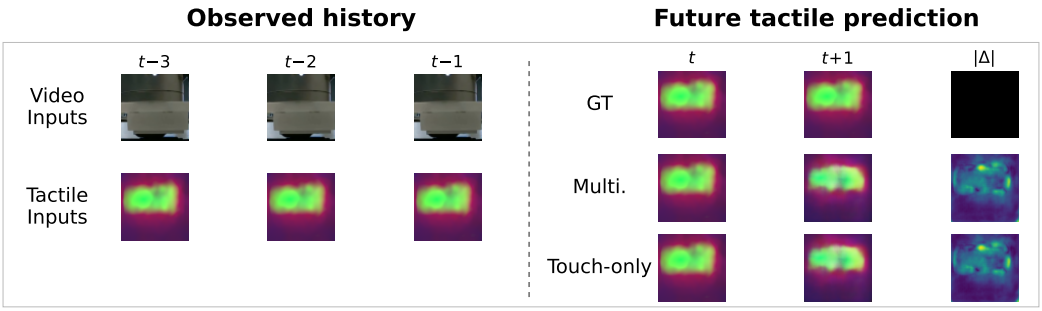


Figure 11: Future tactile prediction in the style of Touch-and-Go, rendered from TouchReason streams. The model observes RGB and tactile frames at indices 105–108 and predicts the tactile response at frame 109. The right panel compares the ground truth, the multimodal prediction conditioned on RGB and tactile history, and the touch-only prediction obtained by dropping the RGB stream at inference; the last column shows the absolute error $|\Delta|$.

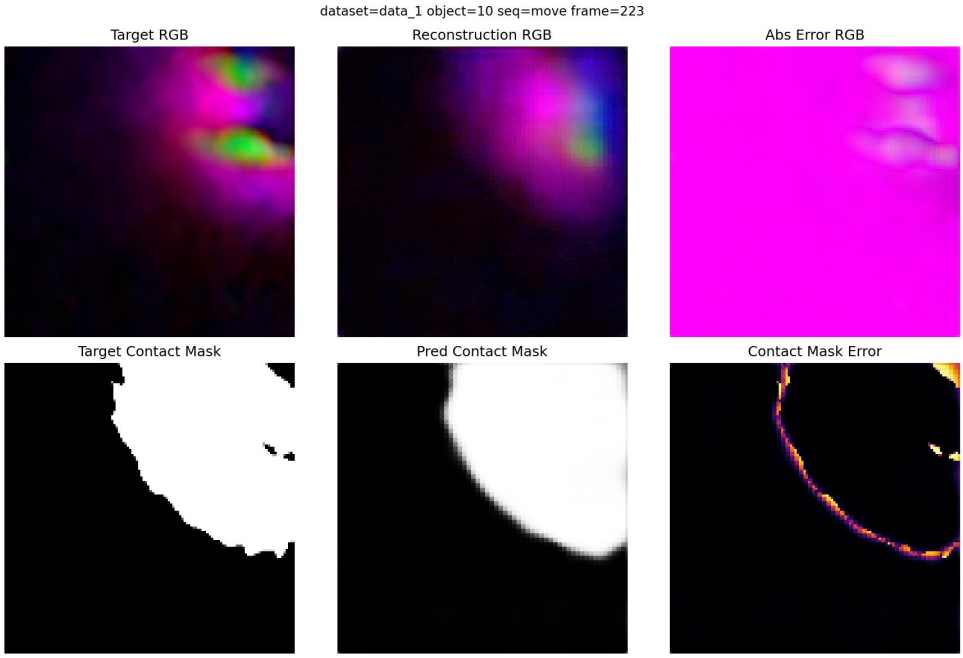


Figure 12: Representative Stage-1 tactile reconstruction on a held-out move frame from the full-data run. The panel compares the target tactile RGB frame, reconstructed RGB frame, absolute reconstruction error, target contact mask, predicted contact mask, and mask error. The sample obtains MAE $= 0.018$, RMSE $= 0.060$, contact IoU $= 0.951$, and center error $= 0.0016$, indicating that the Stage-1 encoder preserves the dominant contact geometry used by later tactile reasoning stages.

the full-data run. Figure 12 shows a representative move sequence frame from the reconstruction/tokenizer branch. The model reconstructs the dominant RGB tactile response and recovers the main contact region with low pixel error (MAE $= 0.018$) and high contact overlap (IoU $= 0.951$). The residual map shows that most remaining errors concentrate near fine contact boundaries and high-frequency texture, which is consistent with the smoothing artifacts expected from reconstruction-based tactile representation learning.

We also include a synchronization diagnostic in Fig. 15. The tactile stream is used as the reference timeline; force values are interpolated to each tactile frame, IMU readings are associated by nearest-neighbor matching or interpolation, and RGB-D frames are selected by nearest timestamp. In the displayed held-out pressing sequence, the force offset is sub-millisecond at the median and the RGB offset is within a single-digit millisecond range. The aligned tactile frames track the rise and release

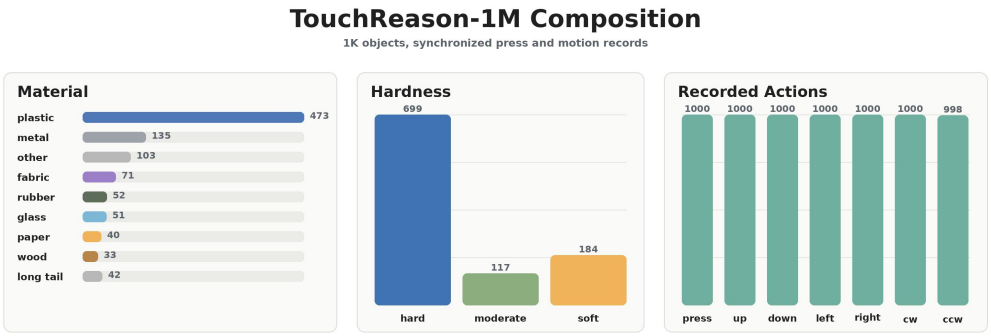


Figure 13: TouchReason composition. The panel follows the compact dataset-summary style used in Touch-and-Go: material and hardness bars summarize object-level metadata, while the action bars summarize the recorded pressing and exploratory motion coverage. Counts are shown at release scale for the 1000+ object dataset.

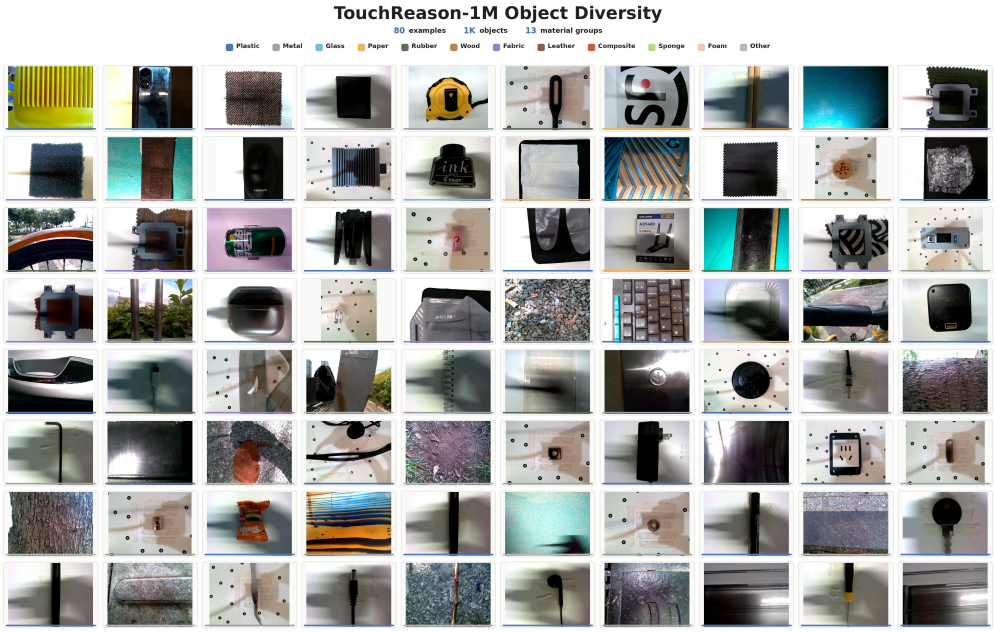


Figure 14: TouchReason object diversity gallery. The panel shows 80 non-duplicate object-family examples sampled from the object collection. Each tile contains an object-level RGB reference, and the thin color strip denotes the material group.

of both normal force and contact area, indicating that the synchronized streams preserve the contact dynamics needed by force-conditioned tactile reasoning.

F Details for multimodal future tactile prediction

The future tactile prediction diagnostic is built on top of the Stage-1 tactile autoencoder. Each tactile frame is encoded as a six-channel tensor: normalized horizontal and vertical deformation, normalized depth, normalized horizontal and vertical shear, and a contact mask derived from deformation, shear, and depth magnitude. A temporal predictor consumes four past tactile latents. The multimodal variant additionally encodes the aligned RealSense RGB history with a convolutional image encoder, concatenates the visual features with the tactile latent history and force context, and predicts the latent of the next tactile frame. The touch-only variant reuses the same predictor while

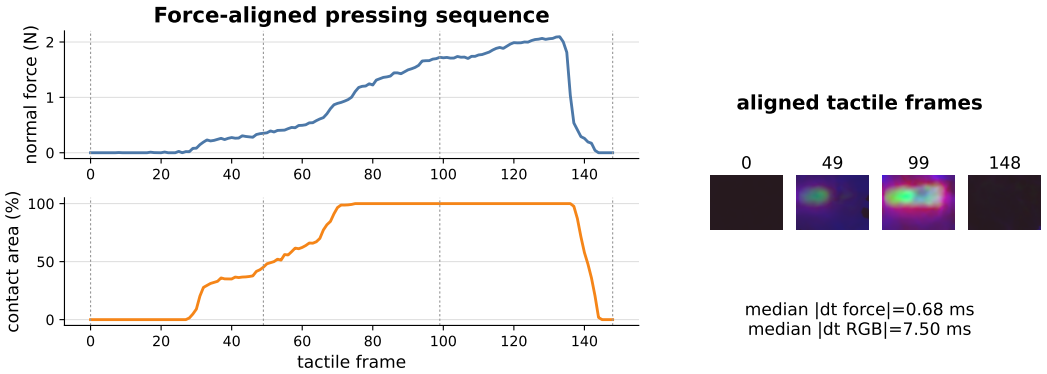


Figure 15: Synchronization diagnostic for a held-out TouchReason pressing sequence. The left panels plot compensated normal force and tactile contact area over the tactile-frame timeline; the right panels show representative aligned tactile frames and median offsets to the force and RGB streams.

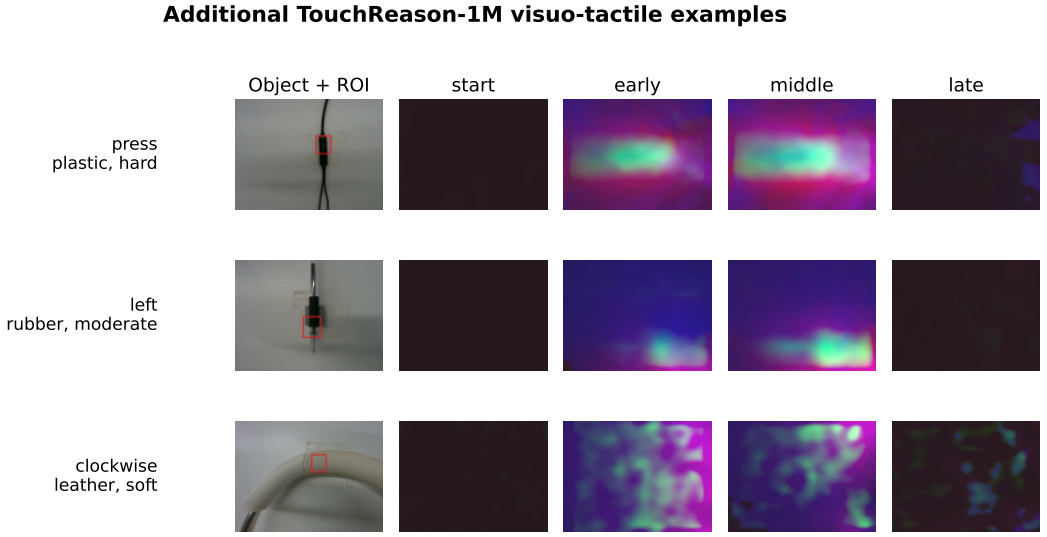


Figure 16: Additional TouchReason examples. Each row shows the object-level RGB image with the contact ROI and four tactile frames sampled across one interaction sequence. The tactile rendering uses the same channels as Fig. 11: deformation magnitude, depth, shear magnitude, and contact mask. These examples complement the acquisition figures by showing the raw visual-tactile diversity available before instruction generation.

dropping the visual stream at inference, which isolates the contribution of RGB context without changing the tactile history.

We train the future module with Adam at learning rate 2×10^{-4}, following the Touch-and-Go future-prediction setting but adapting the input representation to our force-aligned optical tactile fields. The full-data run uses four history frames, one future step, 224×224 RGB crops, 128×128 tactile fields, global tactile normalization, and object-level train/validation/test splits. On the full all-data test split, the temporal predictor obtains MAE $= 0.0220$ and contact IoU $= 0.9255$, while the world-model variant obtains MAE $= 0.0249$, contact IoU $= 0.8920$, and a lower artifact score than the temporal predictor. For the selected qualitative sample, the world model improves over the temporal predictor (MAE 0.0241 vs. 0.0326, IoU 0.9624 vs. 0.9555), which is why we use it as the primary qualitative future-prediction checkpoint.

G Prompt templates

System. "You are an assistant specialized in physical-property reasoning from optical tactile sensing. The input is a frame sequence from one of the TouchReason optical tactile sensors, where each frame is an elastomer-deformation snapshot. Output in three segments: `<perceive>` describes marker displacement amplitude, texture visibility, and deformation extent per frame; `<compare>` compares candidate answers along the ordinal axis and explicitly notes any visual-tactile contradiction; `<conclude>` returns the final structured attributes."

Level 1 hardness prompt. `[Sensor: GelSight Mini] [8 frames] What is the hardness of this object? A) Soft B) Moderate C) Hard. Strictly follow the three-segment format.`

Level 4 V-Trick prompt. `[Visual photo] + [Sensor: DM-Tac X, 8 frames] The visual photo suggests this object appears metallic. The tactile sequence may contradict the appearance. Determine the actual material from tactile evidence: A) Fabric B) Food C) Metal D) Plastic E) Rubber.`

H Complete TouchReason-Bench question examples

H.1 Level 1 single-property perception

Type	Question	GT	Scoring
H-Acc 3-way	A) Soft B) Moderate C) Hard	B	exact match
H-Acc Y/N	Is this object hard?	No	exact match
R-Acc	A) Smooth B) Slightly rough C) Rough	A	exact match
P-Acc	A) None B) Small C) Large protrusion	B	exact match
Mat-Acc	A) Fabric B) Food C) Metal D) Plastic E) Rubber	D	exact match
OMAE	Hardness rating, integer 1–3	2	$r{=}\mathbb{1}[d{=}0]\cdot 1.0{+}\mathbb{1}[d{=}1]\cdot 0.4$

H.2 Level 2 joint description

L2-EM. The full template "hardness=[X], roughness=[Y], protrusion=[Z], material=[M]" must be filled with all four attributes correct to receive credit.

L2-MCQ4. Three distractors differ from the correct option by one or two attributes. For `leather_wallet`, option A reads moderate, smooth, none, fabric and is wrong on roughness; option B reads moderate, slightly rough, none, fabric and is the ground truth; option C reads hard, slightly rough, none, fabric and is wrong on hardness; option D reads moderate, slightly rough, none, plastic and is wrong on material.

L2-MS. Multi-select from 14 attribute terms, scored by F1.

H.3 Level 3 comparative reasoning

SFD-Pair. "Compare A and B; which is softer?" single choice.

SFD-Attr. Pairwise comparison along a specified attribute.

SOI-Best. "Among three sequences, which has the strongest protrusion?"

SOI-Rank. Order four sequences from softest to hardest, scored by Spearman ρ with $\rho = 1.0$ giving full credit.

FRC-Compare. Same object at 1 N versus 10 N. "Which attribute changes most? A) hardness B) roughness C) protrusion D) almost no change." Dataset-unique.

FRC-Numerical. Estimate the maximum normal force in a pressure-ramp sequence within ± 2 N tolerance.

H.4 Level 4 cross-modal reasoning

CSC. "Two sequences from different optical tactile sensors show the same object. Are the inferred attributes consistent? Identify any divergence."

VTC. "Visual cues suggest metal but tactile cues suggest low hardness. What is the true material? Explicitly state the visual-tactile contradiction."

V-Trick. Silicone apple replica. "This object visually matches a real apple. Determine the true material from tactile evidence alone."

I Additional case studies

Case 1, hardness on Tac3D, towel. GPT-4o predicts "hardness = hard", incorrect. Touch-R1: `<perceive>` "markers shift almost entirely on initial press; deformation in the third frame exceeds 60% of the field of view" `<compare>` "such large deformation only matches low compliance, contradicting the hard option" `<conclude>` "hardness = soft; material = fabric". Correct.

Case 2, cross-sensor consistency. The GelSight Mini sequence shows large marker displacement; the DM-Tac X sequence presents the same contact as a different raw color and deformation-field pattern because of its distinct illumination and optical geometry. Vanilla GRPO predicts "soft" on GelSight Mini and "hard" on DM-Tac X, the canonical M2 failure. Touch-R1 writes in the compare segment that the two sensors differ in appearance but agree on strong contact deformation, producing "soft" on both ends.

Case 3, FRC-Numerical. Ground truth maximum force is 7 N over 12 frames. Touch-R1: "radial marker displacement saturates by the eighth frame; given the marker grid spacing and elastic-modulus prior, the maximum force is approximately 6 N". The 1 N error is within tolerance and receives full reward.

J Data annotation protocol

Each object cycles through the four optical tactile sensors used in TouchReason. For each sensor, the acquisition protocol includes a continuous normal pressing ramp up to 20 N and six exploratory motions at two target normal-force levels, 5 N and 15 N. Sequences are captured at the native frame rate of each sensor; only frames near keyframes that satisfy a sensor-specific marker-displacement threshold $|\Delta\mathbf{m}| > \tau$ are retained, with eight frames before and after, stored as lossless PNG files. Two trained annotators independently label each object, with disagreements resolved by a third reviewer. Hardness is calibrated by Shore-A durometer with thresholds < 30 for soft, 30–70 for moderate, and > 70 for hard. Roughness uses the optical profilometer Sa value with < 10 μm for smooth, 10–50 μm for slightly rough, and > 50 μm for rough. Protrusion uses height with < 0.5 mm for none, 0.5–2 mm for small, and > 2 mm for large. Material follows manufacturer specifications. Cohen's κ is 0.83 for hardness, 0.79 for roughness, 0.81 for protrusion, and 0.94 for material.

Table 8: TouchReason attribute annotation scheme. Ordinal attributes are evaluated with distance-aware rewards in Touch-R1; material is treated as a categorical attribute.

Attribute	Level 0 / low	Level 1 / middle	Level 2 / high	Primary evidence
Hardness	Soft, large compliance	Moderate deformation	Hard, small object compliance	Force-depth slope, deformation extent
Roughness	Smooth	Slightly rough	Rough	Shear fluctuation, texture-induced marker motion
Protrusion	Flat/no protrusion	Small local protrusion	Large/localized protrusion	Contact-mask shape, depth peaks
Elasticity	Little rebound	Partial rebound	Clear rebound	Release trajectory and hysteresis
Geometry	Concave	Convex	Mixed concave-convex	RGB-D view, contact ROI, depth field
Material	Food, fabric, metal, plastic, rubber and other release categories			Metadata, visual reference, tactile response

K Question template generation

Table 9 lists representative TouchReason template families. Unlike pure tactile-caption templates, our prompts require a three-segment answer and include adversarial visual-only or tactile-only dis-

tractors when applicable. The final benchmark instances are generated from these templates and then filtered by text-only and visual-only shortcut tests, as described in the main paper.

Table 9: Representative TouchReason template families.

Family	Goal	Template sketch
Tactile feature assessment	Direct property recognition	"Given `<sensor>` tactile frames, determine `<attribute>` and cite tactile evidence."
Joint description	Multi-attribute consistency	"Fill hardness, roughness, protrusion, and material; all fields must agree with the sequence."
Surface feature distinction	Pairwise tactile comparison	"Between `<seq1>` and `<seq2>`, which is `<comparison>` along `<attribute>`?"
Surface optimality identification	Listwise ranking/selection	"Among `<seq1>`, `<seq2>`, and `<seq3>`, select the softest or most protruded object."
Force-response comparison	Dynamics under load	"Compare low-force and high-force clips; identify the attribute whose tactile response changes most."
Cross-sensor consistency	Sensor-invariant reasoning	"Two sensors touch the same object. State whether the inferred physical properties agree."
Visual-tactile conflict	Revision from misleading appearance	"The RGB view suggests `<visual material>`; use tactile evidence to decide the true property."

L More experimental results

L.1 Additional implementation details

Stage 1 uses the full tactile-sequence pool to learn a tactile representation before language fine-tuning. Tactile frames are resized to 448×448 for MLLM training, while the future-prediction side diagnostic uses 128×128 tactile fields and 224×224 RGB crops. During cold-start SFT, all examples are formatted with `<perceive>`, `<compare>`, and `<conclude>` segments. During T-GRPO, we sample eight rollouts per prompt and compute rewards from final-answer correctness, ordinal distance, cross-sensor consistency, and format validity. The consistency reward is evaluated only when paired sensors or tactile-control rollouts are available, so it does not create spurious pressure on single-sequence examples.

L.2 Additional ablations and robustness

The main paper reports the cumulative effect of SFT, ordinal reward, consistency reward, and format reward. Table 10 reorganizes these results in the supplementary style used by prior tactile-language work. We also report the VTV150K transfer comparison from the main experiments to make the external-dataset result easy to find in the supplement.

Table 10: Supplementary ablation summary. The rows reuse the same TouchReason-Bench evaluation protocol as Table 2; Avg is the aggregate score used throughout the paper.

Setting	OMAE↓	CSC	Avg	Main effect
Qwen2.5-VL-7B zero-shot	0.83	31.4	28.0	Visual-language prior only
+ cold-start SFT	0.46	52.6	47.8	Teaches the output format and tactile vocabulary
+ binary GRPO	0.42	54.3	50.1	Reinforces final-answer correctness
+ ordinal reward	0.31	56.1	54.6	Reduces distance along ordered attributes
+ consistency reward	0.27	69.2	58.3	Improves cross-sensor agreement
+ format reward	0.26	70.4	59.2	Stabilizes the three-segment response
Touch-R1	**0.24**	**71.3**	**60.1**	Full recipe

Table 11: External tactile-language robustness on VTV150K. The numbers match the VTV150K protocol in the main experiments and are repeated here for completeness.

Model	Hardness	Protrusion	Elasticity	Friction	Combined	Average
Qwen2.5-VL-7B	25.3	28.9	17.3	15.9	1.4	20.6
LLaVA-Video-Qwen2-7B	30.4	29.7	28.9	18.1	2.1	25.0
VTV-LLM-7B	73.9	75.0	67.3	56.5	35.6	60.4
Touch-R1	**78.6**	**79.3**	**71.8**	**61.2**	**41.9**	**65.5**

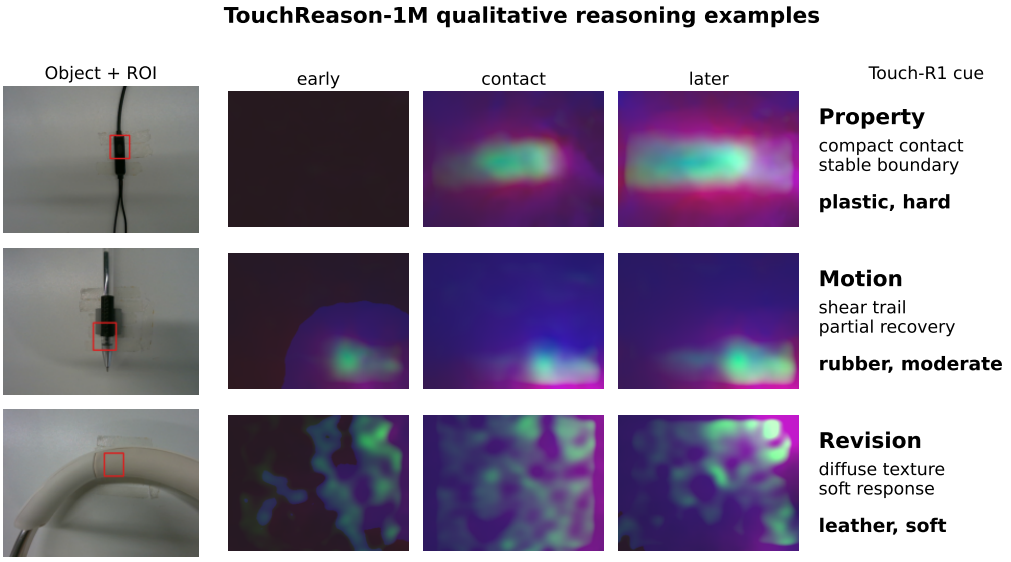


Figure 17: Additional qualitative TouchReason reasoning examples. Each row pairs an object-level RGB reference with three tactile frames and a concise Touch-R1-style evidence cue. The examples cover direct property perception, motion-induced shear reasoning, and visual-tactile revision.

L.3 Additional qualitative reasoning examples

Figure 6 in the main paper shows compact qualitative results. Table 12 provides additional text-only examples in the style of prior supplements. The examples are intentionally concise so that the physical evidence remains visible: each correct answer cites a concrete tactile cue, such as deformation magnitude, shear response, force-conditioned contact growth, or cross-sensor agreement.

Table 12: Additional qualitative examples of Touch-R1 reasoning.

Task	Query type	Typical baseline failure	Touch-R1 behavior
Tactile feature assessment	Hardness from one sequence	Chooses hard from object category prior	Cites large deformation and concludes soft
Surface distinction	Which object is rougher?	Treats blurred tactile colors as visual noise	Compares shear fluctuation and contact texture
Surface optimality	Select most protruded object	Picks visually salient object	Uses localized depth/contact peak to select protrusion
Force response	Low-force vs high-force clips	Ignores normal-force context	Notes contact-area growth and force-depth slope
Cross-sensor consistency	Same object, two sensors	Predicts different labels across sensors	Separates sensor appearance from shared deformation geometry
Visual-tactile conflict	Silicone fruit / metal-like object	Follows RGB appearance	Revises conclusion from tactile compliance evidence

M Computational resources and carbon footprint

Total training cost is approximately 76 GPU-days on H200 140G. With PUE 1.2 and 0.4 kg CO_2e per kWh, the estimate is approximately 360 kg CO_2e. Part of the electricity comes from renewable sources.

N NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: The abstract and introduction state the paper's main contributions, including Touch-R1, TouchReason-1M, TouchReason-Bench, tactile-specific RL failure modes, and the corresponding reward and grounding designs. The claims are supported by the experiments on TouchReason-Bench, ablation studies, tactile video prediction, and VTV150K transfer evaluation.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes].

   Justification: The paper discusses limitations regarding the current focus on optical tactile sensors, the finite set of sensor types and material categories, the cost of multi-stage MLLM training, and the need for further validation in long-horizon real-robot interaction. These limitations are included in the limitations discussion and supplemental material.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A].

   Justification: The paper does not present formal theoretical results, theorems, or proofs. The main contributions are methodological and empirical.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes].

   Justification: The paper describes the training pipeline, model backbone, datasets, data splits, reward definitions, evaluation benchmarks, metrics, and hyperparameters needed to reproduce the main results. Additional implementation details are provided in the appendix, and code and data will be released.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes].

   Justification: We will release TouchReason-1M, TouchReason-Bench, preprocessing scripts, training code, evaluation scripts, and instructions for reproducing the main experiments. The release will be anonymized during submission if required and made public after acceptance.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes].

   Justification: The experimental setup specifies the three-stage training protocol, disjoint data splits, model backbone, input resolution, number of tactile frames, optimizer, learning rates, GRPO group size, KL coefficient, compute hardware, benchmark construction, and evaluation metrics. Additional hyperparameters and implementation details are provided in Appendix B.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No].

   Justification: We do not report full error bars for all main experiments because multi-stage training of MLLMs with tactile sequences is computationally expensive. To improve reliability, we evaluate on multiple benchmarks, include extensive ablations, analyze data and reward sensitivity, and compare against several closed-source, open-source, and tactile-specialist baselines.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes].

   Justification: The paper reports the main compute environment, including 16 NVIDIA H200 GPUs with 140GB memory, and provides training-stage details for tactile pretraining, cold-start SFT, and GRPO. Additional compute and runtime details are provided in the supplemental material.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes].

   Justification: The research conforms to the NeurIPS Code of Ethics. The work focuses on tactile reasoning for embodied interaction, does not involve personal data or human-subject experiments, and will release data and code with documentation and responsible-use instructions.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes].

    Justification: The paper discusses positive impacts such as improving tactile understanding for robotics, assistive manipulation, material recognition, and embodied interaction. It also discusses potential risks from incorrect tactile reasoning in real-world robotic systems, including unsafe manipulation decisions or object damage, and emphasizes the need for deployment-time safety validation.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A].

    Justification: The released assets are tactile reasoning data, benchmarks, and models specialized for physical-property understanding, and they do not contain personal information, unsafe scraped content, or generative capabilities with a high risk of misuse. We will still provide documentation, licenses, and intended-use notes with the release.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes].

Justification: The paper cites and credits the existing models, datasets, and baselines used in the experiments, including Qwen2.5-VL, Touch-and-Go, VTV150K, and prior tactile-language or tactile-video methods. License and usage information for external assets is documented in the supplemental material and respected in our release.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes].

    Justification: The paper introduces TouchReason-1M and TouchReason-Bench, and documents their sensor types, object splits, task categories, annotation format, evaluation metrics, and intended use. Documentation, preprocessing scripts, and benchmark evaluation instructions will be released with the assets.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A].

    Justification: The paper does not involve crowdsourcing experiments or research with human subjects. The data are collected from objects and tactile sensors rather than human participants.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A].

    Justification: The paper does not involve human subjects, human-subject data, crowdsourcing studies, or participant interaction. Therefore, IRB approval or equivalent human-subject review is not applicable.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes].

    Justification: The paper describes the use of Qwen2.5-VL-7B-Instruct as the MLLM backbone and includes closed-source and open-source MLLMs as evaluation baselines. Gemini 2.5 Pro was used to generate candidate reasoning-style QA pairs and rationales from human-labeled tactile data pairs, followed by rule-based filtering and human evaluation. LLM-assisted language polishing, if used, was limited to writing and editing and did not affect the methodology, experiments, or scientific claims.