Towards Large Tactile-Language Models

Samson Yu[†], Kelvin Lin[†], Anxing Xiao[†], Jiafei Duan[§], and Harold Soh[†‡]
Email: {samsonyu,harold}@comp.nus.edu.sg
[†]Dept. of Computer Science, National University of Singapore
[§]University of Washington, [‡]NUS Smart Systems Institute

***Abstract*—In this work, we investigate combining tactile perception with language, which enables embodied systems to obtain physical properties through interaction and apply common-sense reasoning. We contribute a new dataset PHYSICLEAR, which comprises both physical/property inference tasks and annotated tactile videos obtained using a GelSight tactile sensor. We then introduce OCTOPI, a system that leverages both tactile representation learning and large vision-language models to process tactile inputs with minimal language fine-tuning. Our evaluations on PHYSICLEAR show that OCTOPI is able to effectively use intermediate physical property predictions to improve physical reasoning in both trained tasks and for zero-shot reasoning. PHYSICLEAR and OCTOPI are available on https://github.com/clear-nus/octopi.**

I. INTRODUCTION

For humans, touch is a crucial sense that provides physical information beyond what vision can provide (e.g., material properties, texture information, temperature) especially during occlusion. This in turn improves our ability to perform physical reasoning [11, 2] and act in our world. Here, we are interested in enabling general purpose robots, specifically those empowered with large-language models (LLM), to perform similar physical reasoning. Recent work has demonstrated that LLMs and large vision-language models (LVLMs) can provide an impressive level of common-sense and physical inference [15, 19, 1, 18] but focus on either visual or text input modalities.

This extended abstract describes our approach towards giving LVLMs the sense of touch. We hypothesize that incorporating a tactile modality into LVLMs will enable better physical inference/reasoning in real-world environments. As an example, Fig. 1 illustrates how commonsense knowledge is applied together with tactile information to complete a novel physical task in a zero-shot manner. Here, the robot leverages its tactile inputs together with the LLM's commonsense knowledge (that ripe avocados are soft) to correctly select the ripe avocado (despite not having been trained with tactile inputs from avocados). Here, we use visual-tactile sensors, i.e., the GelSight [21], that provides image frames that reveal physical object properties such as texture and hardness [23]. However, there remains a significant domain gap between tactile data and the natural images that typical LVLMs are trained with.

II. PHYSICLEAR: A DATASET FOR TACTILE-BASED PHYSICAL PROPERTY REASONING

To bridge this gap, we created the PHYSICLEAR dataset, which comprises GelSight images on a variety of real world objects, along with object labels and part annotations. PHYSICLEAR complements existing tactile datasets [23, 22, 8, 20, 7] as it provides three physical property annotations, specifically hardness, roughness, and bumpiness, that have been used in prior research [14, 9, 12, 4, 3, 10] and can be potentially inferred from the GelSight data. PHYSICLEAR includes an training and evaluation suite comprising five reasoning tasks, which can serve as a benchmark for the research community:

Fig. 1. Zero-shot avocado ripeness selection by combining tactile information with commonsense knowledge. Using inputs from its tactile sensor, OCTOPI identifies the left avocado as softer. Using commonsense reasoning, OCTOPI infers that it is ripe and fulfils the user's request. *Note: intermediate prompts removed due to space.*

- **Object Property Description (OPD).** This task addresses property-based description: generating both unstructured and structured descriptions of an object's hardness, roughness, and bumpiness from tactile videos.
- **Property Comparison (PC).** Given two tactile videos, each of a different object, and a specified physical property, and its comparative adjective, determine whether the comparative adjective describes the two videos.
- **Property Superlative Selection (PSS).** For three tactile videos, each of a different object, and a specified physical property and its superlative adjective (e.g. hardest for the hardness property), choose the video that the superlative adjective best describes.
- **Property-object Matching (POM).** This task requires matching physical properties to objects: given three tactile videos and three specified objects, the goal is to correctly associate each video with an object.
- **Property Scenario Reasoning (PSR).** We provide two

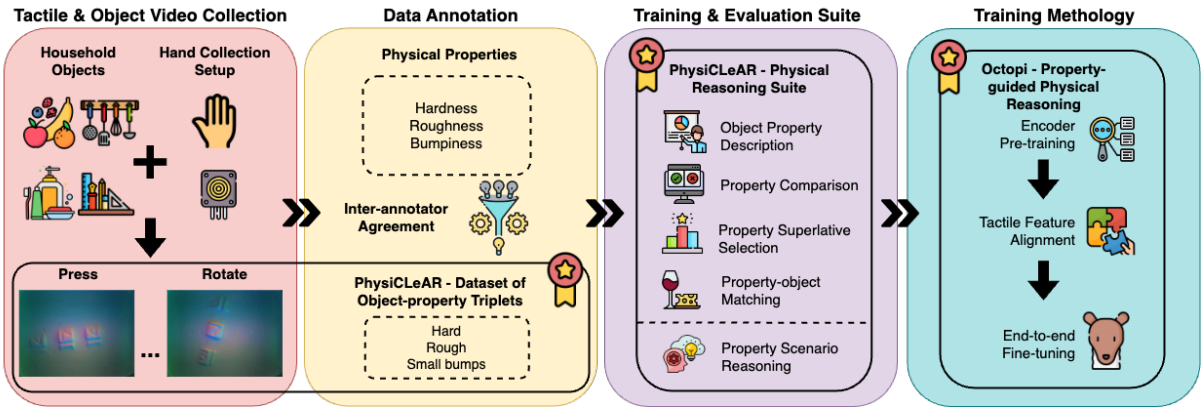

Fig. 2. **PHYSICLEAR and OCTOPI.** The full pipeline of our system with key contributions starred. We collect tactile videos for everyday household objects by hand with two exploratory procedures: pressing and rotation. The videos are annotated by three annotators for three physical properties: hardness, roughness and bumpiness. PHYSICLEAR leverages the videos and annotations for five language-driven physical description and understanding tasks. OCTOPI is a LVLM fine-tuned on PHYSICLEAR for tactile-grounded physical inference.

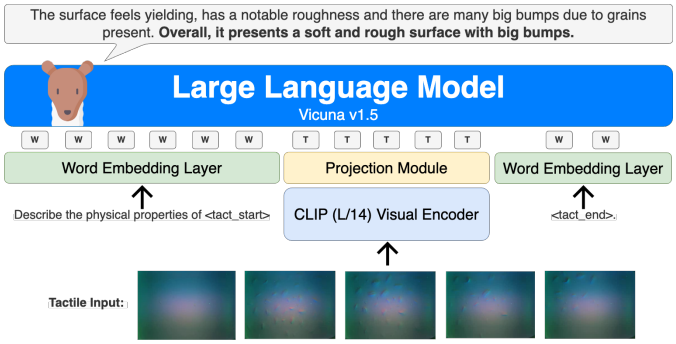

Fig. 3. **OCTOPI Framework.** Our framework consists of CLIP's visual encoder, a projection module with two linear layers and Vicuna v1.5 as the LLM. Language embeddings are derived through tokenization and then Vicuna's word embedding layer, with <tact_start> and <tact_end> indicating the start and end of a tactile frame sequence from a single tactile sensor. Tactile frames are fed into the visual encoder and then the projection module to derive tactile embeddings with the same dimension as the word embeddings.

tactile videos, each showcasing a different object, along with a real-world scenario that relies on one or more of our defined physical properties. The task is to choose the video that represents the object whose physical properties best meet the scenario's demands.

III. OCTOPI: A TACTILE-LANGUAGE MODEL

Using PHYSICLEAR, we developed OCTOPI (Object Comprehension with Tactile Observations for Physical Intelligence, Fig. 3). OCTOPI is a LLaMA-based [16, 17] LVLM (Vicuna [5]) equipped with a CLIP-based [13] tactile encoder, whose representations have been aligned via projection. In experiments, we show that OCTOPI is able to use its tactile modality to predict project properties and reason about scenarios (Tbl. I), and even predict avocado ripeness (Tbl. II) better than a visual-language model [6]. More broadly, these results suggest that incorporating the tactile modality helps large models better infer object properties.

TABLE I. **Results on PHYSICLEAR Physical Understanding Tasks.** OCTOPI's performance on physical understanding tasks improves with object property descriptions (OPD).

	Random	OCTOPI-7b	OCTOPI-7b (no OPD)
PC	33.33	**48.10**	46.51
PSS	33.33	**74.67**	39.88
POM	16.67	**44.39**	23.23
PSR	50.00	**69.57**	63.04
	Random	**OCTOPI-13b**	**OCTOPI-13b (no OPD)**
PC	33.33	**55.06**	40.70
PSS	33.33	**84.00**	39.88
POM	16.67	**60.43**	18.71
PSR	50.00	**67.39**	39.13

TABLE II. **Zero-Shot Avocado Property Prediction and Ripeness Classification Results.** OCTOPI-13b predicts avocado properties reasonably well with only a pressing motion. For avocado ripeness classification, OCTOPI-13b is able to leverage its commonsense knowledge to use both *hardness* and *bumpiness* properties.

	Random	OCTOPI-13b	VLM Baseline
Property Prediction			
Hardness	33.33	**57.50**	37.50
Roughness	33.33	**71.00**	3.00
Bumpiness	33.33	**64.00**	9.50
Ripeness Classification	50.00	**63.00**	-*

* The VLM would always predict the first object.

IV. CONCLUSIONS AND FUTURE WORK.

In this work, we have contributed the following:

- A new GelSight dataset, PHYSICLEAR, that exhibits property diversity, object diversity, and material diversity for selected physical properties.
- OCTOPI, a framework for physical reasoning that leverages vision-based optical tactile sensors, pre-trained vision models, and the reasoning capabilities of LLMs.
- An accompanying training and evaluation suite spanning five tasks and baseline results using OCTOPI.

We hope that PHYSICLEAR and OCTOPI will spur research in tactile-enabled physical reasoning for embodied AI systems. We are currently working on improving OCTOPI's performance and incorporating the visual modality.

References

[1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

[2] Anton Bakhtin, Laurens van der Maaten, Justin Johnson, Laura Gustafson, and Ross Girshick. PHYRE: A New Benchmark for Physical Reasoning. 2019.

[3] Wouter M. Bergmann Tiest. Tactual perception of material properties. *Vision Research*, 50(24):2775–2782, 2010. Perception and Action: Part I.

[4] X. Chen, Fei Shao, Cathy Barnes, Tom Childs, and Brian Henson. Exploring Relationships between Touch Perception and Surface Physical Properties. *International Journal of Design*, 3:67–76, 08 2009.

[5] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023.

[6] Jensen Gao, Bidipta Sarkar, Fei Xia, Ted Xiao, Jiajun Wu, Brian Ichter, Anirudha Majumdar, and Dorsa Sadigh. Physically grounded vision-language models for robotic manipulation. *arXiv preprint arXiv:2309.02561*, 2023.

[7] Ruohan Gao, Yiming Dou, Hao Li, Tanmay Agarwal, Jeannette Bohg, Yunzhu Li, Li Fei-Fei, and Jiajun Wu. The ObjectFolder Benchmark: Multisensory Learning With Neural and Real Objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17276–17286, June 2023.

[8] Ruohan Gao, Zilin Si, Yen-Yu Chang, Samuel Clarke, Jeannette Bohg, Li Fei-Fei, Wenzhen Yuan, and Jiajun Wu. Objectfolder 2.0: A multisensory object dataset for sim2real transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10598–10608, 2022.

[9] Yang Gao, Lisa Anne Hendricks, Katherine J Kuchenbecker, and Trevor Darrell. Deep learning for tactile understanding from visual and haptic data. In *2016 IEEE international conference on robotics and automation (ICRA)*, pages 536–543. IEEE, 2016.

[10] Jiaqi Jiang and Shan Luo. Robotic perception of object properties using tactile sensing. In *Tactile Sensing, Skill Learning, and Robotic Dexterous Manipulation*, pages 23–44. Elsevier, 2022.

[11] Andrew Melnik, Robin Schiewer, Moritz Lange, Andrei Muresanu, Mozhgan Saeidi, Animesh Garg, and Helge Ritter. Benchmarks for Physical Reasoning AI. *arXiv preprint arXiv:2312.10728*, 2023.

[12] Matthew Purri and Kristin Dana. Teaching cameras to feel: Estimating tactile physical properties of surfaces from images. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*, pages 1–20. Springer, 2020.

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

[14] Kuniyuki Takahashi and Jethro Tan. Deep visuo-tactile learning: Estimation of tactile properties from images. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8951–8957. IEEE, 2019.

[15] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[16] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[17] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[18] Haoran Wei, Lingyu Kong, Jinyue Chen, Liang Zhao, Zheng Ge, Jinrong Yang, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. Vary: Scaling up the Vision Vocabulary for Large Vision-Language Models. *arXiv preprint arXiv:2312.06109*, 2023.

[19] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*, 2023.

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

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

[22] Wenzhen Yuan, Yuchen Mo, Shaoxiong Wang, and Edward H Adelson. Active clothing material perception using tactile sensing and deep learning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4842–4849. IEEE, 2018.

[23] Wenzhen Yuan, Mandayam A. Srinivasan, and Edward H. Adelson. Estimating object hardness with a gelsight touch sensor. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 208–215, 2016.