TouchThinker: Scaling Tactile Commonsense Reasoning to the Open World with Large-scale Data and Action-aware Representation

Kailin Lyu[1,2], Di Wu[1], Pengwei Zhang[1,2], Yuhang Zheng[3], Yingxin Lai[4], Long Xiao[1], Kangyi Wu[5], Pengna Li[5], Chen Gao[3], Lianyu Hu[6], Xiaobin Hu[3,†], Jie Hao[1,†], Ce Hao[2,†], Weihao Yuan[7], Shuicheng Yan[3,†]

[1]Institute of Automation, Chinese Academy of Sciences
[2]Zhongguancun Academy, [3]National University of Singapore, [4]Xiamen University
[5]Xi'an Jiaotong University, [6]Nanyang Technological University, [7]Nanjing University
[†]Corresponding authors

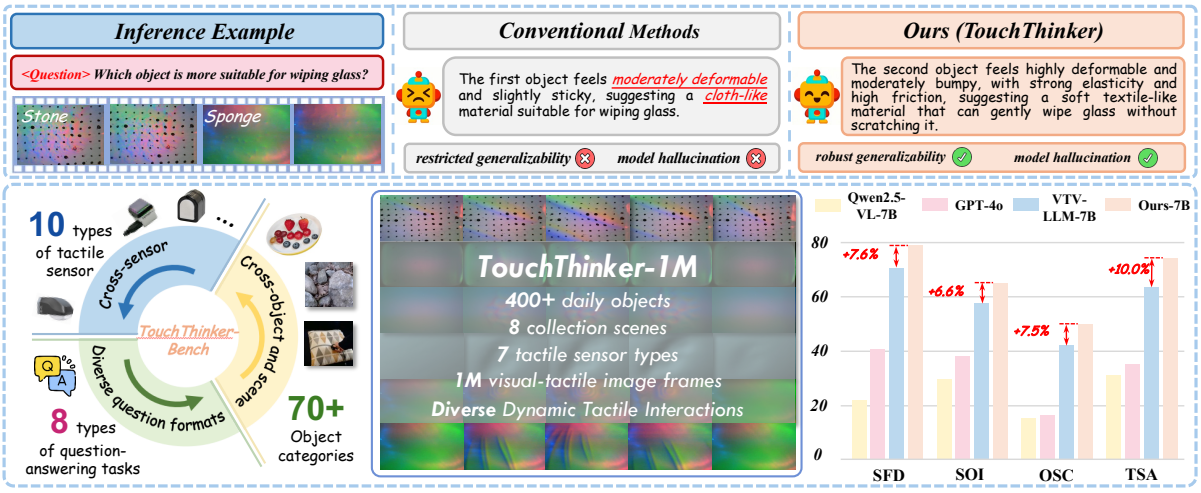


Figure 1: **The core contributions of TouchThinker.** (1) We propose an action-aware tactile encoder to enable efficient tactile representation learning. Upon it, we construct TouchThinker-1M, a million-scale, multi-source visuo-tactile dataset that expands the scope of tactile reasoning. (2) We further introduce TouchThinker-Bench to support multi-dimensional evaluation of open-world tactile reasoning. (3) Across multiple tasks, TouchThinker achieves significant improvements over prior methods.

Abstract

Touch is a key modality for embodied agents to understand the physical world. Although recent work has incorporated tactile signals into language systems for tactile commonsense reasoning, scaling such systems to realistic open-world settings remains challenging due to two key bottlenecks: (1) current tactile reasoning datasets remain limited in format and scale, providing insufficient supervision for reasoning from tactile observations to physical commonsense and hindering the learning of transferable tactile commonsense; (2) tactile signals are inherently redundant and action-specific, yet existing methods often overlook these properties, resulting in inefficient representations with limited semantic expressiveness. To address these limitations, we propose TouchThinker, a tactile-language framework that scales tactile commonsense reasoning to the open world from both data and representation perspectives. First, we construct TouchThinker-1M, a million-scale, multi-source tactile reasoning dataset covering **415** objects, **8** scenarios, and **7** sensor types, providing a solid data foundation for open-world generalization. We further introduce TouchThinker-Bench, an open-world benchmark with more realistic and diverse tasks. Then, we propose an action-aware modeling mechanism to improve tactile representation efficiency and enable efficient reasoning. Experimental results demonstrate that TouchThinker achieves competitive performance against state-of-the-art models across multiple datasets. Our code and dataset will be made available at: https://github.com/lvkailin0118/TouchThinker.

1 Introduction

Among human senses, touch is fundamental for perceiving and interacting with the physical world. It provides physical cues that vision alone cannot reliably capture, including material properties, surface

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 43841–43858
October 24-29, 2026 ©2026 Association for Computational Linguistics

texture, and contact state, thereby supporting physical reasoning and action decisions (Li et al., 2024b; Lyu et al., 2026a; Xiao et al., 2026). For example, when a person touches a hard grain of rice, they can infer that it is under-ripe; similarly, when touching a soft sponge, they can deduce that it is suitable for wiping based on the context. Therefore, incorporating tactile information into commonsense reasoning frameworks and developing tactile reasoning in open-world scenarios are crucial for advancing embodied intelligence.

Several recent studies have begun to integrate tactile perception with large language models (Yu et al., 2024, 2025; Xie et al., 2026; Cheng et al., 2026), enabling robots to perform tactile understanding tasks under natural language instructions. Despite these advances, existing approaches still struggle to achieve reliable and robust tactile reasoning in open-world environments, primarily due to limitations in data and representation. (1) **On the data side**, existing public tactile reasoning datasets remain limited in both format and scale (Yu et al., 2024, 2025; Xie et al., 2026), hindering models' ability to perform tactile reasoning in open-world environments. Specifically, most rely on limited predefined attributes and question-answering templates, lacking causal reasoning supervision from tactile observations to physical properties and further to physical commonsense, which can induce hallucinations (Ishmam et al., 2024). Moreover, these datasets typically cover only one to three sensor types, making it difficult for models to distinguish sensor-specific representational biases from physical property variations shared across sensors, thereby limiting their ability to learn transferable tactile commonsense and cross-sensor physical reasoning. (2) **On the representation side**, existing methods still struggle to effectively model tactile features. First, compared with visual signals, tactile interaction streams typically contain numerous static and transitional frames, whereas task-relevant attributes are localized within only a few highly informative segments (Abad and Ranasinghe, 2020), resulting in substantial redundancy. In addition, tactile signals are inherently action-specific: pressing primarily reveals hardness, sliding captures friction, and rotation exposes texture. However, existing methods typically use uniform sampling or full-frame encoding, treating all frames equally. This makes it difficult to suppress noise from low-information frames and explicitly align action types with question semantics, resulting in inefficient representations with limited semantic expressiveness.

To address these challenges, we introduce TouchThinker, an open-world tactile-language framework that advances both data and representation. On the data side, we construct TouchThinker-1M, which integrates multi-source visuo-tactile data and applies systematic preprocessing, annotation unification, and question-answer expansion to provide semantically consistent supervision for tactile reasoning. On the representation side, we design an action-aware modeling mechanism that first fuses tactile tokens under question guidance and then performs action-aware Gaussian MoE modeling to identify query-relevant action segments and reduce representational redundancy. Based on it, we further develop a two-stage tactile-language training paradigm that aligns tactile evidence with language-based reasoning in the LLM. Experiments across multiple benchmarks, including the VTV-150K (Xie et al., 2026) and our newly constructed TouchThinker-Bench, demonstrate that TouchThinker achieves higher reasoning accuracy than state-of-the-art methods and supports reliable open-world tactile inference.

In summary, our contributions are as follows:

- **Framework**: We propose **TouchThinker**, a tactile-language reasoning framework that aligns tactile cues with task semantics through an action-aware modeling mechanism, enabling efficient tactile representations.
- **Datasets**: We construct **TouchThinker-1M**, a million-scale, multi-source dataset that expands tactile data across multiple dimensions to support open-world tactile reasoning. We further introduce **TouchThinker-Bench**, a systematic benchmark covering diverse tactile sensors and task types for rigorous multi-dimensional evaluation of open-world tactile commonsense reasoning.
- **Practice**: TouchThinker outperforms existing tactile-language models across multiple mainstream datasets and subtasks, while demonstrating more reliable physical reasoning. These results highlight its potential for robotic interaction in open-world environments and provide new insights for future research on tactile intelligence.

2 Related work

Tactile Sensing and Perception. In recent years, tactile sensing has evolved from early capacitive arrays, piezoresistive sensors, and magnetic sensors to advanced vision based systems capable of capturing high resolution contact information (Meribout

et al., 2024; Li et al., 2025). visuo-tactile sensors have attracted growing attention because they can record fine grained spatiotemporal deformation on contact surfaces. Representative systems include DIGIT, GelSight, and Tac3D (Li et al., 2024b; Lyu et al., 2026e). Building on these advances, many studies have used visuo-tactile sensing to infer multidimensional tactile properties, thereby supporting dexterous manipulation tasks such as material classification, grasping, and insertion (Lyu et al., 2026d; Zheng et al., 2026; Zhang et al., 2026). Recent research has shifted toward representation learning for tactile data. UniTouch (Yang et al., 2024) and T3 (Zhao et al., 2024) adopt visual self supervised objectives for fine grained feature extraction, while UniT (Xu et al., 2025) introduces VQGAN (Esser et al., 2021) for compact tactile latent space modeling. To address sensor heterogeneity, several studies employ joint training and alignment strategies to encourage consistent representations across sensors (Xie et al., 2026; Feng et al., 2025). In contrast, we identify representational redundancy in tactile signals and emphasize the action specific nature of tactile information. Based on this observation, we aim to learn more efficient tactile representations and unlock their potential for complex tactile reasoning.

Tactile Commonsense Reasoning. Multimodal large language models jointly model language and visual information, substantially improving cross modal reasoning and reshaping research paradigms (Yin et al., 2024; Li et al., 2024a; Wu et al., 2025). While early work primarily focused on vision language models (Du et al., 2022; Wu et al., 2026; Lyu et al., 2026b; Luo et al., 2026; Lyu et al., 2026c), recent studies have begun to leverage the reasoning and understanding capabilities of large language models to model tactile signals, gradually establishing a touch language modeling paradigm for embodied interaction. Representative studies such as Octopi (Yu et al., 2024), Octopi 1.5 (Yu et al., 2025), and VTV-LLM (Xie et al., 2026) have introduced important methods and benchmark datasets for tactile reasoning. However, taken together, existing studies remain constrained by limited datasets (e.g., sensor, scale, and formats), and directly transferring vision-based methods overlooks the unique properties of tactile perception. Motivated by this gap, we curate and release TouchThinker-1M, the largest tactile commonsense reasoning dataset to date to the best of our knowledge. It substantially surpasses existing benchmarks by scaling the data across multiple dimensions, thereby providing a stronger data foundation. Coupled with our methodological improvements, TouchThinker enhances the practical applicability and generalization of tactile commonsense reasoning in open-world scenarios.

3 Curating TouchThinker-1M Dataset

Multi-Source Tactile Data Collection. To address the limitations of existing tactile reasoning datasets in format and scale, which constrain models' reasoning capabilities, we construct **TouchThinker-1M**, a large-scale, multi-source visuo-tactile dataset for open-world tactile commonsense reasoning. It systematically integrates mainstream visuo-tactile data sources from nine datasets, including representative sources such as TacQuad (Feng et al., 2025), PHYSICLEAR (Yu et al., 2024), and VTV-150K (Xie et al., 2026), comprising approximately **1M frames** in total. The incorporation of heterogeneous sources substantially broadens the distribution of tactile signals and provides a foundation for learning generalizable tactile semantic representations. As shown in Figure 2, TouchThinker-1M significantly improves the scale and diversity of tactile data along five complementary dimensions: sensor type, interaction action, object coverage, sample size, and task formulation. Dataset details are provided in the **Appendix** A.1.1.

Tactile Attribute Annotation and Unification. Since annotation schemas vary across tactile datasets, directly merging them would lead to an inconsistent semantic space and hinder learning of stable tactile concepts. To address this, we therefore adapt the VTV-150K (Xie et al., 2026) schema into a unified four-dimensional tactile attribute space covering Hardness, Protrusion, Elasticity, and Friction. For data with existing annotations, we map the original labels to this unified schema; For unannotated data, five annotators independently assign three-level labels for Hardness, Protrusion, Elasticity, and Friction by jointly considering object appearance, tactile observations, and contact-induced deformation. The annotations are cross-validated, and disagreements are resolved through adjudication. Samples with insufficient tactile evidence are excluded. The inter-annotator agreement scores (Yu et al., 2024) for Hardness, Protrusion, Elasticity, and Friction are 0.907, 0.924, 0.929, and 0.875, respectively, indicating high annotation consistency.

Standardized Processing of Tactile Videos. We

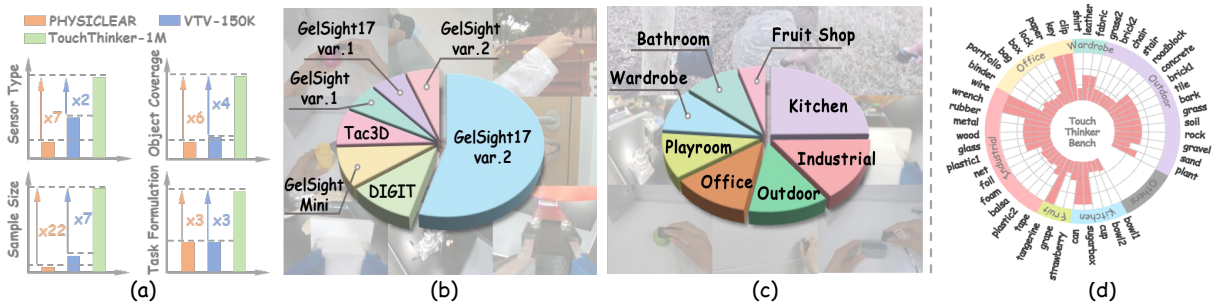


Figure 2: **Data statistics of TouchThinker-1M and TouchThinker-Bench.** (a) Scale comparison between TouchThinker-1M and representative tactile reasoning datasets in terms of sensor type, object coverage, sample size, and task formulation. (b) Distribution of tactile sensor types in TouchThinker-1M. (c) Distribution of scene categories in TouchThinker-1M. (d) Semantic taxonomy and category distribution of TouchThinker-Bench.

standardize all samples into a unified video format to capture tactile interaction dynamics during contact. For existing tactile videos, we retain valid contact intervals and remove noncontact, noisy, or redundant segments. For static tactile images, we construct short video sequences by organizing frames according to object category and contact state to capture deformation evolution. All videos are further processed through contact region cropping, frame rate resampling, interpolation, temporal truncation, and length normalization, yielding clips of approximately 6 to 8 seconds that allow TouchThinker-1M to encode diverse tactile dynamics in a consistent format.

Question Answer Format Expansion. To reduce shallow semantic learning from templated tactile question answering (Yu et al., 2024; Xie et al., 2026), TouchThinker-1M additionally introduces two complementary formats: chain-of-thought reasoning and open-ended question answering. Tactile chain-of-thought instances require an intermediate rationale and a final answer using `<think>...</think><answer>...</answer>`, where `<think></think>` records tactile evidence and reasoning paths, and `<answer></answer>` provides attribute judgments or commonsense conclusions. Open-ended tactile question answering covers free form description, comparative reasoning, attribute explanation, interaction prediction, and open-world decision making, enabling more realistic tactile language reasoning. Construction details are provided in the **Appendix** A.1.2.

4 TouchThinker-Bench

Existing tactile reasoning benchmarks are often constrained by limited object categories, templated question answering formats, and restricted sensor coverage (Yu et al., 2024; Xie et al., 2026), making it difficult to assess whether models acquire transferable tactile semantics. To comprehensively evaluate tactile commonsense reasoning in open-world scenarios, we introduce TouchThinker-Bench. It is designed for more realistic open-world evaluation and incorporates unseen objects, unseen sensors, and diverse tasks, thereby enabling systematic assessment of tactile understanding, commonsense reasoning, and generalization across sensors.

Dataset creation and selection. We first select samples from TouchThinker-1M using an object-level 6:1 train-test split. To evaluate cross-sensor generalization, we further incorporate self-collected data and several additional datasets (Feng et al., 2025; Zhao et al., 2024), covering three sensor types unseen during training. These sensors differ substantially in imaging mechanisms and exhibit intrinsically distinct data characteristics. Then, we apply data processing procedures similar to those described in Section 3 and conduct manual verification, yielding a total of 82 test objects across the unseen-object and unseen-sensor settings, with their distribution shown in Figure 2(d). Further dataset collection processes and dataset details are provided in the **Appendix** A.2.1.

Task Taxonomy. TouchThinker-Bench comprises three core task categories. *(1) Basic tactile property understanding.* This task requires models to identify fundamental tactile attributes, including hardness, roughness, elasticity, and friction. *(2) Basic tactile reasoning.* Following prior benchmarks such as Octopi (Yu et al., 2024) and VTV (Xie et al., 2026), this task covers surface feature distinction (SFD), surface optimality identification (SOI), object sensation correlation (OSC), and tactile scenario analysis (TSA), enabling stable quantitative comparison. *(3) Open-ended tactile commonsense reasoning.* It requires models to generate free-form responses for tactile phenomenon description, at-

tribute explanation, and related reasoning tasks, assessing whether they can integrate tactile evidence with real-world commonsense rather than merely matching fixed templates. Detailed task definitions and examples are provided in the **Appendix** A.2.2.

5 Methodology

5.1 Tactile Action-aware Modeling

As discussed in Section 1, tactile signals contain substantial redundancy, with task-relevant evidence concentrated in only a few action segments. Modeling all frames indiscriminately amplifies representational redundancy and hinders downstream reasoning. To address this issue, we propose a tactile action-aware modeling mechanism that localizes query-relevant subintervals and extracts meaningful tactile representations, as shown in Figure 3.

Question-Guided Token Fusion. Given a tactile video $V = \{I_t\}_{t=1}^{T}$, we first fine-tune a base tactile backbone f_{tac} initialized from ViFi-CLIP (Radford et al., 2021). Specifically, we attach an attribute classifier and optimize a cross-entropy loss to predict tactile attributes, including hardness, protrusion, elasticity, and friction. After training, f_{tac} is frozen and used to extract frame-wise tactile features, yielding a tactile token sequence $F = [f_1, \ldots, f_T] \in \mathbb{R}^{T\times d}$. To incorporate task-specific priors, we encode the question q with a frozen text encoder $f_{\text{txt}}(\cdot)$, obtaining word-level features $Q_w \in \mathbb{R}^{L\times d_t}$ and a sentence-level feature $q_s \in \mathbb{R}^{d_t}$, which are projected into the tactile space as $\tilde{Q}_w = Q_w W_w$ and $\tilde{q}_s = q_s W_s$. We then use tactile tokens as queries and word-level question features as keys and values for cross-modal attention, followed by self-attention:

$$F_{\text{qa}} = \text{SelfAttn}\left(\text{CrossAttn}\left(F, \tilde{Q}_w, \tilde{Q}_w\right)\right), \tag{1}$$

The output $F_{\text{qa}} \in \mathbb{R}^{T\times d}$ provides question-aware tactile representations that align temporal tactile cues with task semantics while suppressing irrelevant noise (e.g., non-contact segments), thereby reducing tactile representational redundancy.

Action-aware Gaussian Temporal MoE. To further focus on reasoning-relevant tactile evidence, we adapt the MoE framework by introducing a set of tactile experts, each specialized in distinct action patterns. Rather than discrete experts (Dai et al., 2024; Lin et al., 2026), we use context-dependent routing weights to softly combine their contributions, capturing the continuous dynamics of tactile interactions. Specifically, given the question-aware tactile tokens $F_{\text{qa}} = [f_{\text{qa},1}, \ldots, f_{\text{qa},T}]$, a question-driven router predicts the expert weights $\pi = \{\pi_k\}_{k=1}^{K}$ from the sentence-level question representation $\tilde{q}_s$. Each expert aggregates local tactile evidence through a Gaussian temporal window:

$$f_{\text{moe}} = \sum_{k=1}^{K} \pi_k \sum_{t=1}^{T} \alpha_{k,t} E_k(f_{\text{qa},t}), \tag{2}$$

$$\alpha_{k,t} = \text{Softmax}_t\left(-\frac{(\hat{\tau}_t - \mu_k)^2}{2\sigma_k^2}\right), \tag{3}$$

Here, $\alpha_{k,t}$ denotes the temporal attention weight assigned by the k-th expert to the t-th token, and $\hat{\tau}_t \in [0, 1]$ denotes the normalized temporal position. The parameters μ_k and σ_k denote the center and width of the temporal window for the k-th expert, respectively. They are generated from the D-dimensional question representation through a linear layer, where the predicted center offset is added to an initial temporal center. This design enables the model to focus on relevant action segments and capture the action-specific nature of tactile signals, thereby supporting downstream tactile reasoning.

5.2 Tactile-Language training paradigm

In the previous section, we introduced a tactile action-aware mechanism for generating more efficient tactile representations. We next propose a two-stage training paradigm that enables the LLM to integrate tactile and linguistic embeddings for multimodal reasoning and response generation. The training pipeline is illustrated in Figure 3.

Stage I: Tactile Token Adaption for LLM. In this stage, we align tactile encoder outputs with the LLM text embedding space through a tactile-language adapter, enabling the LLM to interpret tactile inputs. Trained on TouchThinker-1M attribute question answering data, the adapter projects tactile representations into textual token embeddings while the LLM (Wu et al., 2025) remains frozen, focusing optimization on cross-modal alignment. Given a tactile video V and a text prompt p, the tactile embedding is computed as

$$E_V = \text{Proj}(\text{Enc}(V, p)), \tag{4}$$

where Enc denotes the tactile encoder and Proj denotes the tactile-language adapter. The LLM then generates a response $\hat{Y} = \text{LLM}_\phi(E_V, p)$ based on the tactile embedding and the text prompt. The

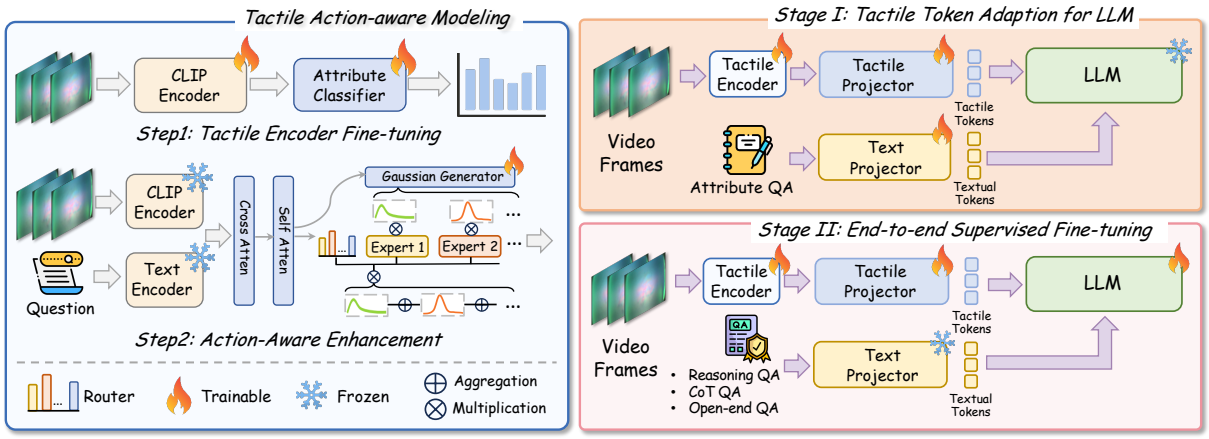


Figure 3: **Overview of the proposed TouchThinker framework and its training pipeline.**

model is trained by minimizing the autoregressive cross entropy loss:

$$\mathcal{L}_{\text{ce}} = -\mathbb{E}_{(V,p,Y)\sim\mathcal{D}_{\text{align}}} \sum_{i=1}^{M} \log \pi_\theta(y_i \mid E_V, p, y_{<i}), \tag{5}$$

where $\mathcal{D}_{\text{align}}$ is the attribute question answering dataset. $Y = [y_1, \ldots, y_M]$ is the target answer sequence, y_i is the target token at position i, $y_{<i}$ denotes the preceding target tokens, and π_θ denotes the predictive distribution. We use teacher forcing during training to stabilize tactile-text alignment.

Stage II: End-to-end Supervised Fine-Tuning. In this stage, we perform LoRA-based supervised fine-tuning on diverse instruction data, including chain-of-thought and open-ended question answering, to elicit evidence-driven and logically consistent tactile reasoning. To preserve stable perceptual representations, we freeze the tactile encoder and update only the tactile-language adapter and the LoRA parameters inserted into the LLM self-attention layers. Each instruction instance is represented as (V, p, Y), where V is a tactile video, p is the text prompt, and Y is the target output. For chain-of-thought instances, Y follows the structured format `<think>...</think><answer>...</answer>`, where the reasoning trace T precedes the final answer A. For open-ended question answering, Y is an open-format natural-language response. Under multi-format supervision, the model aligns dynamic tactile evidence with language-based reasoning. The supervised fine-tuning objective is defined as

$$\mathcal{L}_{\text{sft}} = -\mathbb{E}_{(V,p,Y)\sim\mathcal{D}_{\text{inst}}} \sum_{i=1}^{M} \log \pi_\theta(y_i \mid E_V, p, y_{<i}), \tag{6}$$

where $\mathcal{D}_{\text{inst}}$ denotes the instruction dataset containing attribute question answering, chain-of-thought, and open-ended question answering instances. This stage strengthens alignment between tactile perception and language reasoning, improving generalization in open-world tactile question answering.

6 Experiments

6.1 Settings

Datasets. We conduct experiments on two benchmarks: the recently recognized VTV-150K (Xie et al., 2026) benchmark and our self-constructed TouchThinker-Bench. All test sets contain objects unseen during training, enabling comprehensive evaluation across diverse tactile reasoning tasks. In addition, TouchThinker-Bench further includes unseen sensors to assess cross-sensor generalization.

Implementation. All models are trained on four NVIDIA A100-80G GPUs with a batch size of 16. Tactile-language training follows a two-stage paradigm: Stage I trains the Tactile-Language Adapter on TouchThinker-1M attribute question-answering data using AdamW with a learning rate of 2×10^{-4}; Stage II fine-tunes the adapter and the LLM LoRA parameters on more complex question-answering data, including chain-of-thought and open-ended question answering, with the same optimizer and learning rate. We set the LoRA scaling factor to 256, rank to 128, and maximum training steps to 10,000. The LLM backbone is Qwen2.5, including 7B and 14B (Wu et al., 2025). Additional details are provided in the **Appendix** A.3.1.

Evaluation Metrics. For basic tactile property understanding and reasoning, we report subtask-level accuracy by exactly matching the final conclusions, without considering reasoning processes or seman-

Table 1: **Performance comparison of TouchThinker with state-of-the-art methods on the VTV-150K dataset.** The evaluation covers multiple tasks, with results reported in percentages (%); boldface denotes the best performance. To ensure fair comparison and validate the effectiveness of our method, TouchThinker is trained and evaluated on VTV-150K following the same protocol (Xie et al., 2026), and results are reported as the average over multiple runs.

Models	Hardness	Protrusion	Elasticity	Friction	Combined	SFD	SOI	OSC	TSA	Average
GPT-4o	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	36.2	34.7	39.1	21.0	4.3	42.6	29.4	18.5	40.0	29.5
LLaVA-OneVision-7B	27.5	32.6	26.0	20.2	0.7	40.9	28.2	11.7	30.0	24.2
LLaVA-Video-Qwen2-7B	30.4	29.7	28.9	18.1	2.1	33.6	29.4	17.2	36.0	25.0
InternVL2.5-VL-8B	18.1	23.9	21.0	13.7	0.0	24.5	17.9	11.1	24.0	17.1
VideoLLaMA3-7B	15.2	21.7	14.4	10.8	0.0	11.4	12.8	7.4	20.0	12.6
Qwen2.5-VL-7B	25.3	28.9	17.3	15.9	1.4	22.9	28.2	16.0	30.0	20.6
VTV-LLM-7B	73.9	75.0	67.3	56.5	35.6	71.3	57.6	43.2	64.0	60.4
VTV-LLM-14B	72.1	78.2	68.1	52.8	38.2	72.1	59.7	45.9	72.0	62.1
TouchThinker-7B	79.1	**80.8**	75.3	**63.4**	40.7	78.9	64.2	50.7	74.0	67.5
Δ (vs. VTV-LLM-7B)	+5.2	+5.8	+8.0	+6.9	+5.1	+7.6	+6.6	+7.5	+10.0	+7.0
TouchThinker-14B	**81.2**	77.1	**78.2**	61.5	**46.4**	**80.5**	**68.3**	**52.9**	**80.0**	**69.6**

tic equivalence. For open-ended questions, we measure semantic similarity with METEOR (Banerjee and Lavie, 2005) and use GPT-5 (Singh et al., 2025) and DeepSeek-V4 (DeepSeek-AI, 2026) for multi-dimensional evaluation.

6.2 Main Results

The main experimental results demonstrate that TouchThinker scales tactile commonsense reasoning to open-world scenarios with three key ***adv***antages: ***[Adv.1]*** improved tactile property prediction and basic tactile commonsense reasoning, ***[Adv.2]*** accurate and coherent open-ended response generation, and ***[Adv.3]*** robust generalization to unseen sensors and objects.

[Adv.1] To ensure fair comparison and validate the effectiveness of our method, we follow the VTV-LLM evaluation protocol (Xie et al., 2026) and quantitatively compare TouchThinker with state-of-the-art models on 500 question-answer pairs sampled from VTV-150K. The results are summarized in Table 1. On Tactile Feature Analysis task, TouchThinker-7B outperforms VTV-LLM-7B by 5.2%, 5.8%, 8.0%, and 6.9% in hardness, protrusion, elasticity, and friction prediction, respectively, demonstrating the effectiveness of its action-aware tactile representations. We further evaluate TouchThinker on basic tactile commonsense reasoning tasks, including SFD, SOI, OSC, and TSA. Compared with VTV-LLM-7B, TouchThinker consistently achieves performance gains. Notably, TouchThinker-7B also surpasses VTV-LLM-14B in overall performance and on most subtasks despite using substantially fewer parameters, highlighting its efficient tactile representation learning and stronger reasoning capability.

[Adv.2] Prior methods perform reasonably on tactile property prediction and templated reasoning, but their reliance on fixed answer spaces and exact matching limits the evaluation of open-world tactile explanation and reasoning. Following the implementation details described in Section 6.1, we further evaluate TouchThinker on TouchThinker-Bench for open-ended tactile question answering, using METEOR (Banerjee and Lavie, 2005) and LLM-based evaluation. As shown in Table 2, Large-scale multi-source data and diversified QA training enable TouchThinker to reduce shallow semantic matching and spurious associations, producing responses that are better grounded in tactile evidence and more logically consistent. The qualitative comparison in Figure 4 further shows that TouchThinker better interprets tactile signals, connects physical properties with real-world commonsense, and performs open-world tactile reasoning.

[Adv.3] We evaluate state-of-the-art Touch-Language models on TouchThinker-Bench for tactile property prediction and basic tactile reasoning. Table 3 shows that TouchThinker remains robust to unseen sensors and objects, while Octopi and VTV-LLM degrade substantially. This indicates stronger adaptation to heterogeneous sensor imaging mechanisms, signal distributions, and object properties. This advantage stems from the large-scale, multi-source construction of TouchThinker-1M, which enables models to learn sensor-invariant tactile semantics rather than sensor-specific appearance biases, thereby improving generalization to unseen sensing platforms and open-world objects.

Impact of the Action-aware Modeling. As shown in Table 4, removing the action-aware mechanism leads to consistent performance degradation on

Table 2: **Subtask performance comparison on TouchThinker-Bench.** We evaluate the models' open-ended tactile question-answering capabilities. The best results are in **bold**, and the second-best ones are underlined. TAU: Touch Attribute Understanding. TIU: Touch Interaction Understanding. TKU: Touch Knowledge Reasoning.

Model	TAU			TIU			TKU		
	METEOR	GPT-5	DeepSeek-V4	METEOR	GPT-5	DeepSeek-V4	METEOR	GPT-5	DeepSeek-V4
Octopi-7B (Yu et al., 2024)	26.47	6.45	6.42	25.26	6.89	6.94	20.84	6.23	6.14
Octopi-13B (Yu et al., 2024)	31.43	7.58	7.87	24.41	7.11	**7.79**	26.76	6.39	6.87
VTV-LLM-7B (Xie et al., 2026)	27.93	6.53	6.64	27.45	6.87	6.91	22.17	6.19	6.12
TouchThinker-7B (Ours)	**34.06**	**8.17**	**8.33**	**28.71**	**7.49**	7.21	**27.43**	**7.87**	**7.81**

Table 3: **Performance comparison on TouchThinker-Bench under unseen sensor and object.** Other models suffer substantial performance degradation due to limited generalization, whereas our method maintains robustness.

Models	Hardness	Protrusion	Elasticity	Friction	Combined	SFD	SOI	OSC	TSA	Average
Octopi-7B	43.2	47.1	39.5	32.6	13.9	31.4	33.9	22.4	48.0	34.7
Octopi-13B	44.8	46.5	41.9	36.1	22.1	35.5	38.7	30.6	56.0	39.1
VTV-LLM-7B	59.2	60.1	62.8	46.5	28.4	51.6	43.5	39.7	52.0	49.3
TouchThinker-7B	**68.3**	**69.7**	**70.6**	**51.7**	**37.2**	**68.2**	**52.1**	**47.2**	**62.0**	**58.6**

Table 4: Ablation study of TouchThinker components.

Settings	SFD	SOI	OSC	TSA	Average
w/o Action-aware Modeling	72.9	57.1	46.3	70.0	61.6
w/o stage 1	65.3	58.7	41.2	68.0	58.3
w/o stage 2	61.1	51.3	34.9	66.0	53.3
Ours	**78.9**	**64.2**	**50.7**	**74.0**	**67.0**

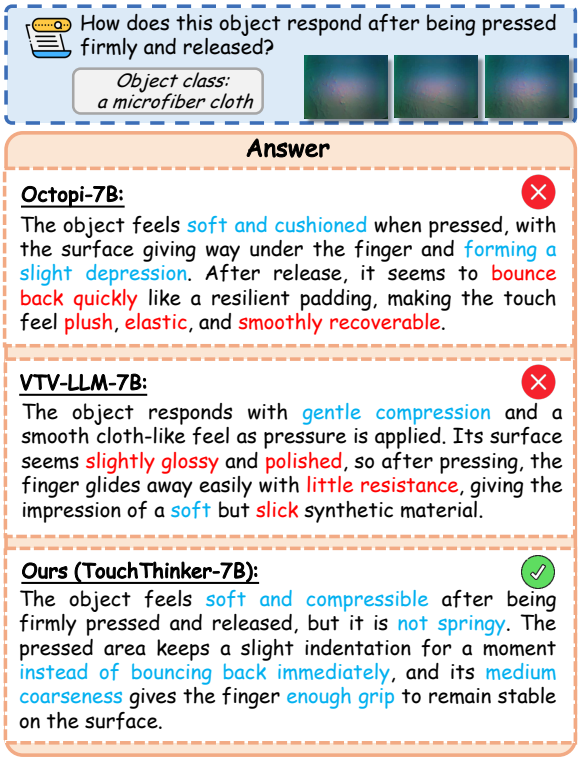


Figure 4: Qualitative comparison of responses from TouchThinker and other models. Errors are highlighted in **red**, while accurate content is marked in **blue**.

both attribute prediction and reasoning tasks, due to the inherent redundancy and action specificity of tactile signals. In contrast, our action-aware mechanism optimizes tactile representations. To

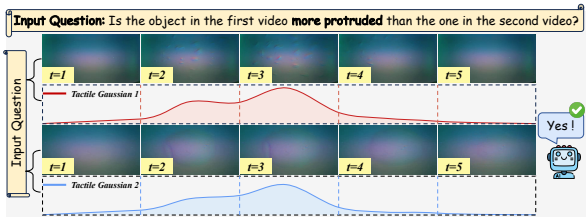


Figure 5: Visualization of action-aware temporal weights from tactile Gaussian experts, integrated to focus on question-relevant frames for accurate reasoning.

intuitively illustrate this effect, we present visualizations in Figure 5. When the question is related to the protrusion attribute, the action-aware mechanism dynamically localizes pressing-related action segments, yielding efficient tactile representations and enhancing tactile understanding.

Impact of the Training Paradigm. Table 4 validates our two-stage training paradigm through ablation studies on VTV-150K. Removing Stage I leads to misalignment between tactile and textual tokens, reducing the average performance to 58.3%. Omitting Stage II removes end-to-end fine-tuning, destabilizes LLM token generation, and causes a sharp performance drop to 53.3%. These results confirm that progressive training is crucial for improving model capability.

7 Conclusion

In this work, we present TouchThinker, an open-world tactile-language reasoning framework that advances tactile reasoning from both data and methodological perspectives. First, we construct TouchThinker-1M, a million-scale multi-source visuo-tactile dataset, and TouchThinker-Bench, a realistic benchmark for evaluating tactile understanding, commonsense reasoning, and cross-

sensor generalization. Then, TouchThinker employs a tactile action-aware modeling mechanism that combines question-guided fusion with action-aware Gaussian Temporal MoE to align tactile cues with task semantics, reduce redundancy, and focus on reasoning-relevant action segments. A two-stage tactile-language training paradigm further aligns tactile evidence with LLM-based reasoning. Extensive experiments show that TouchThinker outperforms existing tactile-language models across benchmarks, with stronger robustness to unseen objects and sensors. These results highlight the value of scalable tactile data and action-aware dynamics for reliable open-world tactile reasoning.

8 Limitations

Although TouchThinker demonstrates promising capability in tactile commonsense reasoning, several limitations remain. First, its supervision is constrained by the diversity of annotated tactile attributes. The current schema covers common and complementary attributes, including hardness, protrusion, elasticity, friction, and their combinations, while real-world tactile perception involves additional properties such as malleability and prickliness. This may limit comprehensive tactile understanding in open-world settings. Second, our current implementation focuses mainly on short-term tactile interactions. Since TouchThinker-1M primarily consists of 6–8-second contact clips, extending the framework to long-horizon tactile manipulation remains important for more complex tasks (Son et al., 2024; Tegin and Wikander, 2005). Third, TouchThinker is evaluated with 7B and 14B LLM backbones, whose computational cost increases with model scale and may hinder deployment on resource-constrained robotic platforms. Developing lightweight and accelerated variants is a promising direction for future work.

9 Broader Impacts and Ethics Statement

This work advances open-world tactile commonsense reasoning by aligning language models with tactile perception, enabling embodied agents to acquire more reliable and interpretable physical understanding. It aligns with broader NLP interests in multimodal grounding, human-AI interaction, trustworthy evaluation, and responsible data curation. Potential benefits include improved assistive technologies for visually impaired individuals and applications in manufacturing, quality control, and robotic manipulation (Liu et al., 2020). We also acknowledge potential ethical risks. (1) Tactile data may encode fine-grained information about object materials, interaction patterns, or usage contexts, raising privacy and sensitive-information concerns if improperly collected, stored, or released. (2) Although TouchThinker-1M integrates multi-source tactile data, its sensor types, object categories, and attribute space cannot fully capture real-world variability, which may lead to generalization biases for low-resource sensors, rare materials, or safety-critical scenarios (Zhang et al., 2024). To mitigate these risks, we emphasize lawful data sources, standardized annotation, cross-sensor evaluation, and open-world generalization analysis, and recommend further real-world validation of safety and feasibility before deployment.

10 Acknowledgments

This work was supported by the Zhongguancun Academy (Grant No. c20250502) and the Guangzhou Key Research and Development Program (Grant No. 2025B01J4002).

References

Alexander C Abad and Anuradha Ranasinghe. 2020. Visuotactile sensors with emphasis on gelsight sensor: A review. *IEEE Sensors Journal*, 20(14):7628–7638.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Ning Cheng, Jinan Xu, Jialing Chen, Bin Fang, and Wenjuan Han. 2026. 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.

Heather Culbertson, Juan José López Delgado, and Katherine J Kuchenbecker. 2014. One hundred data-driven haptic texture models and open-source methods for rendering on 3d objects. In *2014 IEEE haptics symposium (HAPTICS)*, pages 319–325. IEEE.

Damai Dai, Chengqi Deng, Chenggang Zhao, RX Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Yu Wu, and 1 others. 2024. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1280–1297.

DeepSeek-AI. 2026. Deepseek-v4: Towards highly efficient million-token context intelligence.

Yifan Du, Zikang Liu, Junyi Li, and Wayne Xin Zhao. 2022. A survey of vision-language pre-trained models. *arXiv preprint arXiv:2202.10936*.

Patrick Esser, Robin Rombach, and Bjorn Ommer. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883.

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

Ruohan Gao, Yiming Dou, Hao Li, Tanmay Agarwal, Jeannette Bohg, Yunzhu Li, Li Fei-Fei, and Jiajun Wu. 2023. The objectfolder benchmark: Multisensory learning with neural and real objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17276–17286.

Carolina Higuera, Akash Sharma, Chaithanya Krishna Bodduluri, Taosha Fan, Patrick Lancaster, Mrinal Kalakrishnan, Michael Kaess, Byron Boots, Mike Lambeta, Tingfan Wu, and 1 others. 2024. Sparsh: Self-supervised touch representations for vision-based tactile sensing. *arXiv preprint arXiv:2410.24090*.

Md Farhan Ishmam, Md Sakib Hossain Shovon, Muhammad Firoz Mridha, and Nilanjan Dey. 2024. From image to language: A critical analysis of visual question answering (vqa) approaches, challenges, and opportunities. *Information Fusion*, 106:102270.

Haoran Li, Yijiong Lin, Chenghua Lu, Max Yang, Efi Psomopoulou, and Nathan F Lepora. 2025. Classification of vision-based tactile sensors: A review. *IEEE Sensors Journal*.

Ming Li, Keyu Chen, Ziqian Bi, Ming Liu, Xinyuan Song, Zekun Jiang, Tianyang Wang, Benji Peng, Qian Niu, Junyu Liu, and 1 others. 2024a. Surveying the mllm landscape: A meta-review of current surveys. *arXiv preprint arXiv:2409.18991*.

Tong Li, Yuhang Yan, Chengshun Yu, Jing An, Yifan Wang, and Gang Chen. 2024b. A comprehensive review of robot intelligent grasping based on tactile perception. *Robotics and Computer-Integrated Manufacturing*, 90:102792.

Bin Lin, Zhenyu Tang, Yang Ye, Jinfa Huang, Junwu Zhang, Yatian Pang, Peng Jin, Munan Ning, Jiebo Luo, and Li Yuan. 2026. Moe-llava: Mixture of experts for large vision-language models. *IEEE Transactions on Multimedia*.

Huaping Liu, Di Guo, Fuchun Sun, Wuqiang Yang, Steve Furber, and Tengchen Sun. 2020. Embodied tactile perception and learning. *Brain Science Advances*, 6(2):132–158.

Kaiwen Luo, Zhenhong Zhou, Leyan Wang, Liang Lin, Tianyu Shao, Yuanhe Zhang, Yang Xiao, Yuxuan Li, Miao Yu, Kailin Lyu, and 1 others. 2026. A survey of large audio language models: Generalization, trustworthiness, and outlook. *arXiv preprint arXiv:2605.20266*.

Kailin Lyu, Di Wu, Long Xiao, Jianning Zeng, Jianwei He, Chang Lin, Lianyu Hu, Lin Shu, Jie Hao, and Ce Hao. 2026a. Tacreasoner: A dynamic tactile-language framework for interactive reasoning in real-world scenarios. *arXiv preprint arXiv:2607.05131*.

Kailin Lyu, Kangyi Wu, Pengna Li, Xiuyu Hu, Qingyi Si, Cui Miao, Ning Yang, Zihang Wang, Long Xiao, Lianyu Hu, and 1 others. 2026b. Himemvln: Enhancing reliability of open-source zero-shot vision-and-language navigation with hierarchical memory system. *arXiv preprint arXiv:2603.14807*.

Kailin Lyu, Kangyi Wu, Pengna Li, Wenxuan Song, Di Wu, Jianwei He, Junting Chen, Ning Yang, Zebin Han, Kaiwen Luo, and 1 others. 2026c. From instruction following to cognitive navigation: A survey on the evolution of vision-and-language navigation. Technical report, CSPaper OpenPrint.

Kailin Lyu, Long Xiao, Jianing Zeng, Junhao Dong, Xuexin Liu, Zhuojun Zou, Haoyue Yang, Lin Shu, and Jie Hao. 2026d. Touchformer: A robust transformer-based framework for multimodal material perception. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 18496–18504.

Kailin Lyu, Long Xiao, Jianing Zeng, Di Wu, Lin Shu, and Jie Hao. 2026e. Vq-touch: A data-efficient tactile generation framework across sensors and scenarios. *arXiv preprint arXiv:2607.14728*.

Mahmoud Meribout, Natnael Abule Takele, Olyad Derege, Nidal Rifiki, Mohamed El Khalil, Varun Tiwari, and Jing Zhong. 2024. Tactile sensors: A review. *Measurement*, 238:115332.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, and 1 others. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.

Changil Son, Jinyoung Kim, Dongwon Kang, Seojoung Park, Chaeyeong Ryu, Dahye Baek, Geonyoung Jeong, Sanggyun Jeong, Seonghyeon Ahn, Chanoong Lim, and 1 others. 2024. Behavioral biometric optical tactile sensor for instantaneous decoupling of dynamic touch signals in real time. *Nature Communications*, 15(1):8003.

Sudharshan Suresh, Haozhi Qi, Tingfan Wu, Taosha Fan, Luis Pineda, Mike Lambeta, Jitendra Malik, Mrinal Kalakrishnan, Roberto Calandra, Michael Kaess, and 1 others. 2024. Neuralfeels with neural fields: Visuo-tactile perception for in-hand manipulation. *Science Robotics*, 9(96):eadl0628.

Sudharshan Suresh, Zilin Si, Stuart Anderson, Michael Kaess, and Mustafa Mukadam. 2023. Midastouch: Monte-carlo inference over distributions across sliding touch. In *Conference on Robot Learning*, pages 319–331. PMLR.

Johan Tegin and Jan Wikander. 2005. Tactile sensing in intelligent robotic manipulation–a review. *Industrial Robot: An International Journal*, 32(1):64–70.

Universal Robots. 2016. UR5 Technical Specifications. https://www.universal-robots.com/media/50588/ur5_en.pdf. Technical specifications, Item no. 110105, EN 09/2016.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, and 1 others. 2025. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*.

Kangyi Wu, Pengna Li, Kailin Lyu, Xi Lin, Lin Zhao, Qingrong He, Jinjun Wang, and Jianyi Liu. 2026. Dual-anchoring: Addressing state drift in vision-language navigation. *arXiv preprint arXiv:2604.17473*.

Long Xiao, Kailin Lyu, Jianwei He, Jianing Zeng, Lin Shu, and Jie Hao. 2026. Tacexpert: A pseudo-temporal mixture-of-experts framework for open-set tactile object recognition. In *2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3761–3765. IEEE.

Yifan Xie, Mingyang Li, Shoujie Li, Xingting Li, Guangyu Chen, Fei Ma, Fei Yu, and Wenbo Ding. 2026. Universal visuo-tactile video understanding for embodied interaction. *Advances in Neural Information Processing Systems*, 38:127864–127883.

Zhengtong Xu, Raghava Uppuluri, Xinwei Zhang, Cael Fitch, Philip Glen Crandall, Wan Shou, Dongyi Wang, and Yu She. 2025. Unit: Data efficient tactile representation with generalization to unseen objects. *IEEE Robotics and Automation Letters*.

Fengyu Yang, Chao Feng, Ziyang Chen, Hyoungseob Park, Daniel Wang, Yiming Dou, Ziyao Zeng, Xien Chen, Rit Gangopadhyay, Andrew Owens, and 1 others. 2024. Binding touch to everything: Learning unified multimodal tactile representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26340–26353.

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

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403.

Samson Yu, Kelvin Lin, and Harold Soh. 2025. Demonstrating the octopi-1.5 visual-tactile-language model. *arXiv preprint arXiv:2507.09985*.

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

Pengwei Zhang, Bin Xie, Xinpan Meng, Xinyu Guo, Ce Hao, Fang Deng, Long Cheng, and Tiancai Wang. 2026. Feeling the unexpected: Restacvla for contact-rich manipulation via residual tactile representation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*.

Yichi Zhang, Yao Huang, Yitong Sun, Chang Liu, Zhe Zhao, Zhengwei Fang, Yifan Wang, Huanran Chen, Xiao Yang, Xingxing Wei, and 1 others. 2024. Multitrust: A comprehensive benchmark towards trustworthy multimodal large language models. *Advances in Neural Information Processing Systems*, 37:49279–49383.

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

Yuhang Zheng, Songen Gu, Weize Li, Yupeng Zheng, Yujie Zang, Shuai Tian, Xiang Li, Ce Hao, Chen Gao, Si Liu, and 1 others. 2026. Omnivta: Visuo-tactile world modeling for contact-rich robotic manipulation. *arXiv preprint arXiv:2603.19201*.

A Appendix

A.1 Details of TouchThinker-1M Dataset

A.1.1 Source Data of TouchThinker-1M

TouchThinker-1M integrates nine tactile datasets across seven tactile sensing platforms, four major acquisition actions, and 415 deduplicated objects, comprising 1,001,344 tactile frames in total. Detailed object categories are provided in Figure 11. We next provide an overview of the source datasets that constitute TouchThinker-1M.

- **VTV150K** (Xie et al., 2026). It contributes 150,000 visuo-tactile video frames from 100 common objects, manually captured with GelSight Mini, DIGIT, and Tac3D sensors through normal pressing, rotational movement, and sliding motion.
- **PhysiCLEAR** (Yu et al., 2024). It retains 48 everyday objects after duplicate removal and uses a GelSight var.1 tactile sensor for handheld

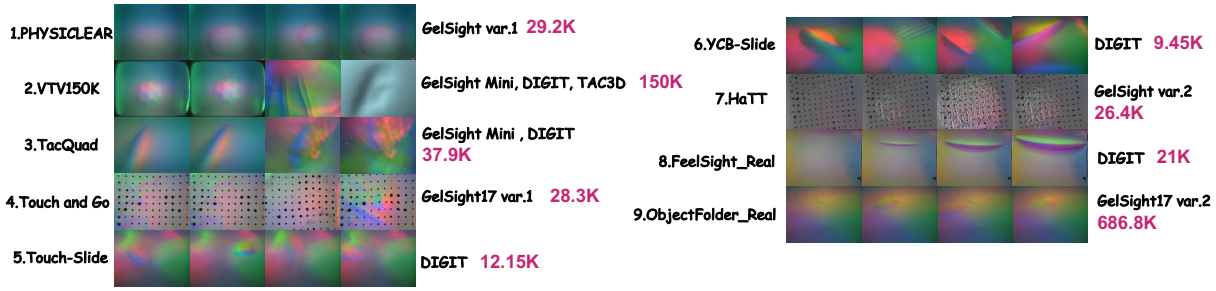


Figure 6: Overview and representative examples of the source datasets in TouchThinker-1M.

pressing and rotation, comprising 29,211 tactile frames after curation.

- **Touch and Go** (Yang et al., 2022). It spans 18 material categories, where human collectors probe real-world objects with a GelSight17 var.1 tactile sensor while recording egocentric video, comprising 28,300 tactile frames after curation.
- **TacQuad** (Feng et al., 2025). It uses the coarse-grained spatially aligned subset covering 92 objects, collected with GelSight Mini and DIGIT image-based tactile sensors through sequential handheld pressing at the same contact location with additional twisting motions, comprising 37,955 tactile frames after curation.
- **Touch-Slide** (Higuera et al., 2024). It covers 9 toy-kitchen objects collected as human sliding interaction trajectories with a DIGIT tactile sensor, comprising 12,150 tactile frames after curation.
- **YCB-Slide** (Suresh et al., 2023). YCB-Slide covers 7 YCB objects collected by sliding a handheld DIGIT tactile sensor across fixed object surfaces and recording trajectories, comprising 9,450 tactile frames after curation.
- **FeelSight-Real** (Suresh et al., 2024). It covers 5 real-world objects collected with four DIGIT vision-based tactile sensors mounted on the fingertips of an Allegro hand, using an in-hand rotation policy to record vision, touch, and proprioception, comprising 21,000 tactile frames after curation.
- **ObjectFolder-Real** (Gao et al., 2023). It covers 70 real-world household object instances, collected with GelSight17 var.2 tactile sensors mounted on a Franka Emika Panda arm by contacting selected surface points along their normal directions and recording gel deformation, comprising 686,880 tactile frames after curation.
- **HaTT** (Culbertson et al., 2014). It covers 66 material textures using GelSight var.2 single-press videos augmented from the original HaTT dataset as tactile data, comprising 26,398 tactile frames after curation.

A.1.2 Tactilc Instruction Data Synthesis

Template-based instructions. Following Octopi (Yu et al., 2024) and VTV-LLM (Xie et al., 2026), we construct template-based tactile instructions by instantiating predefined templates with attribute labels and tactile descriptions. The generated data cover basic tactile property understanding and question-answering tasks, including tactile feature analysis (TFA), surface feature distinction (SFD), surface optimality identification (SOI), object sensation correlation (OSC), and tactile scenario analysis (TSA), enabling models to learn tactile-semantic correspondences across surface properties, object sensations, and interaction contexts. However, this template-driven paradigm is constrained by template coverage and expressiveness; its fixed wording and label-based supervision may encourage shallow correlations rather than grounded tactile reasoning.

Tactile Chain-of-Thought Instructions. Prior work (Yu et al., 2024, 2025; Xie et al., 2026) mainly relies on attribute classification or template-based question answering, yielding discrete labels or brief responses without explicitly modeling contact dynamics or causal dependencies. This limitation induces shallow tactile-semantic associations. To address it, we construct a high-quality tactile chain-of-thought reasoning dataset grounded in intrinsic object-state changes during dynamic tactile interactions. The construction pipeline is detailed below. **Stage 1**: Data Preparation. We first curate tactile videos collected under a unified acquisition protocol, retaining key contact and motion segments that reveal object-property variations. We then standardize task metadata, including question categories, task instructions, and answer annotations, to provide a consistent training format. **Stage 2**: Chain-of-Thought Generation. We reformulate tactile understanding as cross-modal question answering and attribute-description generation. Using prede-

fined prompt templates, we guide a large language model (DeepSeek-AI, 2026) to generate reasoning traces grounded in deformation dynamics, contact-region changes, and geometric or texture cues observed in tactile videos. Each output is organized as ...<think>...</think><answer>...</answer>, where the reasoning is aligned with key interaction phases and the answer summarizes the inferred physical property. **Stage 3**: Manual Filtering. We conduct consistency checking and manual review to ensure data quality. Samples are discarded if their reasoning fails to capture key interaction stages or if the intermediate reasoning conflicts with the final answer, thereby preserving logical consistency between chain-of-thought supervision and attribute prediction. **Stage 4**: Formatting and Data Integration. Finally, we integrate tactile videos, task annotations, structured prompts, chain-of-thought texts, and final answers into a unified training-ready format. This standardized pipeline enables supervision for dynamic tactile reasoning and improves tactile understanding and decision reliability in open-world scenarios.

Open-ended Tactile QA Instructions. To enhance semantic understanding and interaction-oriented reasoning in open-ended tactile scenarios, we construct tactile instruction data in an open-ended question-answering format, enabling models to learn aligned mappings from tactile context to language responses. Inspired by Cheng et al. (Cheng et al., 2026), we design an automated generation pipeline based on the raw tactile data in TouchThinker-1M. This pipeline builds on object-level tactile data, object identifiers, and manually annotated tactile attribute descriptions, and uses DeepSeek-V4 (DeepSeek-AI, 2026) to generate instruction-following samples for tactile interaction. Specifically, we generate open-ended tactile question-answering data from two complementary perspectives: intrinsic tactile attributes, including hardness, protrusion, elasticity, and friction, and interaction-oriented perceptual experiences, including graspability, pressure feedback, surface discriminability, and manipulation stability. This design covers both low-level tactile properties and realistic touch-based reasoning scenarios. We first construct an exemplar pool, where each exemplar contains a tactile context and a grounded dialogue, as illustrated in Figure 7. The tactile context consists of detailed and attribute-level descriptions, and only question-answer pairs explicitly supported by the given tactile evidence are retained. For each target sample, we use its tactile context as the query and retrieve semantically relevant exemplars as few-shot demonstrations. Because TouchThinker-1M organizes tactile samples, object identifiers, and attribute descriptions at the object level, the retrieved demonstrations help align generated responses with the corresponding tactile evidence. The complete prompt construction is shown in Figure 8. We then convert model outputs into single-turn conversations and manually review and correct them, yielding 5,000 touch-language instruction-following samples.

A.2 Details of TouchThinker-Bench

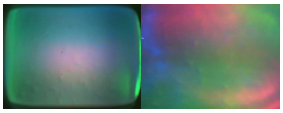
A.2.1 Source Data of TouchThinker-Bench

TouchThinker-Bench consists of two components. The first is derived from the predefined test splits of TouchThinker-1M and contains objects held out from training, enabling evaluation of model generalization to novel objects. The second targets cross-sensor transfer, comprising DuraGel data from the original TacQuad (Feng et al., 2025) test split, whose sensor is absent during training, as well as our self-constructed dataset and VisGel (Zhao et al., 2024), collected with the unseen commercial GelSight Mini and GelSight Var. 3 sensors, respectively. Overall, TouchThinker-Bench covers 10 tactile sensors and 82 test objects, spanning both unseen-object and unseen-sensor settings to provide a comprehensive evaluation of tactile perception, attribute reasoning, and cross-sensor generalization. We provide a detailed description of each component below.

Cross-object generalization evaluation. It consists of multiple existing tactile datasets from the test portion of TouchThinker-1M, as shown in Figure 9(b). The PhysiCLEAR (Yu et al., 2024) portion is collected with a GelSight var.1 tactile sensor and contains 5 daily objects in the test set; the VTV-150K (Xie et al., 2026) portion is collected with three tactile sensors, including GelSight Mini, DIGIT, and Tac3D, and contains 5 common objects; the Touch and Go (Yang et al., 2022) portion is collected with a GelSight17 var.1 tactile sensor and contains 18 material surfaces; the YCB-Slide (Suresh et al., 2023) portion is collected with a DIGIT tactile sensor and contains 3 YCB objects; the Touch-Slide (Higuera et al., 2024) portion is collected with a DIGIT tactile sensor and contains 3 toy-kitchen objects; the HaTT (Culbertson et al., 2014) portion is collected with a Gel-

Tactile Context

Detailed description: The touched object is a sponge. When pressed, its porous body collapses noticeably and then springs back with clear force. The open-cell surface feels mildly uneven, and the soft, grippy material produces strong drag during lateral contact.
Attribute description: highly deformable; moderate protrusion; strong elasticity; strong friction



Conversation

Question: What surface texture is perceived during contact?
Answer: The surface feels porous and mildly uneven rather than smooth. During lateral contact, it produces noticeable drag, and the open-cell structure gives a clear sense of small texture variations across the surface.
Question: How does the object respond under applied pressure?
Answer: The object deforms readily under pressure. Even a light press causes clear compression, indicating high deformability and a soft tactile response.
Question: How stable is the grip during contact?
Answer: The grip is relatively stable. Strong friction and the material's ability to conform to the fingers help maintain contact and reduce slipping during handling.
Question: What rebound response occurs after compression?
Answer: A clear rebound is perceived after compression. The material quickly recovers its shape, producing strong elastic feedback after pressure is released.
Question: Which tactile cue is most prominent during interaction?
Answer: The most prominent tactile cue is the combination of strong compressibility, elastic recovery, and high surface drag. Together, these cues form a distinctive porous and resilient tactile profile.

Figure 7: Example of an open-ended tactile instruction in TouchThinker-1M, using multi-granularity tactile context for grounded reasoning about physical properties and touch-based interactions.

Sight var.2 tactile sensor and contains 9 material textures; the FeelSight-Real (Suresh et al., 2024) subset uses four fingertip-mounted DIGIT sensors on an Allegro hand and contains 1 real object, while ObjectFolder-Real (Gao et al., 2023) uses a GelSight17 var.2 sensor and contains 9 real household objects.

Cross-sensor generalization evaluation. It consists of data collected with three tactile sensors unseen during training, as shown in Figure 9(a). The TacQuad subset is collected using the DuraGel tactile sensor and contains 15 test objects (Feng et al., 2025), while the VisGel subset is collected using the unseen GelSight Var. 3 tactile sensor (Zhao et al., 2024) and includes 4 objects. To further assess performance in real-world deployment scenarios, we additionally collected tactile data for 10 categories of everyday objects using a UR5 robotic arm (Universal Robots, 2016) equipped with the commercial GelSight Mini sensor.

A.2.2 Task Types

Basic Tactile Property Understanding. This task evaluates whether models can identify fundamental tactile attributes, including hardness, roughness, elasticity, and friction. The evaluation covers both unseen sensors and unseen objects, thereby assessing the generalization ability of tactile representations across sensor domains and object categories.

Basic Tactile Reasoning. Following prior benchmarks such as Octopi (Yu et al., 2024) and VTV-LLM (Xie et al., 2026), this task includes surface feature distinction (SFD), surface optimality identification (SOI), object sensation correlation (OSC), and tactile scenario analysis (TSA). These tasks enable standardized quantitative comparison of models' tactile reasoning capabilities.

Open-ended tactile commonsense reasoning. We construct open-ended tactile question-answering samples from TouchThinker-Bench using tactile images and their category labels. Specifically, we design a unified prompt template for three task types and use category labels and tactile attribute cues as conditional inputs to DeepSeek-V4 (DeepSeek-AI, 2026) for generating candidate question-answer pairs, followed by manual verification. The benchmark covers Touch Attribute Understanding, Touch Interaction Understanding, and Touch Knowledge Reasoning. These tasks respectively evaluate models' ability to recognize basic tactile properties, such as hardness, protrusion, elasticity, and friction; reason about interaction-related characteristics, such as graspability, pressure feedback, surface discriminability, and manipulation stability; and infer higher-level tactile knowledge, including material properties, usage patterns, and everyday interaction scenarios, by integrating tactile cues with object categories. As shown in Figure 10, each question is queried independently to avoid unstable formatting and answer-

System Prompt

```
messages = [{"role": "system", "content": f"""You are an AI tactile expert interacting with a single touched object. The
      input provides the tactile context of this object, which contains two complementary descriptions of the same object: a
       detailed description and an attribute description. When answering all questions, treat the tactile context as the
      information perceived during contact.

Design a conversation between you and a person asking about the touched object. The questions should be diverse, and the
      answers must be grounded only in the given tactile context.

The conversation should include questions about the object's physical tactile properties, such as hardness, protrusion,
      elasticity, friction, and related aspects. Only ask questions that can be clearly answered from the tactile context or
       answered reliably.

Also include complex questions related to the touched object, such as questions involving tactile commonsense or interaction
       and perception during touch. These questions may discuss graspability, pressure feedback, surface discriminability,
      manipulation stability, and related aspects. Do not ask about uncertain details. Complex answers should be more
      detailed, clearly structured, and supported by examples or reasoning when necessary."""}]
```

Figure 8: Prompt template for generating open-ended tactile instruction data in TouchThinker-1M. The system prompt guides the model to synthesize evidence-grounded, touch-centric dialogues from the supplied tactile context, covering low-level tactile attributes and higher-level interaction-aware perceptual reasoning.

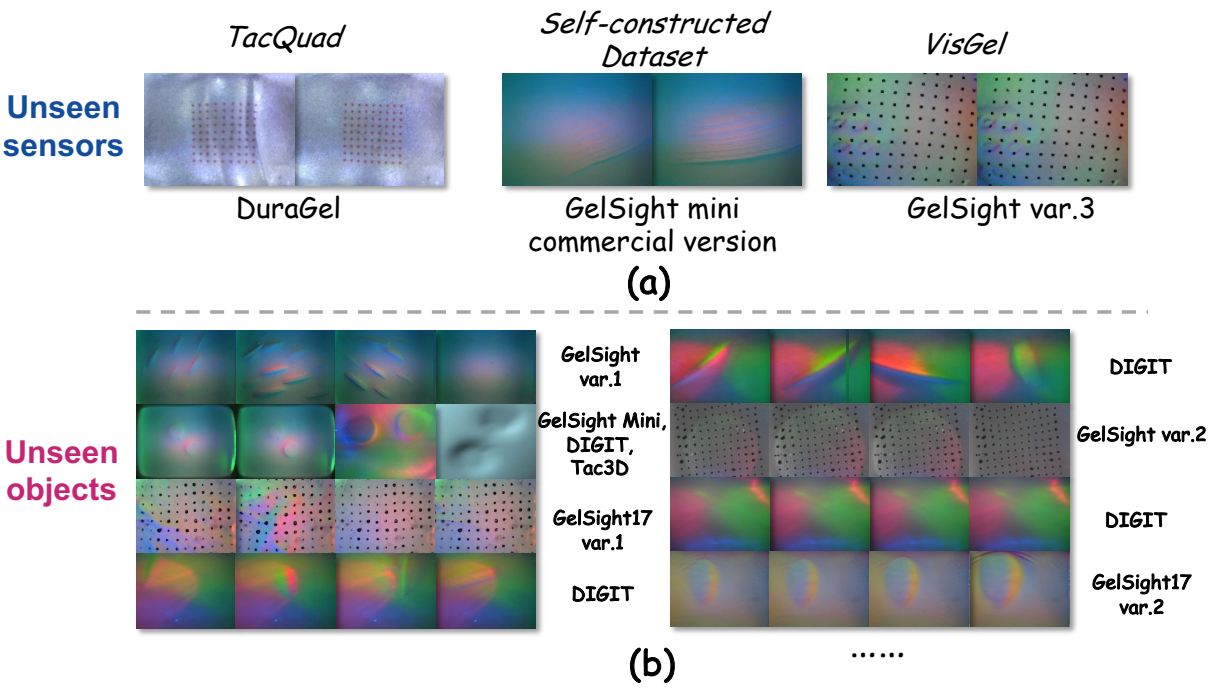


Figure 9: **Overview and representative examples of TouchThinker-Bench.** (a) The unseen-sensor split consists of tactile sensors absent from the training set, including DuraGel from TacQuad (Feng et al., 2025), the commercial GelSight Mini from our self-constructed dataset, and GelSight Var. 3 from VisGel (Zhao et al., 2024). These sensors exhibit varying degrees of domain shift from the training sensors in imaging mechanism, color distribution, and gel deformation characteristics, enabling evaluation of cross-sensor generalization. (b) The unseen-object split consists of objects withheld from training, evaluating whether models can generalize to open-world objects with unseen tactile appearances and physical properties.

segmentation errors when generating multiple long-form responses. We then parse, deduplicate, and format the model outputs into single-turn tactile question-answering samples. Finally, we conduct manual review and correction to ensure that each question is supported by tactile and category information and that the reference answers are consistent with the corresponding objects' tactile attributes and interaction characteristics.

A.3 Details of Implementation

A.3.1 Configuration and Hyperparameters

We present the detailed training configuration and hyperparameters in the two stages in Table 5.

Open-ended Tactile QA Prompt

Text Context
You are an AI tactile assistant interacting with a single touched object. You will understand the touched object from the text prompt, where the object class is `<object class>`. Imagine that you are physically touching this object.

Design conversations between you and a person asking about this object. The answers should be written in a tone indicating that you are touching the object and answering based on tactile perception.

All answers must be based on the confirmed tactile attribute tendencies and the object class, and should not introduce details that cannot be supported by the given information.

Object Class
a microfiber cloth

Task Specific Prompt Type 1: Touch Attribute Understanding
Include questions asking about the typical tactile physical properties of the touched object, including hardness, protrusion, elasticity, and friction. Design a separate conversation for each tactile attribute. Ensure that the conversations contain only confirmed information.

Required format:

```
[
{"question": "<question 1>", "answer": "<answer>"},
{"question": "<question 2>", "answer": "<answer>"},
{"question": "<question 3>", "answer": "<answer>"},
{"question": "<question 4>", "answer": "<answer>"},
{"question": "<question 5>", "answer": "<answer>"}
]
```

Task Specific Prompt Type 2: Touch Interaction Understanding
Include questions asking about interaction and perception characteristics during touch, including graspability, pressure feedback, surface discriminability, and manipulation stability. Randomly select three of these interaction and perception characteristics, and design a separate conversation for each. Each conversation should include one question and 3–5 corresponding ground-truth answers. Ensure that the conversations contain only confirmed information.

Required format:

```
[
{"question": "<question 1>", "answer": "<answer>"},
{"question": "<question 2>", "answer": "<answer>"},
{"question": "<question 3>", "answer": "<answer>"}
]
```

Task Specific Prompt Type 3: Touch Knowledge Reasoning
Include complex reasoning questions involving background knowledge related to the touched object, emphasizing the tactile perspective. Design two conversations. Each conversation should include one question and 3–5 candidate ground-truth answers, with each answer being independent and self-contained. Ensure that the conversations contain only confirmed information. Provide more detailed answers for complex questions. For example, include concrete examples or reasoning steps to make the content more convincing and well-organized. Multiple paragraphs may be used if necessary.

Required format:

```
[
{"question": "<question 1>", "answer": "<answer>"},
{"question": "<question 2>", "answer": "<answer>"}
]
```

Figure 10: Prompt template for open-ended question answering in TouchThinker-Bench. The upper section presents the multimodal input to DeepSeek-V4 (DeepSeek-AI, 2026), including a textual prompt and its paired visual input.

Configuration	Stage I	Stage II
Optimizer	AdamW	AdamW
Learning Rate	2e-4	2e-4
Weight Decay	0.001	0.001
Training Epochs	1	1
Warmup Ratio	0.1	0.1
Learning Rate Scheduler	Linear	Linear
Batch Size Per GPU	16	16
Maximum Token Length	512	512
Unfreeze LLM	✗	✓

Table 5: Training configurations for the two-stage training paradigm. Stage I: Tactile-Text Alignment. Stage II: End-to-end Supervised Fine-Tuning.

A.3.2 GPT-5 and DeepSeek-V4 scoring

Since TouchThinker-Bench is designed for open-ended and free-form tactile commonsense reasoning, model responses may vary substantially in wording, level of detail, and reasoning style. Therefore, conventional n-gram-based metrics such as CIDEr and BLEU-4 are insufficient to fully capture the quality of generated answers. To provide a more comprehensive evaluation, we report METEOR as a semantic-similarity-oriented metric and further employ GPT-5 (Singh et al., 2025) and DeepSeek-V4 (DeepSeek-AI, 2026) as automatic judges for multidimensional response assessment.

Specifically, each generated response is evaluated from five perspectives: semantic correctness,

tactile consistency, commonsense and reasoning plausibility, information completeness, and language quality. For each dimension, the judge assigns a score from 1 to 10 and provides a brief rationale. The final score is computed as the average of the five dimension-level scores:

$$S = \frac{1}{5}\sum_{i=1}^{5} s_i, \tag{7}$$

where s_i denotes the score of the i-th evaluation dimension. The evaluation dimensions are defined as follows:

- **Semantic Correctness:** Whether the response is semantically consistent with the reference answer and free from factual errors.
- **Tactile Consistency:** Whether the response accurately reflects tactile perception, physical properties, or interaction-related sensations, and whether it is consistent with the provided tactile information.
- **Commonsense and Reasoning Plausibility:** Whether the response accords with real-world physical commonsense and presents a reasonable and convincing reasoning process.
- **Information Completeness:** Whether the response covers the key information required by the question and avoids omitting important tactile attributes or explanations.
- **Language Quality:** Whether the response is clear, natural, and fluent, without obvious grammatical or expression errors.

A.4 Additional Ablation Studies

Impact of the model scale. To investigate the impact of model scale on visuo-tactile understanding, we evaluate LLM backbones of varying sizes and report the performance of TouchThinker with Qwen-2.5-14B, as shown in Table 1. The results demonstrate consistent improvements across most subtasks as model size increases, with especially significant improvements on reasoning-oriented tasks such as SFD.

Impact of Question-Guided Token Fusion and Gaussian Temporal MoE in the Action-Aware Modeling Mechanism. We conduct ablation experiments on VTV-150K to validate the effectiveness of the two core components in our action-aware modeling mechanism. As shown in Table 6, both components contribute to tactile reasoning performance. Removing Question-Guided Token Fusion prevents the model from incorporating task semantics during tactile feature aggregation, leading to less focused temporal representations and degraded reasoning accuracy. Removing Gaussian Temporal MoE also leads to performance degradation, indicating that uniform temporal modeling is insufficient for capturing action-specific tactile evidence. In contrast, combining Question-Guided Token Fusion with Gaussian Temporal MoE achieves the best overall performance. This demonstrates that early question-aware fusion helps suppress irrelevant tactile segments, while Gaussian Temporal MoE further localizes task-relevant action intervals, enabling more efficient tactile representations and stronger downstream reasoning.

Question-Guided Token Fusion	Gaussian Temporal MoE	SFD	SOI	OSC	TSA	Avg.
✓	✓	**78.9**	**64.2**	**50.7**	**74.0**	**67.0**
	✓	73.4	51.6	42.2	70.0	59.1
✓		71.2	52.9	45.8	72.0	60.5

Table 6: Ablation study of Question-Guided Token Fusion and Gaussian Temporal MoE in the action-aware modeling mechanism.

Object Categories in TouchThinker-1M

tangerine, carton, playing card, orange, spoon with rice, nectarine, tomato, rubber glove, pen pad, cotton cloth, velvet, sandpaper, chip bag, rubber slipper sole, candle, suede, silk scarf, fascia ball, chalk, clay, avocado, lemon, banana, kiwi, pineapple, plastic bottle, waffle, sponge, plastic basket, balloon, leather glove, building block protrusion, jelly, piano key, blanket, ceramic cup, oven glove, bark, scouring pad, fur, pine cone, ping pong ball, plastic building block, cork, wooden ruler, eraser, velcro, leather wallet, toilet paper roll, shower mat, baseball, golf ball, sticky note, silicone pad, yoga mat, masking tape roll, rubber band, cotton ball, gauze, computer mouse, headphone, face towel, woven watch strap, rubber watch strap, metal watch strap, wooden block, marble, claw, keyboard, remote control button, toothbrush head, vitamin tablet, tennis ball, towel, absorbent cloth, wrist guard, fine bubble film, coarse bubble film, rice, ridge cup, key, screw, circuit board, mold, sponge sheet, wrench, screwdriver handle, aluminum tube, hairpin, steel wool, huamei, candy, grid bag, ridged plastic bottle, scissors, iron ruler, iron clip, tape, injection tube, wire, scissors handle, dishwashing cloth, mesh strainer scoop, toothbrush handle, scissors blade, TV remote back, pillow, disposable water bottle, game controller thumbstick, tissue paper, mop head, rice spatula handle, game controller buttons, hairbrush bristles, leather book cover, bread knife handle, card holder, stress ball, feather duster head, toilet brush handle, potato, paper towel, game controller shoulder buttons, lanyard, insulating holder, basket, rubber slippers, rice spatula scoop, aluminium foil, pen barrel, mesh strainer handle, egg, game controller keypad, bread knife blade, game controller body, bubble wrap, ice block, clothes peg, TSA lock numbers, millet, nylon shirt, feather duster handle, jeans, microfiber cloth, hairbrush handle, toilet brush bristles, hairbrush bristle ends, bath towel, concrete surface, plastic surface, glass surface, wooden surface, metal surface, brick surface, tile surface, leather surface, synthetic fabric surface, rubber surface, paper surface, tree bark surface, grass surface, soil surface, rock surface, gravel surface, sand surface, plant surface, 3D-printed object, USB connector, vacuum cup, vacuum cup lid, glass jar, bowl edge, bowl handle, metal can, metal can bottom, metal can top, CD, coffee cup bottom edge, cup shell edge, cup shell interior, glass bottle, hot dog, red metal box, metal box edge, plush toy, metal box, mouse scroll wheel, onion, paper cup, peach, pencil box, pill bottle bottom edge, pill bottle cap, plastic bottle cap, plastic box bottom, duck toy beak, duck toy wing, rubber rabbit toy, rubber rabbit ear, rubber ball, shampoo bottle, glove, sponge brush handle, sponge brush, spoon, plastic box, plastic cup, plastic glove, tomato leaf, wooden cup, rubber toy, shuttlecock feather, marble slab, bamboo surface, bicycle basket, bicycle handle, bicycle pedal, bicycle seat, bicycle wheel, stone, block-patterned carpet, board brush, cable path, chair, desk, door handle, file envelope, glass door, hinge, bamboo leaf, metal post, metal sign, metal block, paper board, plastic chair, post box, tree branch, stone surface, stone pier, stopcock, tube, van surface, wall corner, well cap, wooden stair surface, wall surface, column, bowl, bowl bottom, coffee cup bottom, grape, paper bag, plastic box bottom edge, fiber chair, lock, metal net, metal stair surface, roadblock, toy banana, toy cheese, toy corn, toy lettuce, toy strawberry, toy tomato, toy bread, toy cookie, toy plum, mustard bottle, cleanser bottle, mug, power drill, hammer, sugar box, tomato soup can, salad plate, dinner plate, hair comb, decorative plate, soup ladle, salad fork, frying spatula, 8-inch skillet, 10.25-inch skillet, 10.5-inch griddle, dutch oven, dutch oven lid, rinsing cup, hand scoop, shovel toy, round plate, square plate, cutting board, wine glass, drinking cup, portion cup, cake pan, loaf pan, pestle, mortar, sculpture, ladle, spatula, decorative cast, small fork, large fork, soap dish, beer glass, large container, medium container, small container, vase, plate handle, plate, plate base, display stand, drop funnel, container lid, food pan, large flowerpot, small flowerpot, green vase, blue vase, orange vase, large swan, small swan, spoon holder, utensil container, potato masher, skimmer, pasta server, solid turner, slotted turner, green glass, red glass, scoop, box lid, Stanford frisbee, kettlebell, trim removal tool, bell pepper, dice, pear, pepper grinder, Rubik's cube, ABS plastic, aluminum, artificial grass, athletic shirt fabric, balsa wood, binder, brick, bubble envelope, CD sleeve, canvas, carbon fiber, ceramic, acrylic, coffee filter, cotton, dot paper, EPDM foam, felt, file portfolio, flannel, fleece, floor tile, wooden floor, folder, frosted acrylic, gift box, glitter paper, greeting card, leather back, leather front, MDF, metal foil, metal mesh, metal shelving, foam, wood, nylon bag, nylon mesh, nylon strap, package foam, painted brick, painted wood, green painted wood, pink foam, plastic mesh, pleather, polyethylene foam, portfolio cover, resin carbon fiber, resume paper, rough paper plate, silk, smooth paper plate, stained wood, stone tile, styrofoam, tarp, terra cotta, textured cloth, textured metal, textured paper, textured rubber, tissue paper, vinyl, wax paper, whiteboard.

Figure 11: Object categories in TouchThinker-1M. The figure lists object categories after removing articles, dataset sources, and non-essential descriptive modifiers for compact presentation.