Robotics: Science and Systems 2025
Los Angeles, CA, USA, June 21–25, 2025

Demonstrating the Octopi-1.5 Visual-Tactile-Language Model

Samson Yu[†], Kelvin Lin[†], and Harold Soh[†‡]
[†]Dept. of Computer Science, National University of Singapore
[‡]NUS Smart Systems Institute
Contact Authors: samson.yu@u.nus.edu, harold@comp.nus.edu.sg

***Abstract*—Touch is recognized as a vital sense for humans and an equally important modality for robots, especially for dexterous manipulation, material identification, and scenarios involving visual occlusion. Building upon very recent work in touch foundation models, this demonstration will feature Octopi-1.5, our latest visual-tactile-language model. Compared to its predecessor, Octopi-1.5 introduces the ability to process tactile signals from multiple object parts and employs a simple retrieval-augmented generation (RAG) module to improve performance on tasks and potentially learn new objects on-the-fly. The system can be experienced live through a new handheld tactile-enabled interface, the TMI, equipped with GelSight and TAC-02 tactile sensors. This convenient and accessible setup allows users to interact with Octopi-1.5 without requiring a robot. During the demonstration, we will showcase Octopi-1.5 solving tactile inference tasks by leveraging tactile inputs and commonsense knowledge. For example, in a Guessing Game, Octopi-1.5 will identify objects being grasped and respond to follow-up queries about how to handle it (e.g., recommending careful handling for soft fruits). We also plan to demonstrate Octopi-1.5's RAG capabilities by teaching it new items. With live interactions, this demonstration aims to highlight both the progress and limitations of VTLMs such as Octopi-1.5 and to foster further interest in this exciting field. All code for Octopi-1.5 and design files for the TMI gripper will be released as open-source resources.**

I. Introduction

Touch has long been recognized as crucial for robots, particularly in contact-rich tasks or situations with visual occlusion. It enables robots to discern latent object properties, such as determining the softness or deformability of objects. In recent years, robot touch has made significant strides, with notable advances in sensing technologies — for example, in visual tactile sensors like GelSight [1] and Digit [2] — and in the interpretation and application of tactile signals for manipulation. Notably, recent work has integrated tactile perception with Vision-Language Models (VLMs), enabling robots to perform various tactile-related tasks using natural language prompts [3, 4, 5].

Building upon this line of research, we have developed an improved visual-tactile language model (VTLM), which we call Octopi-1.5. Compared to existing VTLMs, Octopi-1.5 features three key enhancements:

- a new tactile encoder trained on a combination of existing Gelsight datasets with an expanded PhysiCLeAR dataset [3],
- the use of the Qwen2-VL 7B base VLM [6], offering enhanced interaction and commonsense knowledge capabilities, the visual modality, and
- experimental retrieval-augmented generation (RAG) [7], which not only boosts performance but potentially allows the model to learn new tactile-object pairings by adding them to its database.

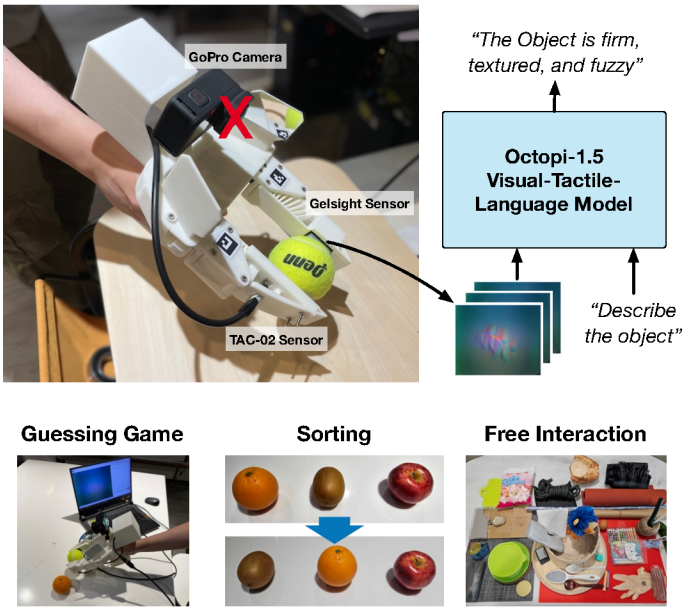


Fig. 1. Octopi-1.5 Demonstrations using a Tactile Manipulation Interface (TMI) gripper. We plan to demonstrate how Octopi-1.5 can be used to describe and make use of tactile sensations in a range of scenarios, e.g., without vision as shown in the top image. We also a (i) **Guessing Game** where Octopi-1.5 is tasked to guess which item is being grasped given only tactile inputs (without vision), (ii) a **Sorting** task where Octopi-1.5 has to sort items according to hardness or fruits according to ripeness, and (iii) a **Free Interaction** setting where users can use visual, tactile, and language modalities together with Octopi-1.5. in a free manner on a variety of objects to explore the strengths and limitations of the system.

At R:SS 2025, we propose to demonstrate Octopi-1.5's capabilities in real-time using a Tactile Manipulation Interface (TMI) (Fig. 1), which is a Universal Manipulation Interface (UMI) [8] modified with tactile sensors. The demo setup will allow users to interact with objects using the TMI and query Octopi-1.5 through natural language and physical interaction. We designed the setup to be portable and not require physical

robot, which eases transportation to R:SS 2025. We envision several interactive scenarios, including a Guessing Game — in which Octopi-1.5 must identify the object being touched from a collection of items — to free interaction where users can interact freely with Octopi and the TMI.

The remainder of this paper details Octopi-1.5 and the demonstration setup. We begin with a brief review of related work in tactile-language models (Sec. II), followed by a description of Octopi-1.5's model architecture, training process, and the TMI (III). Sec. IV outlines the planned demonstrations, supported by preliminary experimental results that validate their feasibility. Finally, we conclude with a summary and a discussion of current limitations (Sec. V).

Our demonstrations aim to showcase advancements in tactile-enabled VLMs and their potential applications in robotics. While Octopi-1.5 and similar models are still under development, current limitations, such as challenges in generalization, remain areas of active research. We hope the demonstration will inspire discussion and further exploration in this field. To support the community, we will release all code for Octopi-1.5 and the TMI gripper design.

II. Background and Related Works

Octopi-1.5 builds upon a wide base of prior work in tactile perception for robotics and multi-modal language models. In the following, we give a brief overview of these areas, focusing on advances in visual-tactile-language models. We refer readers interested in general tactile sensing and perception for robotics to survey articles [9, 10].

Tactile Sensing. Although not as well-developed as vision, there has been significant progress on the development of tactile sensors over the years [11]. Tactile sensors come in a variety of types depending on their sensing modality, e.g., piezoresistive, capacitive, and optical. In this work, we primarily use an optical sensor called the Gelsight Mini [1][1]. Optical tactile sensors provide high-resolution tactile images that capture surface deformations by converting them into visual data [11]. This makes them effective at detecting fine surface features and texture, and their image outputs are easily processed by modern machine learning methods. However, compared to alternative sensor types, optical tactile sensors tend to be larger in size and have lower sampling rates.

Our TMI gripper also incorporates a piezoresistive sensor called the TAC-02[2]. Unlike optical sensors, piezoresistive sensors operate by detecting changes in resistance caused by the deformation of elastic materials under applied forces [11]. The TAC-02 has 64 taxels, which is lower in resolution compared to the Gelsight Mini, but is more compact, has a higher sampling rate (1kHz), and can directly capture static and dynamic pressure.

Visual-Tactile-Language Models. Octopi is closely related to very recent work on multimodal large-language models (MLLMs) [12] that process data from real-world tactile sensors [3, 4, 5] — we refer to these models as Visual-Tactile-Language models (VTLMs). These prior works share similarities in that they involve the collection of tactile, visual, and language modalities into datasets. For example, the TVL Dataset [4] includes over 44,000 paired tactile-visual samples annotated with natural language, while other datasets, such as ObjectFolder [13] and PhysiCLeAR [3], combine high-resolution tactile data with semantic labels for material properties like hardness and roughness. These datasets are used to train tactile encoders, often based on Visual Transformers (ViTs) [14], employing contrastive and regression-based loss functions to align tactile, visual, and language representations. Octopi-1.5 shares these foundational approaches but (i) is trained on a larger dataset compared to its predecessor and (ii) incorporates a retrieval-augmented generation (RAG) module, enabling it to retrieve and utilize similar objects from its database to enhance predictions and allows for on-the-fly learning of new object-tactile pairings.

TABLE I
Octopi-1.5 Training Dataset Statistics

Dataset	Num. Objects	Samples per Object	Num. Tactile Videos
PhysiCLeAR-Plain	100	5–45	2689
PhysiCLeAR-Dotted	68	11–32	1939
Hardness [15]	210	1–133	1860
ObjectFolder-Real [13]	100	30–67	3550

III. System Description: Octopi-1.5 and TMI

In this section, we discuss the main components of our demonstration: Octopi-1.5 and the TMI.

A. Octopi-1.5 VTLM

Octopi-1.5 is a visual-tactile-language model (VTLM). Below, we present the model's architecture, training methodology, and the integration of a simple RAG module.

Octopi-1.5 Model Structure. Fig. 2 A provides a high-level overview of the model architecture, including the GelSight mini tactile encoder. Octopi-1.5 is based on the QWEN2-VL 7B open-source vision-language model (VLM) [6], whereas the previous Octopi version utilized Llama [16]. Encoders are used to transform raw inputs into tokens, which are subsequently processed by the VLM transformer.

A key improvement in Octopi-1.5 is in the tactile encoder, which translates tactile frames from a GelSight mini sensor into tokens for the VLM. Specifically, the tactile encoder is a fine-tuned CLIP [17] module augmented with a projection layer. To optimize computational efficiency, we process only "salient" frames, which are selected using a heuristic that identifies the top 10 frames with the largest differences compared to their preceding frames.

Octopi-1.5 Training. Octopi-1.5 was trained on an expanded PhysiCLeAR dataset (including both marker and markerless GelSight pads), as well as the hardness [15] and ObjectFolder datasets [13]. Table I provides a summary of the dataset statistics. Model training was conducted in two stages: first, we

[1]Gelsight Mini Product Sheet: https://www.gelsight.com/wp-content/uploads/productsheet/Mini/GS_Mini_Product_Sheet_10.07.24.pdf

[2]https://www.tacniq.ai/tac-02-robotic-finger-dev-kit

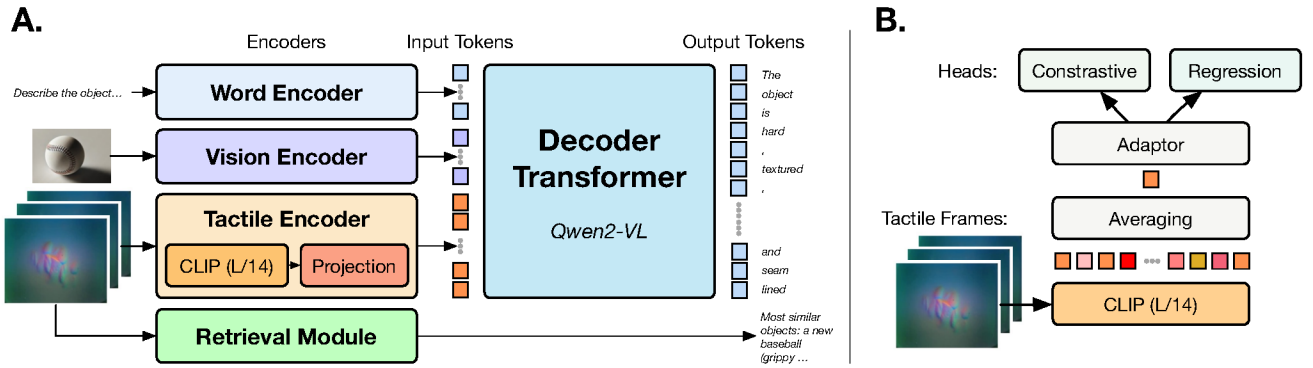


Fig. 2. (**A**) Octopi-1.5 model architecture. Octopi-1.5 is a fine-tuned Qwen2-VL 7B multimodal VLM with a tactile encoder and retrieval module. (**B**) The CLIP module of the tactile encoder is pre-trained before finetuning using contrastive and regression losses on PhysiCLeAR, Hardness, and ObjectFolder datasets. Please see main text for details.

TABLE II
ANNOTATOR SCORING GUIDANCE FOR HARDNESS AND ROUGHNESS.

Property	Score	Description	Example Object
Hardness	0	Easily compressible when pressed with little force	Cotton wool
	5	Compressible upon pressing with moderate force	Foam mattress
	10	Incompressible with high human pressing force	Rock
Roughness	0	Minimal feeling of friction upon finger sliding	Ice
	5	Slight feeling of friction upon finger sliding	Jeans
	10	Significant feeling of friction upon finger sliding	Toothbrush bristles

Training Task Example: Describe and Rank

USER

```
Describe the objects in the following
tactile videos and rank them in
decreasing hardness and roughness.

Object 1
Part 1.1: [tactile frames]
Part 1.2: [tactile frames]
```

OCTOPI-1.5

```
Object 1
Part 1.1: [descriptions]
Part 1.2: [descriptions]

Object parts ranked in decreasing
hardness: 1.1, 1.2
Object parts ranked in decreasing
roughness: 1.2, 1.1
```

Fig. 3. Training Task Example. Octopi-1.5 is trained to describe and rank different parts of an object (in addition to different objects).

RAG Description Example

USER

```
Describe the object in the following
tactile video(s).
```

OCTOPI-1.5

```
Object 1: [Tactile Frames]

Object 1: textured, grippy,
seam-lined, hard.
Most similar objects (in order
of decreasing similarity): a new
baseball's seams (grippy, hard,
seam-lined, textured); ...
```

Fig. 4. The outcome of a tactile description by Octopi-1.5 augmented with RAG. The additional information added by the RAG module is shown in blue.

trained the CLIP module within the tactile encoder, followed by end-to-end fine-tuning of the VTLM. This staged approach resulted in more stable training outcomes.

The CLIP module was trained using both regression and contrastive losses, as illustrated in Fig. 2 B. In the current implementation, the network is trained to predict human-annotated hardness and roughness scores for objects. Annotators followed the guidance shown in Table II for labeling. For the contrastive loss, the network distinguishes tactile inputs from the same object/part (positive class) versus those from other objects/parts (negative class). To ensure consistency, both positive and negative classes were sampled using the same type of GelSight silicon pad (marker or markerless). We trained the CLIP module for 30 epochs, selecting the best encoder based on validation loss using a set of six unseen objects.

In the second phase, the CLIP module is frozen, and the projection layers are trained alongside the decoder. The decoder training employs Visual Prompt Tuning (VPT) [18] on description and ranking tasks as illustrated in Fig. 3. Unlike its

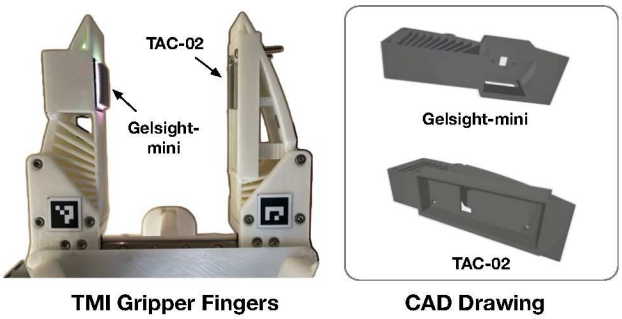


Fig. 5. TMI fingers and CAD drawings with compartments for inserting the Gelsight-mini or the TAC-02 sensors.

predecessor, Octopi-1.5 is explicitly designed to handle objects comprising multiple parts, such as a hairbrush with distinct handle and bristle components.

RAG-Modified Descriptions. We are currently experimenting with the decomposition of a VTLM into two components: one dedicated to processing/inference and another serving as a "memory" of previously seen items. As a step towards this goal, Octopi-1.5 incorporates a simple RAG scheme to enhance its tactile descriptions by augmenting them with textual information from similar objects, thereby aiding downstream tasks. Specifically, the tactile descriptions of objects are supplemented with their labels and the tactile descriptions of similar items, creating a more informative representation (example in Fig. 4).

To generate these augmented descriptions, we follow a straightforward process:

- For a new set of salient tactile images, compute their average embedding using the tactile encoder.
- Perform a cosine-similarity search over an existing dataset of tactile embeddings to identify the top-5 objects with the most similar averaged embeddings.
- Aggregate the unique objects and rank them by the number of retrieved samples to form a prioritized list of matches.

While straightforward, this RAG scheme improved Octopi's performance in our proposed demonstration tasks (see Section IV). Future work could explore more advanced RAG setups, such as retrieving and incorporating tactile embeddings directly; implementing such improvements will likely require retraining the VTLM to optimally utilize the retrieved information.

B. Tactile Manipulation Interface

To facilitate the demonstration, we propose using a modified Universal Manipulation Interface (UMI) [8]. Our Tactile Manipulation Interface (TMI) features fingers equipped with tactile sensors (Fig. 5). Specifically, one finger is mounted with a GelSight Mini sensor, while the other houses a TAC-02 piezoresistive tactile sensor. The GelSight sensor provides high-resolution tactile imaging, whereas the TAC-02 sensor [3] directly captures pressure readings.

[3]https://www.tacniq.ai/tac-02-robotic-finger-dev-kit

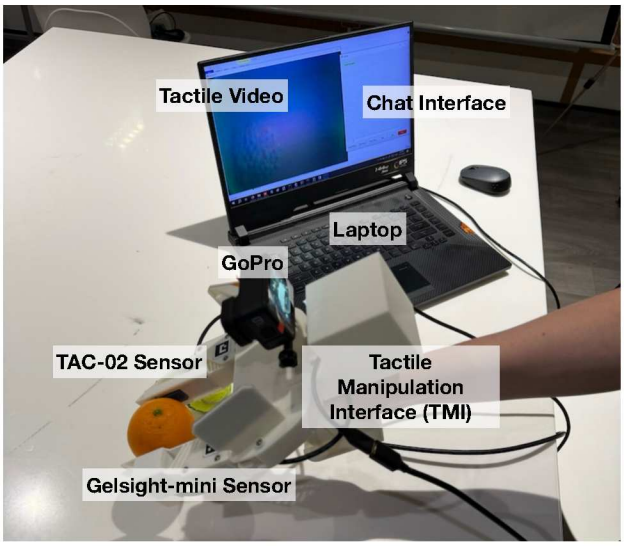


Fig. 6. Our demonstration system is highly portable, consisting primarily of the TMI and a laptop. The laptop is equipped with the necessary software to interface with Octopi-1.5, which runs on a remote workstation or server. A stable internet connection is required, with the option to use a mobile hotspot if needed. The setup can be fully assembled in under 15 minutes. We will provide a variety of items for grasping and participants can use their own objects.

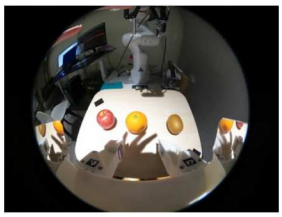
Our demonstration will primarily utilize the GelSight Mini sensor, as it has significantly more data available for model training. Nevertheless, the TMI provides a proof-of-concept integration of multiple sensors, which we believe will spur discussion on integrating heterogeneous tactile modalities. We also hope to show participants how manipulation data can be collected using the TMI. We plan to engage in broader discussions surrounding TMI's benefits — such as ease of use and versatility — and limitations including sensor wear and tear, task constraints, difficulties in annotation, and the variability across different tactile sensors.

IV. Planned Demonstrations

We propose a series of demonstrations to showcase Octopi-1.5's capabilities in aiding tactile-related tasks. These demonstrations will utilize the TMI, connected to a laptop with an internet connection to a remote workstation/server at NUS running Octopi-1.5 (Fig. 6). If available, we would prefer to use a large-screen TV or projector to display presentation slides and Octopi-1.5's feedback.

Our planned format is straightforward: we will begin with a brief overview of Octopi-1.5 and the TMI, followed by an introduction to the demonstration system. Currently, text interaction is facilitated through a simple chat interface, though we are exploring the feasibility of integrating speech-to-text using Whisper [19]. However, we found Whisper can be unreliable in noisy environments with multiple speaking participants, so it may not be included in the final setup.

We will then demo a tactile-based description task (Fig. 1), followed by two structured demonstrations: the Guessing



Guessing Game Example (through Simplified Interface)

USER
Take a picture of the scene.
[System takes a picture using the GoPro; see image above].
What fruits do you see in the image?

OCTOPI-1.5
Object 1: round, red, apple.
Object 2: round, orange, orange.
Object 3: oval, yellow, kiwi.

USER
Touch an item on the table.

OCTOPI-1.5
Please use the gripper to grab the item.
[User grasps apple using the TMI (no visual images taken)].
Finished collecting tactile data.

USER
Describe what you just touched.

OCTOPI-1.5
Object 4: hard, firm, glossy, smooth.

USER
Which fruit is it likely to be?

OCTOPI-1.5
To determine which option the object is likely to be, we need to consider the surface texture descriptions that help to distinguish between the given options. Let's analyze each option:
A) Apple: Apples typically have a smooth, slightly waxy texture on the skin.
B) Orange: Oranges have a rough, bumpy texture on the skin.
C) Kiwi: Kiwis have a fuzzy, textured skin.
Now, let's give a succinct case for each option ...
Answer: (A) Apple is the most likely option because it has a smooth, slightly waxy texture on the skin, which is characteristic of apples.

Fig. 7. User Interaction Example for the Guessing Game.

TABLE III
AVERAGE ACCURACIES (%) ON THE GUESSING GAME.

Method	RAG	Balls	Fruits	Unseen	Unseen (teaching)
Encoder-1.5	–	80.00	100.00	N/A	89.02
Octopi-1 (7B)	No	44.00	42.31	43.90	N/A
+-Octopi-1 (13B)	No	48.00	34.62	53.66	N/A
Octopi-1.5 (8B)	No	56.00	57.69	41.46	N/A
Octopi-1.5 (8B)	Yes	96.00	100.00	73.17	95.12

Game and Sorting Task. Finally, users will have the opportunity to interact more freely with the system to explore its capabilities and limitations.

A. *Guessing Game*

The guessing game is designed to demonstrate how tactile information can be combined with commonsense knowledge to make inferences:

- We will provide octopi with a set of objects (described via language or visual input)
- Users will select an object and provide tactile inputs by grasping the object with the TMI (no visual input provided during grasping)
- Octopi-1.5 will infer which object corresponds to the tactile inputs.
- Users can interact with the system by prompting Octopi-1.5 to guess again if the initial guess was incorrect or asking follow-up questions (e.g., how hard the object should be grasped).

To simplify the interaction, the demo will make use of a chat interface that "hides" the more complex prompts used to elicit proper behavior from Octopi-1.5; see Fig. 7 for an example of the interaction where the user grasps an apple. After the interaction, we will also show the exact prompts that are used "behind-the-scenes".

Preliminary tests support the feasibility of this demonstration. We experimented with three categories of objects: balls (baseball, plush ball, tennis ball) and fruits (apple, partially ripe kiwi, orange) and completely unseen objects (hairbrush bristles, hairbrush handle, microfiber cloth). Accuracy scores are summarized in Table III, with averages computed across multiple samples from the same object (25 samples for balls, 26 samples for fruits, and 41 unseen samples).

As a non-VLM baseline, the Encoder-1.5 model uses cosine similarity of tactile embeddings in the training set. The encoder-only baseline achieves good results on familiar objects but offers no out-of-the-box handling of unseen items without database updates.

Octopi-1.5 shows higher overall accuracy compared to the older Octopi-1. Simple RAG retrieval improves performance and "teaching" further boosts results by adding new item samples to the dataset on the fly. With RAG enabled, accuracy scores for Octopi-1.5 are high (96-100%) for the seen objects (Balls, Fruits) with good performance (73%) for unseen objects. Once teaching is enabled, performance rises to (95%). Along with these items, we hope the demonstration will also

Sorting Example with RAG (through Simplified Interface)

USER
Touch this item.

OCTOPI-1.5
Please use the gripper to grab the item.
[User grasps apple using the TMI (no visual images taken)].
Finished collecting tactile data.

[... Repeated 3 times...]

USER
Describe and rank the objects by their hardness.

OCTOPI-1.5
Object 1: smooth, hard, slightly waxy.
Most similar objects (in order of decreasing similarity): an unpeeled, ripe apple (firm, glossy, slightly waxy, smooth); the back of a TV remote (hard, slightly grainy, smooth);
Object 2: textured, slightly soft, rounded.
Most similar objects (in order of decreasing similarity): an unpeeled, ripe orange (rounded, slightly soft, textured);
Object 3: slightly rough, fuzzy, squishy, soft.
Most similar objects (in order of decreasing similarity): an unpeeled, overripe kiwi (fuzzy, slightly rough, soft, squishy); an unpeeled, spoiled orange (rounded, slightly soft, textured);
Object parts ranked in decreasing hardness: 1 > 2 > 3

Fig. 8. Example User Interaction for the Sorting Task.

show Octopi-1.5's ability to recognize multiple parts of the same object.

B. Sorting

The sorting task is designed to demonstrate Octopi-1.5's ability to distinguish tactile properties of objects. This capability has potential use in applications such as sorting/classifying fruits or other items based-on their material properties.

Given a set of user-chosen items and their corresponding tactile inputs, Octopi-1.5 will be tasked with sorting the items according to properties such as hardness (Fig. 8). Current experiments show that Octopi is able to sort the balls (100%) and fruits (93.18%) but faces difficulty on the unseen objects (43.33%). We plan to highlight failure cases during the demonstration to show current limitations. This will provide an opportunity to discuss various challenges from annotation (e.g., the relative nature of properties when described using language) to sensing and model training.

C. Free Interaction

The free interaction segment will allow participants to explore Octopi-1.5's capabilities in a more open-ended manner. Users will be able to:

- Provide both visual and tactile inputs to Octopi-1.5.
- Engage in an unconstrained chat to ask questions and explore Octopi-1.5's responses.
- Test experimental features, including teaching Octopi-1.5 new items.

We hope this can help reveal other limitations of the current system and highlight interesting ways users may use Octopi and the TMI.

V. Conclusion, Limitations, and Future Work

In summary, we have outlined a series of demonstrations to illustrate the capabilities of Octopi-1.5 and the TMI gripper. We believe the proposed setup is feasible and will be of interest to R:SS attendees, particularly roboticists working on foundation models and tactile robotics. Octopi-1.5 and the TMI showcase interesting features, such as the design of the tactile encoder and the use of RAG, as well as how tactile sensors can be integrated into handheld or robotic grippers. The design and development of such systems remain open questions, and we aim to foster discussions ranging from technical advancements to broader questions about system use and design philosophy.

Limitations. Octopi-1.5 is a work-in-progress toward language-conditioned models that enable robots to better utilize vision and touch modalities. While preliminary results are promising, significant limitations remains. General improvements are needed to improve performance on some tasks (e.g., sorting by roughness remains error-prone). This might be addressed by using larger base LLMs, though this requires substantial computational resources. Alternatively, we are exploring more structured decompositions and architectures that leverage prior knowledge about tactile information to mitigate compute/data limitations. We also plan to investigate LTLMs trained to use other sensors such as the TAC-02.

The current RAG module is based solely on text retrieval. Incorporating tactile sample retrieval may further enhance the system, especially for tactile information that is difficult to encode in text. However, more research is needed to isolate what information is best to retrieve in computationally low-cost manner.

Finally, Octopi-1.5 only outputs text tokens. Directly linking Octopi-1.5 to manipulation tasks is a key area of future work. We are in the process of training a Vision-Tactile-Language-Action (VTLA) model capable of directly outputting robot actions for manipulation tasks, but this is still under development. We hope to discuss these limitations and future directions at R:SS 2025.

ACKNOWLEDGEMENTS

We gratefully acknowledge Schaeffler Pte. Ltd. for their support and special thanks to Geet Jethwani, Kratika Garg, and Han Boon Siew for their insightful input and assistance throughout the project. This research is supported by the National Research Foundation, Singapore under its Medium Sized Center for Advanced Robotics Technology Innovation.

REFERENCES

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

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

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

[4] Letian Fu, Gaurav Datta, Huang Huang, William Chung-Ho Panitch, Jaimyn Drake, Joseph Ortiz, Mustafa Mukadam, Mike Lambeta, Roberto Calandra, and Ken Goldberg. A touch, vision, and language dataset for multimodal alignment. In *International Conference on Machine Learning*, 2024.

[5] Fengyu Yang, Chao Feng, Ziyang Chen, Hyoungseob Park, Daniel Wang, Yiming Dou, Ziyao Zeng, Xien Chen, Rit Gangopadhyay, Andrew Owens, et al. 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, 2024.

[6] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[7] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.

[8] Cheng Chi, Zhenjia Xu, Chuer Pan, Eric Cousineau, Benjamin Burchfiel, Siyuan Feng, Russ Tedrake, and Shuran Song. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. In *Proceedings of Robotics: Science and Systems (RSS)*, 2024.

[9] Shan Luo, Joao Bimbo, Ravinder Dahiya, and Hongbin Liu. Robotic tactile perception of object properties: A review. *Mechatronics*, 48:54–67, 2017.

[10] Peter Roberts, Mason Zadan, and Carmel Majidi. Soft tactile sensing skins for robotics. *Current Robotics Reports*, 2:343–354, 2021.

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

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

[13] Ruohan Gao, Yen-Yu Chang, Shivani Mall, Li Fei-Fei, and Jiajun Wu. Objectfolder: A dataset of objects with implicit visual, auditory, and tactile representations. In *Conference on Robot Learning (CoRL)*, 2021.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[15] 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. IEEE, 2016.

[16] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.

[17] 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.

[18] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022.

[19] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023.