STOLA: Self-Adaptive Touch-Language Framework for Tactile Commonsense Reasoning in Open-Ended Scenarios

Ning Cheng[1,2], **Jinan Xu**[1,2], **Jialing Chen**[1,2], **Bin Fang**[3], **Wenjuan Han**[1,2*]

[1] Key Laboratory of Big Data & Artificial Intelligence in Transportation (Beijing Jiaotong University), Ministry of Education
[2]School of Computer Science and Technology, Beijing Jiaotong University, Beijing 100044, China
[3]Beijing University of Posts and Telecommunications, 100876, Beijing, China

Abstract

This paper explores the challenges of integrating tactile sensing into intelligent systems for multimodal reasoning, particularly in enabling commonsense reasoning about the open-ended physical world. We identify two key challenges: **modality discrepancy**, where existing touch-language models often treat touch as a mere sub-modality of language without further addressing the semantic differences, and **open-ended tactile data scarcity**, where current datasets lack the diversity, open-endedness, and complexity needed for reasoning. To overcome these challenges, we introduce STOLA, a **S**elf-Adaptive **To**uch-**La**nguage framework. STOLA utilizes Mixture of Experts (MoE) to dynamically process, unify, and manage tactile and language modalities, capturing their unique characteristics. Crucially, we also present a comprehensive tactile commonsense reasoning dataset and benchmark featuring free-form questions and responses, 8 physical properties, 4 interactive characteristics, and diverse commonsense knowledge. Experiments show STOLA exhibits competitive performance compared to existing models on the PHYSICLEAR benchmark and self-constructed datasets, proving the effectiveness of the Mixture of Experts architecture in multimodal management and the performance advantages for open-scenario tactile commonsense reasoning tasks.

Code — https://github.com/cocacola-lab/STOLA
Extended version — https://arxiv.org/abs/2505.04201

Introduction

Human interactions with the physical world are fundamentally grounded in touch, a sense that surpasses the constraints of vision and hearing, offering direct, detailed, and multidimensional perception through physical contact (Fulkerson 2013; Paterson 2020; Packheiser et al. 2024). In the field of robotics and artificial intelligence, tactile sensing has been widely recognized as a critical modality for robots to interact with their surroundings, especially in scenarios with visual occlusion (Kappassov, Ramon, and Perdereau 2022; Lenz et al. 2024; Ueda et al. 2024).

However, integrating tactile sensing for reasoning presents significant challenges. These challenges can be distilled into two key issues. Firstly, a fundamental **modality discrepancy** exists: tactile and language modalities possess distinct characteristics, a fact underscored by dedicated neural pathways for touch processing. Current touch-language models often oversimplify this, treating touch as a mere "sub-modality" of language without further addressing the semantic differences. That is, these models use a touch encoder to map tactile data into a representation space that's very similar to the representation space used for text, and then force touch and language representation to fit into one transformer architecture (Yu et al. 2024; Yang et al. 2024), overlooking the fact that the two representations remain semantically distinct despite being mapped into a shared space. This lack of further distinction prevents the model from understanding the nuanced differences between the two modalities. Secondly, we face **open-ended tactile data scarcity**. Natural interaction demands intelligent systems capable of handling free-form queries and responses encompassing a broad spectrum of tactile properties. However, existing datasets, such as the recently recognized PHYSICLEAR (Yu et al. 2024), are limited in scope, focusing on a narrow range of properties and employing templated question-answer formats. This constraint severely restricts the generalization capabilities of models in real-world, open-ended scenarios.

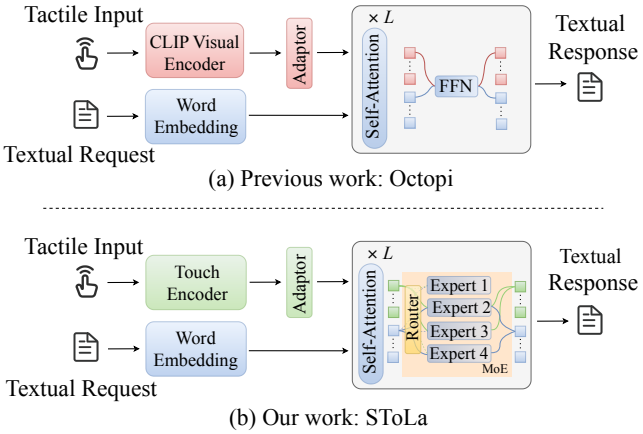


Figure 1: Comparison between (a) previous work and (b) our work across model.

*Corresponding author. Email address: wjhan@bjtu.edu.cn.

To address the challenge of modality discrepancy, we pioneer the exploration of the Mixture of Experts (MoE) to dynamically process diverse token types from both tactile and language modalities within the tactile domain, and propose **STOLA** (**S**elf-Adaptive **To**uch-**La**nguage), a framework capable of effectively handling both single tactile images and tactile sequential data, while leveraging MoE to seamlessly modality integration. As illustrated in Figure 1, STOLA primarily distinguishes itself from typical touch-language models, such as Octopi (Yu et al. 2024), by incorporating an MoE layer within the internal blocks of the Large Language Model (LLM). Each MoE-enabled block has a shared self-attention layer that is applicable to both tactile and language modalities, alongside the routers and feed-forward network (FFN) based experts to dynamically allocate token-level knowledge for both modalities. To further enhance training stability and expert collaboration, we employ a two-stage progressive training strategy. Through this model architecture and training strategy, we obtained an efficient and stable STOLA model.

To address the challenge of open-ended tactile data scarcity, we introduce a comprehensive commonsense reasoning dataset. This novel dataset transcends the limitations of existing resources, encompassing over 8 physical properties, 4 interactive characteristics, and diverse commonsense knowledge. Notably, the dataset features free-form queries and responses, designed to reflect the complexities of general tactile open-ended scenarios. In contrast, the widely recognized PHYSICLEAR (Yu et al. 2024) dataset, while valuable, focuses on a limited scope, exploring commonsense reasoning across only three physical properties: hardness, roughness, and bumpiness. Furthermore, despite its five reasoning tasks, PHYSICLEAR relies heavily on templated question-answer formats. For instance, the object property description subtask has only a few question formats, mainly variations of word substitutions, with responses following a fixed structure. The other four tasks are similar. However, in real-world scenarios, the form of both questions and responses is unpredictable, highlighting a significant gap between the current dataset and the demands of genuine open-ended tactile reasoning.

To validate the effectiveness of the STOLA framework including the model architecture, training strategy, and data set, we compare it with state-of-the-art touch-language models. Our experimental results demonstrate that dynamically managing different modalities through MoE significantly outperforms the traditional approach. Our contributions are summarized as follows:

- *Framework.* We propose STOLA, a pioneering MoE-based touch-language framework capable of processing diverse forms, including individual tactile images and tactile time-series data, as well as accommodating different sensor configurations (GelSight and GelSight Mini). We also present a progressive training paradigm to enhance the LLM's comprehension of the tactile and language modality.
- *Dataset.* We introduce a comprehensive tactile commonsense reasoning dataset for open-ended scenarios, featuring free-form questions and answers. The dataset covers over 8 physical properties, 4 interactive characteristics, and various commonsense knowledge from daily life.
- *Practice.* STOLA surpasses existing touch-language models in overall performance on PHYSICLEAR and TactileBench. STOLA also delivers competitive results across various subtasks.

Related Work

Tactile Commonsense Reasoning

Existing models process tactile signals by leveraging the powerful reasoning capabilities of LLMs. Yang *et al*. (Yang et al. 2024) aligns touch embeddings with image embeddings from the existing vision-language model (Zhang et al. 2023; Gao et al. 2023a) through contrastive learning, resulting in the creation of the Touch-LLM, a touch-language model capable of performing tactile question-answering tasks, including tactile commonsense reasoning tasks. Yu *et al*. (Yu et al. 2024) have significantly advanced tactile commonsense reasoning by formally introducing the PHYSICLEAR dataset, a training and evaluation suite based on three physical properties: hardness, roughness, and bumpiness. This suite includes five tactile commonsense reasoning tasks based on tactile temporal signals, leading to the development of Octopi, a touch-language model based on Vicuna v1.5 (Chiang et al. 2023). Unlike previous work, we focus more on free-form tactile commonsense reasoning in open scenarios that align with real-world distributions.

Mixture of Experts

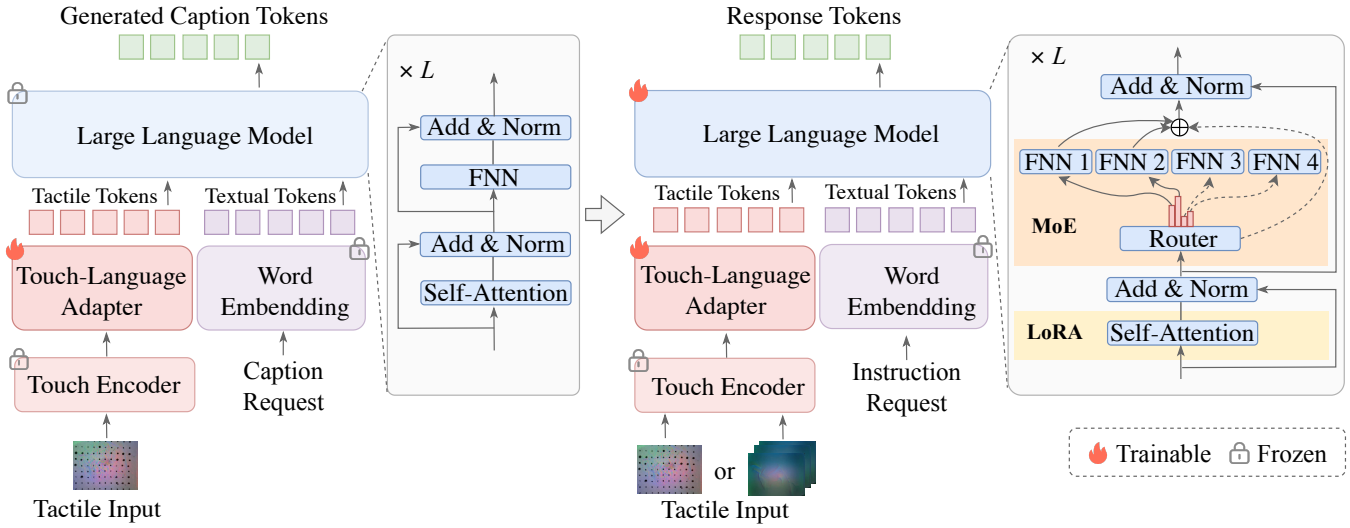
Mixture of Experts (MoE) aims to boost performance by selectively activating a subset of experts via a routing mechanism, enabling efficient handling of diverse data. Fedus *et al*. (Fedus, Zoph, and Shazeer 2022) propose Switch Transformer, Du *et al*. (Du et al. 2022) introduce GLaM, and Komatsuzaki *et al*. (Komatsuzaki et al. 2023) present sparse upcycling method, all of which demonstrate MoE's performance advantages and exceptional efficiency in language models. Additionally, MoE has achieved groundbreaking progress in vision models (Riquelme et al. 2021; Chen et al. 2023, 2024; Zhu et al. 2024). Recently, MoE has been widely applied in multimodal models (Lin et al. 2024; Shu et al. 2024; Li et al. 2025). In this work, we pioneer the integration of MoE into the large touch-language model, enabling finer differentiation, management, and interpretation across tactile and language modalities.

Method

We introduce a **S**elf-Adaptive **To**uch-**La**nguage Framwork (STOLA) for tactile commonsense reasoning, capable of handling different forms of tactile input, but also manages both tactile and language modalities effectively, as illustrated in Figure 2. In this section, we detail the STOLA's model architecture and the two-stage training strategy.

Model Architecture

Overview. As shown in Figure 2, STOLA comprises three key components: a touch encoder, a touch-language adapter,



(a) Stage I: Tactile Token Adaption for LLM

(b) Stage II: End-to-end Fine-tuning with MoE

Figure 2: **STOLA framework.** Our framework consists of a touch encoder, a touch-language adapter, and a LLM. The training process follows a two-stage strategy. Stage I: We train only the touch-language adapter, allowing the LLM to adapt to tactile inputs—static tactile images with spatial details. Stage II: The weights from Stage I are copied, keeping the touch encoder unchanged. The self-attention module of the LLM is fine-tuned using LoRA, while the FFN is upcycled from dense to sparse. Notably, we do not adjust the word embedding layer throughout the process.

and a LLM with MoE blocks. First, the touch encoder processes raw tactile data, transforming it into corresponding embeddings. Subsequently, the touch-language adapter bridges the modality gap, performing a coarse alignment of these embeddings with textual representations. Finally, and crucially, the LLM itself is augmented with Mixture of Experts (MoE) blocks. Next, we elaborate on the model architecture in the following paragraphs.

The *Input Unification* paragraph introduces how the touch encoder unifies diverse tactile signals and how the adapter and LLM integrate embeddings from the tactile and language modalities; *System Design* paragraph introduces the workflow of each component; *MoE Module* paragraph introduces how to apply the MoE module in the touch-language transformer.

Input Unification. To enable the self-adaptive touch-language model, it is essential to unify the model's inputs by seamlessly integrating tactile signals with text, harmonizing individual tactile images alongside time-series data, and accommodating diverse sensor configurations, such as Gelsight and Gelsight Mini. Following previous work (Gao et al. 2023b; Yang et al. 2024; Higuera et al. 2024; Feng et al. 2025; Dave, Lygerakis, and Rueckert 2024), tactile signals are divided into static images (individual tactile images) and dynamic videos (time-series data). To harmonize the two forms of tactile inputs, we process images as single-frame videos. Given a tactile video input $X_{touch} \in \mathbb{R}^{N\times H\times W\times C}$ with N frames, where H and W are the initial resolution of a frame, and C is the number of channels, with $C = 3$ in this case. The touch encoder encoders the N frames independently as a batch of tactile images and produces frame-level tactile token sequences $\mathcal{Z} = [[z_{11}, z_{12}, \cdots, z_{1P}], \cdots, [z_{i1}, z_{i2}, \cdots, z_{ij}, \cdots, z_{1P}], \cdots, [z_{N1}, z_{N2}, \cdots, z_{NP}]] \in \mathbb{R}^{N\times P\times C}$, where ij represents the j-th tactile token in the i-th frame, and $P = \frac{H\times W}{14^2}$ denotes the sequence length of tactile tokens. Each tactile token is a 14×14 patch. Inspired by ViFi-CLIP (Rasheed et al. 2023), these frame-level tactile token sequences are average-pooled to obtain a video-level tactile token sequence $\mathcal{Z}' \in \mathbb{R}^{P\times C}$. This operation aggregates multiple frames, implicitly incorporating temporal patterns. It is noteworthy that all sensor configurations, including both Gelsight and Gelsight Mini images, use the same processing method. Subsequently, the touch-language adapter f_{touch} is applied to transform $\mathcal{Z}' \in \mathbb{R}^{P\times C}$ to $\mathcal{V} \in \mathbb{R}^{P\times D}$, with D denoting the hidden size of the LLM. In addition, the text input is processed through a word embedding layer f_{text}, which maps the text input to the sequence tokens $\mathcal{T} = [t_1, t_2, \cdots, t_M] \in \mathbb{R}^{M\times D}$, where M refers to the length of text token sequence. Finally, we concatenate the tactile tokens and text tokens together and input the resulting sequence into the LLM.

System Design. We leverage the existing tactile representation model from TLV-Link (Cheng et al. 2024), which is well-aligned with the language modality, as our touch encoder. Meanwhile, we use a linear projection layer and Vicuna-7b v1.5 (Chiang et al. 2023) to function as our touch-language adapter and LLM, respectively. Given a tactile-textual instruction conversation (X, Y), STOLA produces response Y as follows:

$$Y = LLM_\phi(Proj_\lambda(Enc_\omega(X_{touch})), X_{text}), \quad (1)$$

where X_{touch} is the tactile input, and X_{text} is the text re-

quest. Enc, $Proj$, and LLM refer to the touch encoder, touch-language adapter, and LLM, respectively, with ω, λ, and ϕ denoting their corresponding parameters.

MoE Module. Considering that the commonsense reasoning task involves diverse tokens from multiple modalities, we introduce MoE layers into the model to dynamically select and activate experts, allowing it to adapt to varying input patterns. As shown in Figure 2, our MoE layer consists of a router and multiple experts. Specifically, we use a linear layer as the router and replicate the FFNs from Stage I to form K experts $\{E_i\}_{i=1}^{K}$. The router is responsible for predicting the activation probability of each expert for each token x, and this process can be formalized as:

$$\mathcal{P}(\mathbf{x}) = Softmax(\text{Top-}k(x \cdot W_r, k)), \tag{2}$$

where $W_r \in \mathbb{R}^{D\times K}$ represents the router's weight matrix, and Top-$k(x \cdot W_r, k)$ selects the top k experts based on the router's weight logits $x \cdot W_r \in \mathbb{R}^{1\times K}$. The Top-$k$ strategy for expert selection can be defined as:

$$\text{Top-}k_i(z,k) = \begin{cases} z_i, & \text{if } z_i \text{ is in the top k values of } z. \\ \infty, & \text{otherwise.} \end{cases} \tag{3}$$

Thus, the MoE layer output is calculated as the weighted sum of the experts' contributions, with the activation probabilities functioning as the weights:

$$\text{MoE}(x) = \sum_{i=1}^{K} \mathcal{P}_i(x) \cdot E_i(x). \tag{4}$$

Two-Stage Training Strategy

Stage I: Tactile Token Adaption for LLM. The goal at this stage is to adapt the tactile tokens converted by the touch encoder into LLM, enabling the LLM to interpret the content within the tactile inputs. The strategy involves using a touch-language adapter to map the tactile tokens into the LLM's text representation space, treating tactile patches as **pseudo**-text tokens. Specifically, the touch-language adapter is trained on tactile images and parallel language descriptions, while the touch encoder and LLM remain frozen. During this stage, the MoE layers are NOT employed to the LLM. We minimize the cross-entropy loss of the generated tokens to optimize the output $\mathcal{Y} = [y_1, y_2, \cdots, y_M] \in \mathbb{R}^{M\times D}$ of the LLM. The objective function is:

$$\mathcal{L}_{\text{ce}} = -\mathbb{E}_{(\mathcal{Y}_i|\mathcal{V},\mathcal{T}_{<i})\sim\pi_\theta}\left[\log \pi_\theta(\mathcal{Y}_i \mid \mathcal{V}, \mathcal{T}_{<i})\right], \tag{5}$$

where the training process employs teacher forcing (Williams and Zipser 1989), and $\pi_\theta(\mathcal{Y}_i \mid \mathcal{V}, \mathcal{T}_{<i})$ represents the likelihood of the predicted token $\mathcal{Y}_i$ conditioned on $\mathcal{V}$ and the first $i-1$ target tokens $\mathcal{T}_{<i}$.

Stage II: End-to-end Fine-tuning with MoE. In this stage, we aim to dynamically assign specialized experts to process diverse token types from both touch and language modalities, thereby enhancing the model's multimodal comprehension and generative abilities. To achieve this, we freeze only the touch encoder and word embedding layer, while fine-tuning the touch-language adapter and the LLM using instruction data. Within the LLM, we employ LoRA (Hu et al. 2022) for parameter-efficient fine-tuning of the self-attention layers, and crucially, replace the traditional feed-forward network (FFN) layers with Mixture of Experts (MoE) layers. This combination enables dynamic, token-level expertise allocation, significantly improving the model's ability to handle complex touch-language inputs. Specifically, for the MoE layers, a linear layer is employed as the router, with the FFN from stage I being replicated multiple times to facilitate the implementation of the experts. When tactile and textual tokens are fed into the MoE layers, the router calculates the weights of each expert for each token. Using the Top-k strategy, only the top-k experts are activated, and each token is processed by these activated experts. The resulting output is the weighted sum of the router's weights and the outputs from the activated experts. In addition to the cross-entropy loss from stage I, we introduce a differentiable load balancing loss (Fedus, Zoph, and Shazeer 2022) to the MoE layer as an auxiliary loss. The formula is as follows.

$$\mathcal{L}_{\text{aux}} = \alpha \cdot K \cdot \sum_{i=1}^{K} \mathcal{F}_i \cdot \mathcal{G}_i, \tag{6}$$

where α is the scaling factor, $\mathcal{F}_i$ represent the probability of tokens assigned to expert E_i and $\mathcal{G}_i$ represents the router probability allocated to expert E_i, respectively. $\mathcal{F}_i$ and $\mathcal{G}_i$ are calculated using the following formulas:

$$\mathcal{F}_i = \frac{1}{P+M} \sum_{i=1}^{K} 1\{\text{argmax}\, \mathcal{P}(\mathbf{x}) = i\}, \tag{7}$$

$$\mathcal{G}_i = \frac{1}{P+M} \sum_{i=1}^{P+M} \mathcal{P}_i(\mathbf{x}). \tag{8}$$

Thus, the objective function for this stage is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ce}} + \mathcal{L}_{\text{aux}}. \tag{9}$$

TactileBench

To evaluate models' reasoning capabilities in open-ended unstructured environments, we designed a comprehensive benchmark encompassing tasks ranging from fundamental understanding to complex reasoning. Unlike PHYSICLEAR, which was designed around specific task objectives with templated questions and answers, our benchmark dataset focuses on the hierarchical cognitive process from perception to reasoning. Rather than following a fixed format, we adopt an open-ended approach that better evaluates the model's depth of understanding and complexity.

Task Category

The underlying tasks involve open-ended responses in a free-form manner. The benchmark consists of three progressively challenging subtasks:

Fundamental Property Understanding. The subtask involves recognizing and describing an object's basic physical properties, including but not limited to hardness, roughness, weight, and texture. The model needs to perceive these properties through tactile signals and convert them into human-understandable textual outputs.

Model	PHYSICLEAR			TactileBench		
	CIDEr	B@4	METEOR	METEOR	GPT-4	DeepSeek-R1
Touch-LLM (Yang et al. 2024)	-	-	-	17.92	6.88	7.06
Octopi-7B (Yu et al. 2024)	138.60	64.16	77.63	21.47	6.91	7.17
Octopi-13B (Yu et al. 2024)	141.20	64.33	77.79	28.83	7.85	7.97
STOLA (Ours)	**195.03**	**68.03**	**82.58**	**30.27**	**8.02**	**8.12**

Table 1: **Overall performance comparison on the PHYSICLEAR and TactileBench benchmark.** The best results are shown in **bold**, and the suboptimal ones are highlighted with underline. For Touch-LLM, which does not support the PHYSICLEAR dataset with interleaved tactile temporal signals and text, the corresponding results are represented with "–".

Model	PC	PSS	POM	PSR	OPD			
					Combined	Hardness	Roughness	Bumpiness
Random	33.33	33.33	16.67	50.00	3.70	33.33	33.33	33.33
Octopi-7B (Yu et al. 2024)	48.10	74.67	44.39	69.57	47.37	71.05	73.68	81.58
Octopi-7B* (Yu et al. 2024)	43.71	63.43	39.49	69.39	20.51	28.21	71.79	**92.31**
Octopi-13B (Yu et al. 2024)	55.06	**84.00**	**60.43**	67.39	**55.26**	**73.68**	78.95	78.95
Octopi-13B* (Yu et al. 2024)	41.92	66.29	50.32	67.35	20.53	30.77	76.92	82.05
STOLA (Ours)	**62.28**	74.86	57.32	**69.80**	48.72	61.54	**82.05**	82.05

Table 2: **Subtasks accuracy comparison with accuracy score on the PHYSICLEAR benchmark.** The results marked with * are the ones reproduced by us using the open-source models and scripts provided in the original paper[1]. The best results are in **bold**, and the second-best ones are underlined. **PC**: Property Comparison. **PSS**: Property Superlative Selection. **POM**: Property-object Matching. **PSR**: Property Scenario Reasoning. **OPD**: Object Property Description. Combined: The generated results for hardness, roughness, and bumpiness are all correct.

Model	FPU			TIP			CDR		
	METEOR	GPT-4	DeepSeek-R1	METEOR	GPT-4	DeepSeek-R1	METEOR	GPT-4	DeepSeek-R1
Touch-LLM (Yang et al. 2024)	15.49	7.01	7.16	17.27	6.42	6.51	24.98	7.24	7.67
Octopi-7B (Yu et al. 2024)	21.87	6.65	7.04	22.55	7.13	7.15	18.82	7.26	7.52
Octopi-13B (Yu et al. 2024)	29.70	7.81	7.96	29.89	7.81	7.87	25.04	**8.00**	**8.15**
STOLA (Ours)	**31.34**	**8.19**	**8.28**	**31.24**	**8.03**	**7.97**	**26.15**	7.61	7.96

Table 3: **Subtasks performance comparison on the TactileBench.** The best results are in **bold**, and the second-best ones are underlined. **FPU**: Fundamental property Understanding. **TIP**: Tactile Interaction Perception. **CDR**: Commonsense-Driven Reasoning.

Tactile Interaction Perception. The subtask involves sensing an object's dynamic characteristics during real-world interactions. These characteristics include graspability, prickliness, bendability, malleability, and more. The model needs to perceive and respond to these complex tactile signals in real time during dynamic interactions with objects.

Commonsense-Driven Reasoning. The subtask requires the model not only to perceive object's tactile properties but also to integrate external commonsense knowledge for reasoning. These tasks often involve understanding an object's behavior, function, or usage in specific scenarios, as well as making high-level decisions based on tactile information.

Data Construction

We use the material classification test set from *Touch and Go* (Yang et al. 2022) as our baseline data. This test set consists of visual images, tactile images, and classification labels. To construct tactile question-answering data for three types of tasks, we designed a unified prompt template using visual images and classification labels, queried GPT-4, and conducted manual verification.

Data Statistics

Given the significance of fundamental property understanding, the auxiliary role in tactile interaction perception, and the challenges of commonsense-driven reasoning, the data are stratified into a 50%-30%-20% proportion to ensure a balanced difficulty distribution. Ultimately, we compile a total of 600 questions, each with 3-5 ground-truth answers, covering 14 objects.

Evaluation Metrics

Since TactileBench is an open-ended, free-form tactile commonsense reasoning dataset, traditional n-gram-based metrics such as CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015) and B@4 (Papineni et al. 2002) are not suitable. Instead, we evaluate the dataset using METEOR (Banerjee and Lavie 2005), which is based on semantic sim-

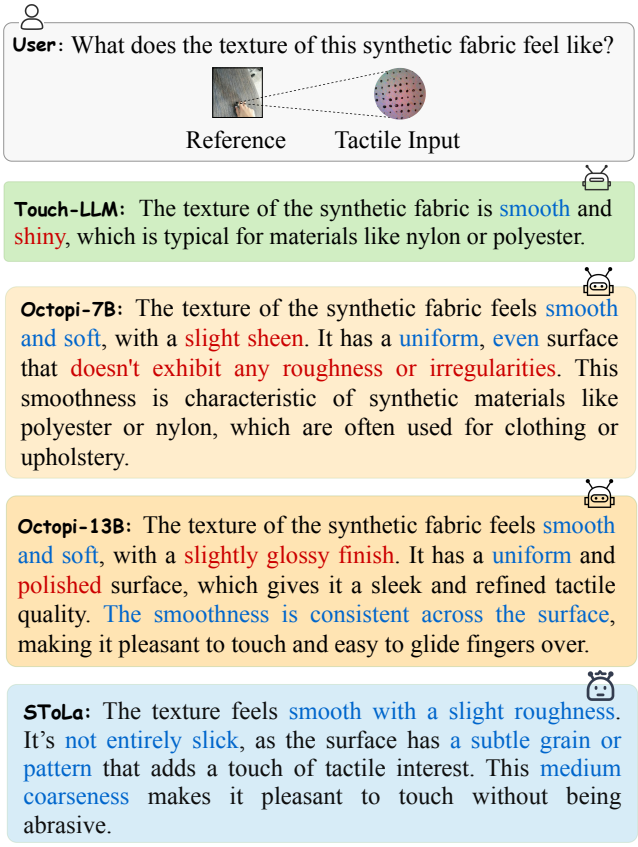


Figure 3: **Qualitative responses of SToLa and existing touch-language models.** Errors are highlighted in red, while detailed and accurate parts are emphasized in blue. For visualization, we include the visual signals of the tactile input (labeled as "reference"), even though it is not used for the model.

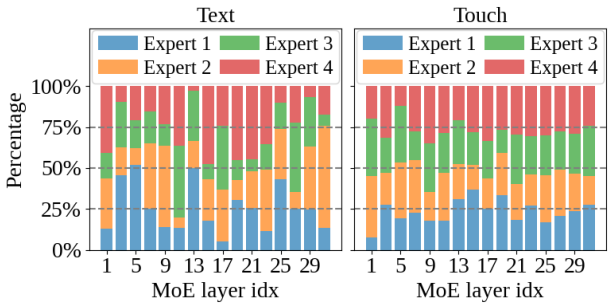


Figure 4: Distribution of modalities across different experts.

ilarity, along with multi-dimensional scoring from GPT-4 (Achiam et al. 2023) and DeepSeek–R1 (Guo et al. 2025).

Experiments

Implementation Details

Models are trained with batch size 16 on an Nvidia A100-80G GPU. During the implementation of our two-stage

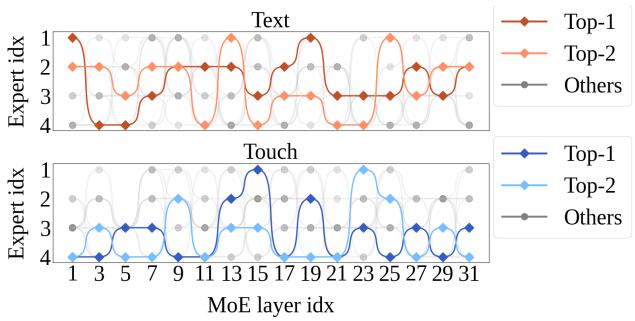


Figure 5: Visualization of activated pathways.

training strategy, we treat Stage I as a "Touch-to-Text" generation task, trained on touch-language pairs, with the objective being to prompt the LLM to generate a corresponding text description given a tactile input. Specifically, we use touch-language pairs from Touch100k. Stage II is regarded as a process of instruction tuning, aimed at enhancing the model's capabilities and controllability. In this stage, we use the video-based PhysiCLeAR dataset and our self-constructed image-based tactile instruction dataset. Both datasets are processed into video units by frames. In particular, the base model of SToLa adopts Vicuna-7B v1.5.

Results

We conducted a quantitative comparison between SToLa and current state-of-the-art touch-language models, including Touch-LLM (Yang et al. 2024), Octopi-7B (Yu et al. 2024), and Octopi-13B (Yu et al. 2024). Table 1 presents the overall performance on the PhysiCLeAR and TactileBench datasets, while Table 2 and Table 3 provide results comparisons for the subtasks in each dataset, respectively.

Compared to Touch-LLM and Octopi-7B, which use the equivalent scale of 7B, SToLa achieves the **best** overall performance in both PhysiCLeAR and TactileBench datasets, as well as across all eight subtasks within these datasets. Since Touch-LLM does not support data in which tactile temporal signals and text appear interleaved, we did not evaluate Touch-LLM's performance on the PhysiCLeAR dataset.

Moreover, SToLa of 7B outperforms Octopi-13B in overall performance and most subtasks on the PhysiCLeAR dataset. Specifically, SToLa outperforms the previous best-performing Octopi-13B by 53.83, 3.70, and 4.79 in the overall performance metrics of CIDEr, B@4, and METEOR, respectively. Additionally, SToLa shows superior performance on the property comparison and property scenario reasoning subtasks while achieving suboptimal results only to Octopi-13B on the property superlative selection, property-object matching, and object property description subtasks, demonstrating the strong competitive capabilities of SToLa with a 7B LLM. It is important to note that we follow the evaluation of Octopi (Yu et al. 2024) and adopt accuracy as the metric for subtasks. That is, based on templated responses, only the conclusion part of the response is evaluated using exact matching, without considering the

Model	PHYSICLEAR			TactileBench		
	CIDEr	B@4	METEOR	METEOR	GPT-4	DeepSeek-R1
STOLA	**195.03**	**68.03**	**82.58**	**30.27**	**8.02**	**8.12**
w/o MoE	176.79	66.46	81.55	28.71	7.44	7.57
w/o LoRA	166.71	64.46	80.39	29.32	7.95	7.97
w/o Stage I	172.52	64.47	80.55	29.27	7.72	7.89

Table 4: Ablation study on PHYSICLEAR and TactileBench.

reasoning process or semantics.

Similarly, on the TactileBench dataset, STOLA of 7B outperforms Octopi-13B in overall performance, whether measured by the METEOR metric or the GPT-4 and DeepSeek-R1 evaluations. Although Octopi-13B performed slightly under on the commonsense-driven reasoning subtask, STOLA demonstrates significant advantages in achieving such remarkable results with a significantly smaller parameter size (7B$<$13B).

Furthermore, we present a qualitative comparison of responses from different models in Figure 3. The comparison shows that our model excels in interpreting tactile signals and performing tactile reasoning, further confirming the advantages of our approach. More quantitative comparisons can be found in the extended version.

MoE Analysis

We analyze how different experts in each MoE layer of STOLA dynamically manage the tactile and text modalities. Specifically, the routing distributions and token pathways of STOLA on TactileBench are visualized in Figure 4 and 5.

Routing Distributions. Figure 4 illustrates the modality distribution handled by different experts, revealing that each expert develops its preferences. Differences in the selection of tactile and text modalities indicate that our model can dynamically adjust expert utilization based on input characteristics, enabling effective processing of multimodal data.

Token Pathways. We track the trajectories of all tokens and analyze expert behavior at the token level, as shown in Figure 5. Following the previous work (Li et al. 2025; Lin et al. 2024), we employ PCA (Pearson 1901) to extract the top 10 pathways for all activated routes. The token pathways also reflect the preferences of different experts for tactile and text modalities across each MoE layer. For a tactile token, STOLA tends to assign it to experts 3 and 4. Meanwhile, for a text token, STOLA usually prefers to assign it to a combination of expert 2 and another expert in the shallow layers, while in the deeper layers, expert 3, along with another expert, tends to be assigned. This indicates that STOLA has learned a specific pattern that enables it to manage the tactile and text modalities in a certain manner.

Ablation Study

In this section, we perform ablation studies to analyze the impact of key factors in STOLA model, including the MoE module, LoRA, and training strategy, as shown in Table 4.

Impact of MoE Module. To evaluate the MoE module's contribution, we compare models' performance with and without MoE. Specifically, we replace the MoE layer with a standard feedforward network while keeping all other configurations unchanged. The results show that removing the MoE module significantly drops model performance, indicating that the expert routing mechanism plays a crucial role in enhancing performance.

Impact of LoRA Fine-tuning. To investigate the impact of LoRA, we ablate low-rank adaptation based on our training processing. The results in Table 4 indicate that LoRA fine-tuning has a significant positive effect on model performance. Without LoRA fine-tuning, the model's performance degrades significantly.

Impact of Training Strategy. We further validate the necessity of Stage I in the training strategy by removing it and directly applying the training of Stage II. The experimental results demonstrate that skipping Stage I leads to a decrease in model performance, which proves that the progressive training strategy is crucial to gradually optimizing model capability.

Conclusion

This work introduces STOLA, a self-adaptive touch-language framework that pushes the boundaries of tactile commonsense reasoning. By incorporating MoE and two-stage training strategy, STOLA effectively bridges the gap between tactile and language modalities, enabling more complex reasoning of the open-ended world. Moreover, we contribute a comprehensive tactile commonsense reasoning dataset that covers a wider range of dimensions, with a distribution that aligns with cognitive levels and features free-form content. Finally, experiments demonstrate that our model achieves outstanding performance in tactile commonsense reasoning in open-ended scenarios.

Although STOLA performs competitive capabilities, computational resource constraints hinder the extension of our method to a 13B LLM, leading to suboptimal performance in certain tasks compared to Octopi-13B. Additionally, we designed the MoE architecture from the modality level. In future work, we will consider assigning experts from a task perspective or a combination of modality and task perspective.

Acknowledgements

The work described in this paper has been supported by Fundamental Research Funds for the Central Universities under Grant No. 2025JBZX058, and by National Natural Science Foundation of China under Grant No. 62376019, 62476023, 62406020, 62573063, 62536001.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Banerjee, S.; and Lavie, A. 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*, 65–72.

Chen, J.; Guo, L.; Sun, J.; Shao, S.; Yuan, Z.; Lin, L.; and Zhang, D. 2024. EVE: efficient vision-language pre-training with masked prediction and modality-aware moe. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 1110–1119.

Chen, T.; Chen, X.; Du, X.; Rashwan, A.; Yang, F.; Chen, H.; Wang, Z.; and Li, Y. 2023. Adamv-moe: Adaptive multi-task vision mixture-of-experts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 17346–17357.

Cheng, N.; Guan, C.; Gao, J.; Wang, W.; Li, Y.; Meng, F.; Zhou, J.; Fang, B.; Xu, J.; and Han, W. 2024. Touch100k: A Large-Scale Touch-Language-Vision Dataset for Touch-Centric Multimodal Representation. *arXiv preprint arXiv:2406.03813*.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3): 6.

Dave, V.; Lygerakis, F.; and Rueckert, E. 2024. Multimodal visual-tactile representation learning through self-supervised contrastive pre-training. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 8013–8020. IEEE.

Du, N.; Huang, Y.; Dai, A. M.; Tong, S.; Lepikhin, D.; Xu, Y.; Krikun, M.; Zhou, Y.; Yu, A. W.; Firat, O.; et al. 2022. Glam: Efficient scaling of language models with mixture-of-experts. In *International conference on machine learning*, 5547–5569. PMLR.

Fedus, W.; Zoph, B.; and Shazeer, N. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120): 1–39.

Feng, R.; Hu, J.; Xia, W.; Shen, A.; Sun, Y.; Fang, B.; Hu, D.; et al. 2025. AnyTouch: Learning Unified Static-Dynamic Representation across Multiple Visuo-tactile Sensors. In *The Thirteenth International Conference on Learning Representations*.

Fulkerson, M. 2013. *The first sense: A philosophical study of human touch*. MIT press.

Gao, P.; Han, J.; Zhang, R.; Lin, Z.; Geng, S.; Zhou, A.; Zhang, W.; Lu, P.; He, C.; Yue, X.; et al. 2023a. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*.

Gao, R.; Dou, Y.; Li, H.; Agarwal, T.; Bohg, J.; Li, Y.; Fei-Fei, L.; and Wu, J. 2023b. The objectfolder benchmark: Multisensory learning with neural and real objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17276–17286.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Higuera, C.; Sharma, A.; Bodduluri, C. K.; Fan, T.; Lancaster, P.; Kalakrishnan, M.; Kaess, M.; Boots, B.; Lambeta, M.; Wu, T.; et al. 2024. Sparsh: Self-supervised touch representations for vision-based tactile sensing. In *8th Annual Conference on Robot Learning*.

Hu, E. J.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Kappassov, Z.; Ramon, J. A. C.; and Perdereau, V. 2022. Tactile-based task definition through edge contact formation setpoints for object exploration and manipulation. *IEEE Robotics and Automation Letters*, 7(2): 5007–5014.

Komatsuzaki, A.; Puigcerver, J.; Lee-Thorp, J.; Ruiz, C. R.; Mustafa, B.; Ainslie, J.; Tay, Y.; Dehghani, M.; and Houlsby, N. 2023. Sparse Upcycling: Training Mixture-of-Experts from Dense Checkpoints. In *The Eleventh International Conference on Learning Representations*.

Lenz, J.; Gruner, T.; Palenicek, D.; Schneider, T.; Pfenning, I.; and Peters, J. 2024. Analysing the Interplay of Vision and Touch for Dexterous Insertion Tasks. In *CoRL Workshop on Learning Robot Fine and Dexterous Manipulation: Perception and Control*.

Li, Y.; Jiang, S.; Hu, B.; Wang, L.; Zhong, W.; Luo, W.; Ma, L.; and Zhang, M. 2025. Uni-moe: Scaling unified multimodal llms with mixture of experts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Lin, B.; Tang, Z.; Ye, Y.; Cui, J.; Zhu, B.; Jin, P.; Huang, J.; Zhang, J.; Pang, Y.; Ning, M.; et al. 2024. Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947*.

Packheiser, J.; Hartmann, H.; Fredriksen, K.; Gazzola, V.; Keysers, C.; and Michon, F. 2024. A systematic review and multivariate meta-analysis of the physical and mental health benefits of touch interventions. *Nature Human Behaviour*, 1–20.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Paterson, M. 2020. *The senses of touch: Haptics, affects and technologies*. Routledge.

Pearson, K. 1901. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11): 559–572.

Rasheed, H.; Khattak, M. U.; Maaz, M.; Khan, S.; and Khan, F. S. 2023. Fine-tuned clip models are efficient video learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6545–6554.

Riquelme, C.; Puigcerver, J.; Mustafa, B.; Neumann, M.; Jenatton, R.; Susano Pinto, A.; Keysers, D.; and Houlsby, N. 2021. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34: 8583–8595.

Shu, F.; Liao, Y.; Zhuo, L.; Xu, C.; Zhang, G.; Shi, H.; Chen, L.; Zhong, T.; He, W.; Fu, S.; et al. 2024. LLaVA-MoD: Making LLaVA Tiny via MoE Knowledge Distillation. *CoRR*.

Ueda, S.; Hashimoto, A.; Hamaya, M.; Tanaka, K.; and Saito, H. 2024. Visuo-Tactile Zero-Shot Object Recognition with Vision-Language Model. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 7243–7250. IEEE.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4566–4575.

Williams, R. J.; and Zipser, D. 1989. A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2): 270–280.

Yang, F.; Feng, C.; Chen, Z.; Park, H.; Wang, D.; Dou, Y.; Zeng, Z.; Chen, X.; Gangopadhyay, R.; Owens, A.; et al. 2024. Binding touch to everything: Learning unified multimodal tactile representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26340–26353.

Yang, F.; Ma, C.; Zhang, J.; Zhu, J.; Yuan, W.; and Owens, A. 2022. Touch and go: learning from human-collected vision and touch. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 8081–8103.

Yu, S.; Lin, K.; Xiao, A.; Duan, J.; and Soh, H. 2024. Octopi: Object Property Reasoning with Large Tactile-Language Models. In *Robotics: science and systems*.

Zhang, R.; Han, J.; Liu, C.; Gao, P.; Zhou, A.; Hu, X.; Yan, S.; Lu, P.; Li, H.; and Qiao, Y. 2023. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*.

Zhu, X.; Guan, Y.; Liang, D.; Chen, Y.; Liu, Y.; and Bai, X. 2024. Moe jetpack: From dense checkpoints to adaptive mixture of experts for vision tasks. *Advances in Neural Information Processing Systems*, 37: 12094–12118.