2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Multi-Modal Representation Learning with Tactile Data

Hyung-gun Chi[†,1], Jose Barreiros[†,2], Jean Mercat[2], Karthik Ramani[1], Thomas Kollar[2]

***Abstract*— Advancements in embodied language models like PALM-E and RT-2 have significantly enhanced language-conditioned robotic manipulation. However, these advances remain predominantly focused on vision and language, often overlooking the pivotal role of tactile feedback which is advantageous in contact-rich interactions. Our research introduces a novel approach that synergizes tactile information with vision and language. We present the Multi-Modal Wand (MMWand) dataset enriched with linguistic descriptions and tactile data. By integrating tactile feedback, we aim to bridge the divide between human linguistic understanding and robotic sensory interpretation. Our multi-modal representation model is trained on these datasets by employing the multi-modal embedding alignment principle from ImageBind which has shown promising results, emphasizing the potential of tactile data in robotic applications. The validation of our approach in downstream robotics tasks, such as texture-based object classification, cross-modality retrieval, and the dense reward function for visuomotor control, attests to its effectiveness. Our contributions underscore the importance of tactile feedback in multi-modal robotic learning and its potential to enhance robotic tasks. The MMWand dataset is publicly available at https://hyung-gun.me/mmwand/.**

I. Introduction

The advent of language models in robotics has enabled robots to achieve a primitive human-like contextual understanding. Pioneering contributions such as PALM-E [1] and RT-2 [2] have driven the development of embodied language models, enhancing robotic manipulation and planning capabilities. Similarly, some of the recent works [3]–[5] have excelled in representation learning for robotics through language. However, while robots are increasingly adept at understanding human language and performing related tasks, much of the current research emphasizes vision and language, often sidelining the critical role of other critical modalities like tactile sensing. Our research addresses this oversight, focusing on a method that integrates tactile information with vision and language. A significant gap in the available data is the lack of language annotations about tactile features. They are essential for merging tactile data with language models. This is why we propose a new dataset with tactile image and text annotations that describe the objects and their tactile properties.

Grasping objects reliably necessitates tactile feedback. While humans naturally rely on sensory feedback from their skin, robotic grippers have incorporated tactile feedback to improve environmental perception as in [6], [7]. Grippers with tactile sensors can discern not only object textures but also the quality of the grasp [8], [9]. These sensors are crucial, offering robots essential perceptual insights. In this work, we aim to merge tactile input with other modalities as illustrated in Fig. 1 by overcoming the challenges of tactile images that are discussed in the next paragraph. This integration seeks to strengthen the connection between perceptual data and linguistic representation, enhancing robotic perception and narrowing the gap between human linguistic comprehension and robotic sensory interpretation. By utilizing pre-trained multi-modal representation models, our goal is to enhance robotic tasks like texture-based object classification, cross-modality retrieval, where one modality is retrieved based on another, and the dense reward function for visuomotor control.

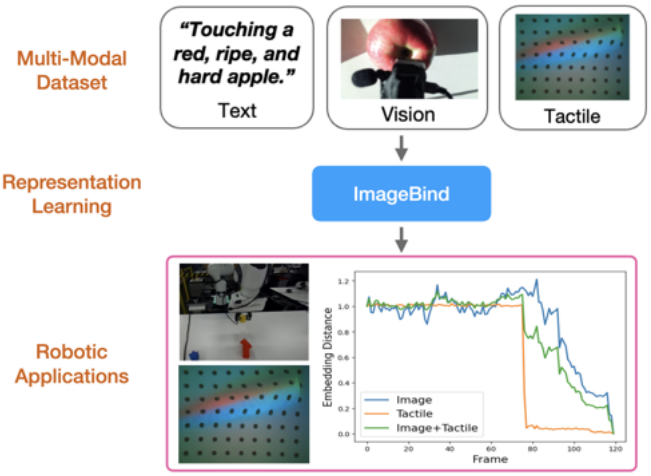

Fig. 1: We integrate tactile data with vision and language by multi-modal representation learning for robotic applications.

Multi-modal representation learning approaches [10], [11] aim to combine data from various modalities into a unified framework. Modalities such as inertial measurement unit (IMU) trajectories, depth, point cloud, video, audio, and text have been explored but tactile data remains largely untapped.

Incorporating tactile sensors into multi-modal representation learning poses unique challenges. There is a noticeable scarcity of tactile datasets, with only a handful publicly available, that we describe in section II-A. These datasets often originate from controlled lab environments, limiting their diversity. The delicacy of tactile sensors can also compromise data quality [12]. GelSight [13], a commonly used tactile sensor, employs a gel that conforms to object features when in contact resulting in precise texture observations. However, due to the gel's durability issues, the tactile images collected are prone to include cracks or deformities due to missing

† Equal contribution.
[1] Purdue University, {chi45, ramani}@purdue.edu
[2] Toyota Research Institute, {jose.barreiros, jean.mercat, thomas.kollar}@tri.global

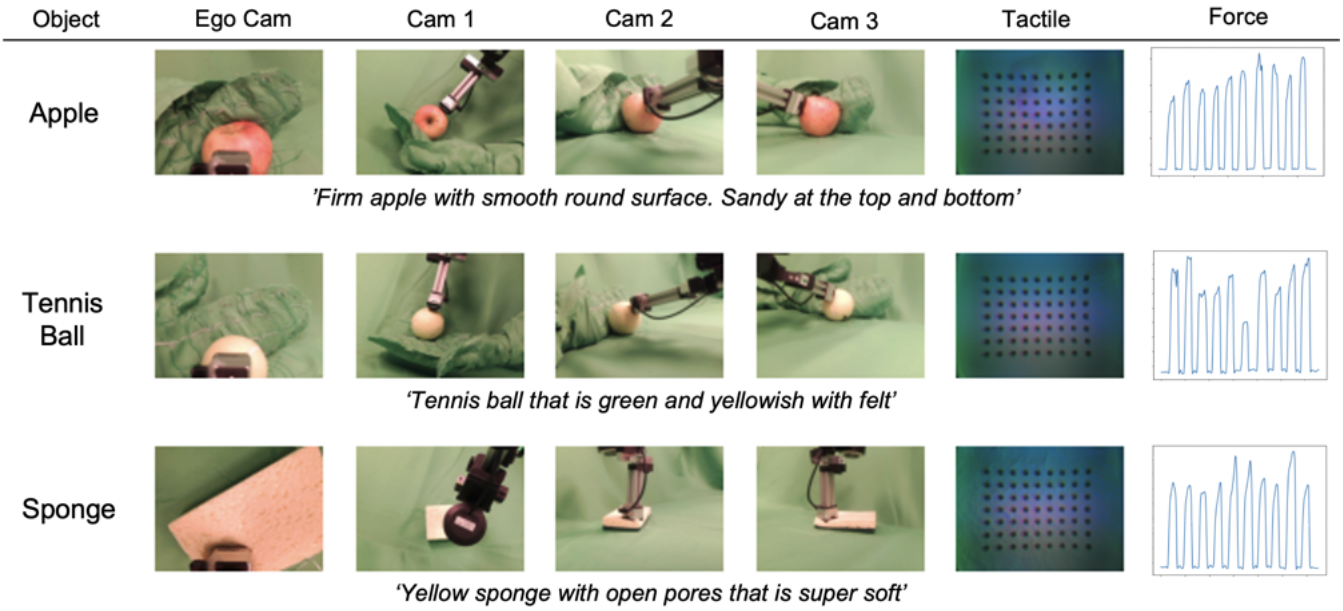

Fig. 2: **Samples of the MMWand dataset.** The dataset comprises multi-view visual images, tactile images, and force values. Crucially, it offers language descriptions, which can facilitate the use of large language models for robotics tasks.

pieces of the gel which came to our attention when analyzing currently available datasets. Furthermore, tactile images often represent limited areas of visual images, demanding accurate alignment. For example, the tactile sensors only include the texture of the local region of the object while the image of the object shows the entire part of the object.

Addressing the constraints of current tactile datasets, we present two strategies for tactile-based multi-modal representation learning. Firstly, we introduce the Multi-Modal Wand (MMWand) dataset, a unique collection enriched with linguistic descriptions and tactile data. This dataset is sourced from a specially designed wand equipped with various sensors: force, tactile, and vision. A key additional feature is our linguistic description of object tactile feeling which was acquired by human labelers that interacted with the same objects as the wand.

We adopt the embedding alignment principle from ImageBind [10] and employ contrastive loss to learn a representation of the touch modality that complements the perceptions from other modalities. Our experimental results highlight that the learned representations offer advantages for robotic tasks.

II. Related Works

A. Datasets with Tactile Modality

Tactile input plays a crucial role in enabling robots to apprehend objects. Several studies have introduced datasets [12], [14]–[20] encompassing tactile information. 'Touch and Go' [12] presents videos capturing tactile interactions with a diverse range of outdoor environments and objects. 'VisGel' [16] gathers tactile data through sensors embedded in robot arms. Among the more recent datasets, 'Object Folder' [17]–[19] offers an exhaustive multisensory perspective, integrating both sound and 3D mesh data. These datasets document tactile interactions with multiple objects, amounting to a large number of touches. Distinct from these datasets, MMWand provides language annotations along with other modalities, which are essential for multi-modal representation learning aiming to align contextual and tactile.

B. Multi-Modal Representation Learning

Pretrained CLIP models [21] are commonly used to instruct other models because of their potent representations combining images and text [22]–[24]. ImageBind [10] utilizes a pretrained CLIP model to align embeddings from diverse modalities into the CLIP embedding space. This facilitates the emergent alignment of previously unseen modality pairs, enabling data retrieval between different modalities without the need for training on paired samples. Meanwhile, meta-transformer [11] introduces a modality-shared encoder for multi-modal learning, which minimizes the parameter count required for various modalities. In our application, we propose to modify the ImageBind framework in order to incorporate tactile input into the unified representation space.

III. Multi-Modal Wand Dataset

We introduce the Multi-Modal Wand (MMWand) dataset which is a collection of a rich array of high-quality multi-modal data (See Fig. 2). Multi-modal representation learning is needed to compensate for the missing information from one modality with the perception from another. To meaningfully align the different modalities, some of the information they contain should be overlapping. However, the image annotations available in open datasets, [12], [14]–[17], are mostly descriptive of the image itself but rarely contain information about its content such as the texture of objects. This gap has hindered the full potential of large-language models [25]–[27] for enhanced representation learning. On this aspect, MMWand stands out distinctly. It includes text

TABLE I: **Comparison with Other Tactile Datasets.** MMWand is the only tactile dataset that includes language descriptions.

Dataset	# Object	# Touch	Data Source	Object Label	Multi-View	Language Annotation	Other Modalities
The Feeling of Success [14]	106	9.3k	Robot	-	-	-	Vision, Tactile
More Than a Feeling [15]	65	6.5k	Robot	-	-	-	Vision, Tactile
VisGel [16]	195	12k	Robot	-	-	-	Vision, Tactile
Touch and Go [12]	3,971	13.9k	Human	✓(Material)	-	-	Vision, Tactile
ObjectFolder [17]	1,100	50k	Synthetic + Real Object	✓	-	-	Vision, Tactile, Sound, Mesh
MMWand (Ours)	102	5.2k	Human	✓	✓**(4 Views)**	✓	Vision, Tactile, Force

Fig. 3: **Backgrounds Synthesized for Data Augmentation**. The first image displays the original with a green screen.

'Firm apple with smooth round surface. Sandy at the top and bottom'
'A small mostly yellow firm waxy apple that has a little give to it'
'A red and yellow apple that feels smooth'
'An apple with a waxy surface is hard and squishy in some places'
'A round red smooth firm apple'

Fig. 4: **Examples of the descriptions for the apple.**

labels that not only describe the images or the object appearances but also the object textures. This particularity bridges the gap between language and touch.

Inevitably, when considering manipulation tasks, the very objects that are touched are occluded by the end effector. This limits the visual information available for training. To overcome this challenge, MMWand includes multiple views cameras and a full video allowing the user to refer to different image frames where the object of interest is not occluded. Our setup includes three third-person view cameras, in addition to the camera on the MultiWand itself.

Most datasets are restricted by static lab environments that do not reflect the diversity of contexts where robots aspire to operate. Drawing inspiration from the FreiHand [28], our data is captured in a greenscreen setup shown in Fig. 5. This allows the user to easily synthesize any background which can significantly enhance the diversity of visual data as the examples in Fig. 3.

Table I summarizes the comparison of MMWand with other datasets. MMWand is the only dataset offering a language description of objects while simultaneously encompassing high-quality touch acquisitions and images. In addition, it includes acquisition from different viewpoints and force measurements. These features are important assets for multi-modal representation learning and are needed to improve contact-rich applications in robotic manipulation. The MMWand dataset is publicly available at https://hyung-gun.me/mmwand/.

A. Multi-Modal Wand Design

The hardware setup of the multi-modal wand is depicted in Fig. 5 *bottom*. Tactile information is gleaned using a GelSight Mini [13], a widely used visuo-tactile sensor that

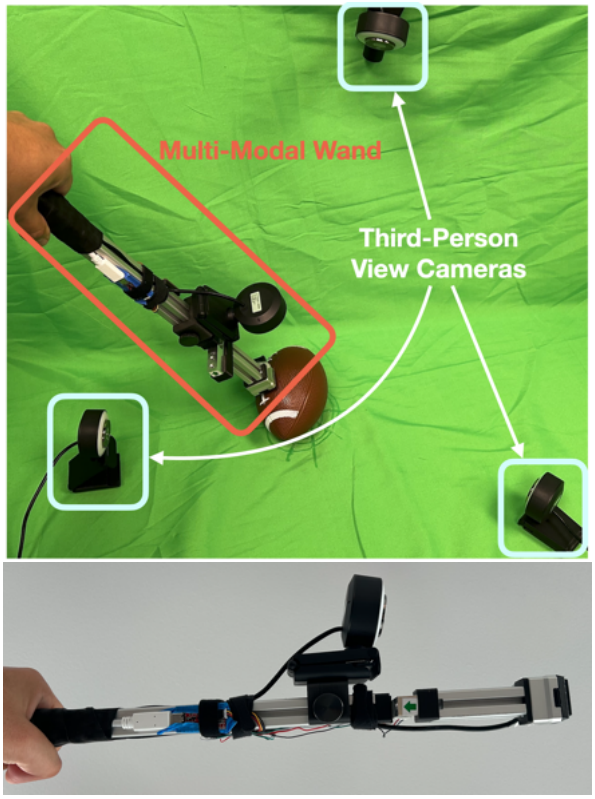

Fig. 5: **Overview of Data Collection Setup.** For data collection, we installed three third-person view cameras. The MMWand is outfitted with tactile and force sensors, as well as an ego-view camera.

uses a camera looking at a deformable gel illuminated with structured light. A strain gauge load cell (5kg, 4-wire, Adafruit, ID: 4541) is integrated to record the total normal force exerted when the MMWand pushes against an object. Additionally, the wand includes a point-of-view RGB camera (Vitade webcam with an LED light ring, Full HD 1080P) located over the wand tip. The experimental bench is covered with a chroma-key green screen and includes three additional cameras that provide overhead and side views of the interaction.

B. Data Collection & Annotation

Our dataset contains a collection of 102 diverse common objects listed in Table V, each chosen for its unique stiffness and texture. A touch data point was collected by following the procedure.

1) Place the object in the middle of the scene (for round objects, we held them manually using a glove made from the aforementioned green screen material)
2) Poke the object with the MMWand in a place and with force chosen randomly by the experiment facilitator, we tried to follow a uniform distribution for the contact surface and force sampled within three ranges - low: [0, 5] N, medium: (5, 10] N, and high: (10, 20] N.
3) Verify the integrity of the tactile sensor gel and replace it when compromised.
4) Move the object and repeat from 2).

What mostly sets MMWand apart is our language annotation of tactile perception. Every object in the dataset is accompanied by 3 to 5 distinct natural language descriptions, each detailing the object's texture and shape, ensuring a comprehensive understanding of each item. We asked 10 participants (6 male and 4 female-identified, 30-42 y.o.) to touch each of the objects and provide a text description with a focus on the tactile feeling. The task was to touch the object and describe it, focusing on the tactile sensations. Examples of the language annotations are provided in Fig. 4. We ensure an overlap of information between various modalities, maintaining their complementary nature. Hence, the text includes tactile details, but it isn't necessary to exactly mirror the touch sensor's perception.

IV. Multi-Modal Representation Learning

We trained a multi-modal representation learning model on each individual dataset: MMWand dataset and the separately augmented existing tactile dataset. Our approach is influenced by Imagebind [10], which integrates embeddings from various modalities into a cohesive image embedding space. In line with Imagebind, while the pre-trained CLIP image and text encoders remain static, the tactile encoder undergoes training using the InfoNCE loss [29]. Recognized for its efficacy in contrastive learning, the InfoNCE loss ensures the convergence of similar data points and the divergence of dissimilar ones. The InfoNCE loss, calculated from the embeddings of modalities M_1 and M_2, is defined as follows:

$$L_{\text{InfoNCE}}^{(M_1 \leftrightarrow M_2)} = -\log \frac{\exp\left\{s(z_i^{(M_1)}, z_i^{(M_2)})/\tau\right\}}{\sum_{j=1}^{B} \exp\left\{s(z_i^{(M_1)}, z_j^{(M_2)})/\tau\right\}}. \quad (1)$$

Here, B represents the mini-batch size, i is the index of the positive pair, z_M denotes the embeddings of modality M, τ is a scalar temperature parameter that affects the smoothness of the softmax distribution, and $s(\cdot,\cdot)$ is the similarity function which in our case is the cosine similarity.

Inspired by the architecture choices from Imagebind, we employed ViT-H as a tactile encoder [30]. The objective of this encoder is to align the tactile embeddings with the pre-trained CLIP [21] embeddings. We train it to minimize the sum of two InfoNCE losses. One for the image-tactile pair $L_{\text{InfoNCE}}^{(tact \leftrightarrow img)}$ and the second for the tactile-text pair $L_{\text{InfoNCE}}^{(tact \leftrightarrow txt)}$. The total loss for our multi-modal representation learning is as follows:

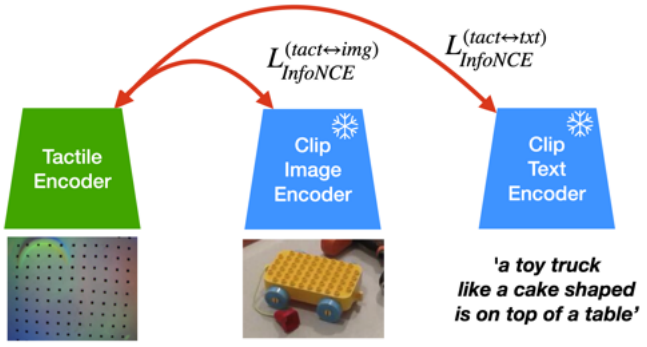

Fig. 6: **Multi-Modal Representation Learning Scheme**. We align the tactile embedding with image and text embeddings from the CLIP image and text encoders. The CLIP encoders are pretrained and remain frozen during training.

$$L_{\text{total}} = L_{\text{InfoNCE}}^{(tact \leftrightarrow img)} + \lambda L_{\text{InfoNCE}}^{(tact \leftrightarrow txt)}. \quad (2)$$

Here, $tact$, img, and txt represent the tactile, image, and text modalities, respectively, while λ is a scalar coefficient.

V. Experiments

To highlight the advantages of aligned tactile embedding, we conducted experiments focusing on object classification and cross-modality retrieval. We further utilize tactile embedding for robotics applications.

Implementation Details. We adopted the open-source implementation of ImageBind [10] to train our model. All experiments were executed on 8 NVIDIA V100 GPUs. We trained our model for 100 epochs using the Adam optimizer [31], setting the learning rate to $1e-4$ and weight decay to 0.05. Also, we employed a batch size of $B = 64$, a temperature parameter of $\tau = 0.2$, and loss coefficient $\lambda = 1$.

A. Datasets

MMWand. We designate approximately 80% of the MMWand dataset for training, which amounts to 4,134 samples, and reserve the remaining 20%, or 1,048 samples, for testing. Throughout the training process, the descriptions linked to each object are randomly shuffled and matched with the scenes where they appear.

VisGel. We further employed the VisGel dataset [16] to evaluate the generalizability of our approach. Given that the dataset features multiple objects in each scene without accompanying language descriptions, we annotated the VisGel dataset using frame-matching techniques and the BlipV2 model for scene description. A VisGel annotation sample is provided in Fig. 7. Detailed information on the VisGel dataset annotation framework is available in Appendix A. For the data split, we adhered to the division used in the touch2vision task as detailed in [16]. Data with a low instance detection confidence value and a low IOU between the object and sensor regions were excluded. Additionally, we omitted data with minimal tactile image deformation.

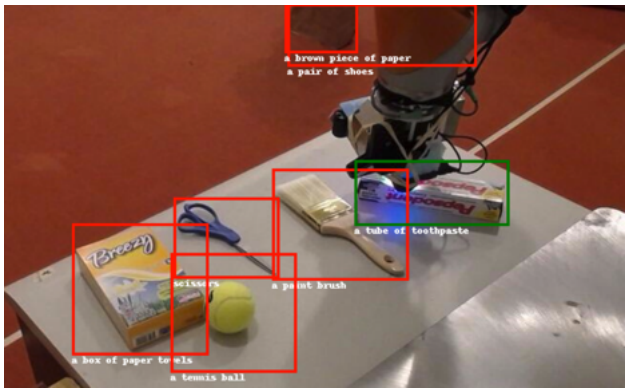

Fig. 7: **Visgel Dataset Sample**. Objects identified as positive are highlighted within green bounding boxes, while negative samples are delineated by red bounding boxes. Additionally, language descriptions for each object are provided below their respective bounding boxes.

Consequently, our final training set consists of 7,772 samples while the testing set includes 889 samples for seen objects and 850 for unseen objects.

B. Multi-Modal Object Classification

In Table II, we evaluated the classification performance using different modality inputs (image, tactile, and image+tactile) on the MMWand with well-known image classifiers [30], [32], [33]. The classifiers, initially pretrained on ImageNet [34], were further fine-tuned for our study. We then compared these outcomes with the results from ImageBind, which was trained as described in Sec. IV. Notably, we used a linear classifier to evaluate the embeddings produced by ImageBind. We note that the classification accuracy of ImageBind with image input is quite low, primarily because it isn't trained for classification and its image encoder remains static during training. Conversely, the performance with tactile input surpasses that of image input, indicating that it effectively learns representations, even with a suboptimal image encoder for the MMWand dataset.

C. Object Classification on VisGel Dataset

To assess the generalizability of our model trained on the MMWand dataset, we conducted object classification on the VisGel dataset. In Table III, we present the tactile-driven classification accuracy on VisGel testset. Given that VisGel doesn't provide labels for objects, we determine classification accuracy by comparing the similarity between an object's tactile or image representation and potential text candidates. These candidates are derived from the captions associated with object instances in the corresponding image. We compare the performance of models trained on MMWand and VisGel. Notably, since our training approach kept the image encoder frozen, the image-based classification performance remains consistent, irrespective of the training dataset. We note that the model trained on MMWand exhibits impressive performance, especially with unseen objects, even though it was trained on a different dataset. The performance disparity with the model on VisGel could be attributed to variations in tactile images between the two datasets.

TABLE II: Classification accuracy (%) across various modality inputs on MMWand testset.

Model	Modality		
	Image	Tactile	Image + Tactile
VGG [33]	94.94	**79.48**	96.66
ResNet [32]	**96.28**	**79.48**	97.70
ViT [30]	95.51	76.90	**97.99**
ImageBind [10]	3.24	46.94	47.22

TABLE III: Classification accuracy (%) across various modality inputs on VisGel testset.

Train Dataset	Modality	Seen		Unseen	
		Top-1	Top-5	Top-1	Top-5
-	Image	53.95	93.49	9.97	48.52
VisGel	Tactile	14.58	65.20	13.83	60.54
MMWand	Tactile	7.38	50.08	11.79	51.02

TABLE IV: Cross-modality retrieval mean Average Precision (mAP) on MMWand.

Input Modality	Retrieved Modality	mAP	
		w/ $L_{\text{InfoNCE}}^{(tact\leftrightarrow txt)}$	w/o $L_{\text{InfoNCE}}^{(tact\leftrightarrow txt)}$
Image	Tactile	**31.52**	27.08
	Text	**43.87**	**43.87**
Tactile	Image	**38.71**	31.73
	Text	**52.25**	29.73
Text	Image	**46.73**	**46.73**
	Tactile	**44.10**	22.37

D. Cross-Modality Retrieval

In cross-modality retrieval, as outlined in [17], the model is tasked with taking input from one modality and sourcing the corresponding data from another modality. For example, when presented with a tactile image of an apple, the "tactile2language" model is designed to retrieve the relevant linguistic description of the apple from a pool containing sentences for objects. For evaluation, we utilize the mean Average Precision (mAP) score range [0, 100] following [17]. We present the results of our model trained on the MMWand dataset in Table IV.

We further explored the impact of the InfoNCE loss between tactile and text modalities, denoted as $L_{\text{InfoNCE}}^{(tact\leftrightarrow txt)}$. This was done by comparing the model's performance when the $L_{\text{InfoNCE}}^{(tact\leftrightarrow txt)}$ term was removed from L_{total} in Eq. (2). We can note that $L_{\text{InfoNCE}}^{(tact\leftrightarrow txt)}$ enhances mAP across all retrieval types, confirming the efficacy of $L_{\text{InfoNCE}}^{(tact\leftrightarrow txt)}$ in aligning modalities.

E. Robotics Application

To highlight the benefits of our work in robotic applications, we measure the embedding distance from a current state to a goal state. This embedding distance can be used as a dense reward/cost function, as proposed in [4], [35]. We extend it with tactile information. Due to page limitations, we present only a selection of sample results here, but additional results are available in our supplementary video. A perfect

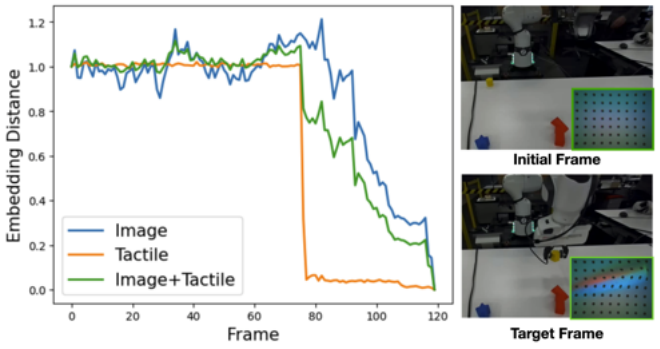

Fig. 8: Embedding distance curves for robot control on different modalities.

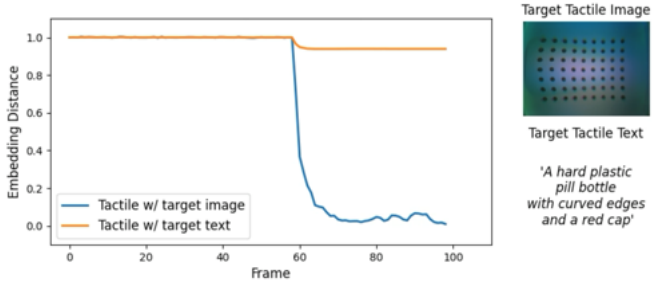

Fig. 9: Embedding distance curves for robot control on different target modalities.

dense cost function would be monotonically and smoothly decreasing for a successful trajectory such that a decrease in the cost is informative that the completion of the task is getting closer.

In our experiment, we execute a pick-and-place task using a Franka Research 3 robotic arm. We equipped the gripper with a 3D-printed mount[1] that holds the tactile sensor. Additionally, the scene is recorded with an external camera. In Fig. 8, we define the last frame of the sequence as the target and compute the embedding distances of each frame with the target frame. The embedding distances were compared using only image, only tactile, and both image and tactile. The tactile input curve includes a step-like transition because the tactile image remains unchanged until the robot arm makes contact with an object. In contrast, the image input curve is smoother but displays noticeable "peaks", which translates into a less informative cost function. By integrating both modalities, we achieve improved curves with fewer peaks and an earlier drop in the embedding distance. This feature enhances its applicability as a dense cost for visuomotor control since it encompasses the act of grasping and provides an earlier indication that the trajectory is correctly solving the task.

In Fig. 9, we delve deeper by comparing the embedding distance plots, this time substituting the target tactile image with the object's texture described in language form. Owing to our model's training on aligning various modalities within the same embedding space, it is observed that even when the target tactile image is replaced with a textual texture description, the embedding distance continues to decrease following the tactile sensor's contact with the object. This result suggests that the tactile embedding is well aligned with the language embedding generated by a text encoder trained on large-scale language data.

[1] https://www.gelsight.com/wp-content/uploads/2022/10/Panda-Adapter-GS-Mini.stp

TABLE V: Object list of the MMWand dataset.

1. Apple	2. Clementine
3. Pineapple	4. Mango
5. Avocado	6. Onion
7. Banana	8. Sweet Potato
9. Slice Bread	10. Tennis Ball
11. Soccer Ball	12. Scrubbing Pad
13. Sponge	14. Bubble Wrap
15. Yarn (Cotton/Polyester Blend)	16. Wine Glass
17. Wire Strainer	18. Wooden Box
19. Russett Potato	20. Coffee Mug
21. Wooden Slotted Spoon	22. Baseball Cap
23. Wicker Basket	24. Loofah
25. Beads	26. Wire Cup (Pencil Holder)
27. Bath Sponge	28. Pill Bottle
29. Comb	30. Calculator
31. Toilet Paper Roll	32. Can of Beans
33. Spoon	34. Battery
35. Power Plug	36. Rice
37. USB Plug	38. Toothpaste
39. Paint Brush	40. Wrench
41. Bowl	42. Flathead Screwdriver
43. Toothbrush	44. Nut (hardware)
45. Rope	46. Mason Jar
47. Carpet	48. Packaging Tape
49. Painters Tape	50. Whisk
51. Hair Brush	52. Screw
53. Fork	54. Marker
55. Dish Towel	56. Wire Cutter
57. Mustard Bottle	58. Knife
59. Pen	60. Baking Pan
61. Eraser	62. Peanut Butter Jar
63. Plate	64. Cardboard Box
65. Candle	66. Styrofoam
67. Peach	68. Waffle Cracker
69. M&Ms	70. Oven Mitt
71. Shaving Cream Can	72. Paper Towel
73. Broccoli	74. Brussel Sprout
75. Soda Can	76. Cucumber
77. Scissors	78. Complex Wrench
79. Philips Head Screwdriver	80. Nail Polish Bottle
81. Book	82. Plastic Fork
83. Plastic Knife	84. Rubber Duck
85. Lightbulb	86. Green Bell Pepper
87. Nuts (pile)	88. Oreo Cookie
89. Popcorn (pile)	90. Human Hand
91. Human Arm	92. Watch
93. Ring	94. Butter Knife
95. Drink Bottle	96. Key
97. Deodorant	98. Forearm
99. Middle Finger	100. Remote
101. Computer Mouse	102. Glasses

VI. Conclusion & Discussion

We introduced the MMWand dataset, improving multimodal learning by aligning tactile input with vision and language using the ImageBind principle. Our experimental results demonstrate significant benefits for robotic tasks, though cross-dataset generalization remains a challenge. Our works's contributions to haptics and multimodal robot learning highlight the importance of integrating tactile feedback with other sensory modalities, advancing robotic perception

and interaction. Future research can build on these findings to further explore the synergy between sensory modalities in robotics.

APPENDIX

A. VisGel Dataset Annotation

Among available datasets, VisGel [16] stands out due to its high-quality tactile images. However, while the dataset is rich in visual and tactile data, it presents certain limitations. For instance, each scene in the dataset contains multiple objects, but only a fraction of the visual image corresponds to a tactile representation. Additionally, the dataset lacks language descriptions that align with the tactile and visual images. We designed our MMWand dataset to address these gaps but we would like to compare it to existing datasets. Thus, we have implemented a pipeline to augment an existing dataset with automatically generated labels. We applied this augmentation pipeline to the VisGel dataset in order to use it in the same way as our MMWand dataset and produce comparable results. We assemble image-tactile pairs from VisGel using two matching algorithms: 1) frame-matching and 2) instance-matching. Then, we generate linguistic descriptions for the identified pairs.

Frame-Matching. A single video in the dataset contains multiple tactile and visual frames. However, since the touch occurs once and all tactile images are close to identical during the touch, we reduce redundancy by pairing one image frame with one tactile frame, as illustrated in Fig. 10. Frames where the tactile sensor is inactive or merely replicates another frame are excluded. When the tactile sensor interacts with an object it often occludes it due to the camera's perspective. Therefore, we select the first unobstructed image frame from the video as a reference frame. We select the contact frame as the one that maximizes the deformation with respect to the reference tactile frame (no-contact condition). The distance is the norm of the pixel difference between two frames. Finally, we match the reference image frame with the contact tactile frame. This forms the matching pair for representation learning. Additionally, we observed that robots do not consistently make contact with the object or sometimes fail to touch it entirely. Consequently, we filtered out the scenes that lacked sufficient deformation.

Instance-Matching. The tactile image corresponds only to a small specific region of the object and thus, to the overhead camera image. To account for this, we match the tactile images to a crop of the object that is in contact with the sensor. We first produce object crops using GLIP [36]. This model returns object crops given text queries. We used the query *'object'* to extract all the objects from the reference image frame. The GLIP model gave the best results in our comparison with other state-of-the-art object detectors [36]–[39]. To find which object is touched, we match the object crops with the sensor location in the image contact frame. We employ the object tracking algorithm [40] to track the sensor location. At the contact frame, we select the object crop that yields the maximum Intersection Over Union (IOU) with the sensor crop. An example of this process is given in Fig. 11.

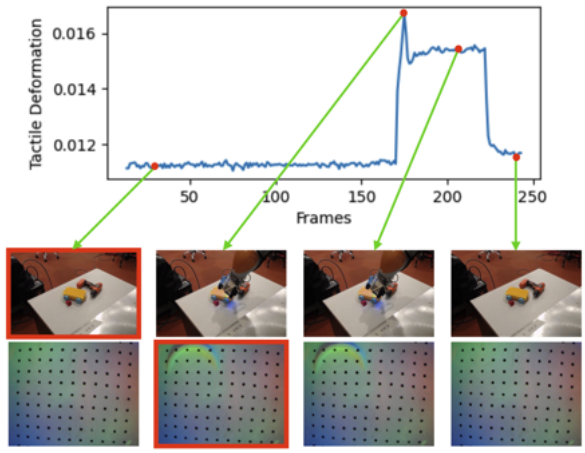

Fig. 10: **Frame Matching Process.** We pair an image frame with a tactile frame that exhibits the maximum deformation for each touch video. (Highlighted with red boxes)

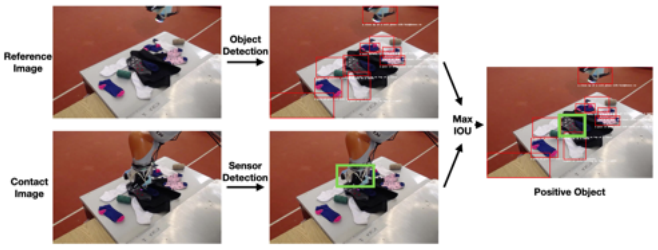

Fig. 11: **Instance Matching Process.** To eliminate unrelated backgrounds from the touch image, we crop the object instance highlighted by a green box in the image, determined by the IOU between the object instance and the touch sensor.

Language Description Generation. To enrich the VisGel dataset further, we automatically generate captions for the selected object crops using BLIP-2 [41]. The captioning process is enhanced by prompting the system with the query '*What specific object is presented in the scene?*'. We observed qualitatively a marked improvement in caption quality compared to the unprompted way. The automatically generated caption do not match the quality of the human-created labels of our MMWand dataset. In particular, these captions never contain any texture information about the object. We believe that some overlap of information is important to learn a meaningful alignment of the different modalities. Therefore this approach is limited and we expect MMWand to yield better results.

Acknowledgement. We acknowledge Feddersen Chair Funds and the US National Science Foundation (FW-HTF 1839971, PFI-TT 2329804) for Dr. Karthik Ramani.

REFERENCES

[1] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, *et al.*, "Palm-e: An embodied multimodal language model," *arXiv preprint arXiv:2303.03378*, 2023. 1

[2] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn, *et al.*, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," *arXiv preprint arXiv:2307.15818*, 2023. 1

[3] S. Karamcheti, S. Nair, A. S. Chen, T. Kollar, C. Finn, D. Sadigh, and P. Liang, "Language-driven representation learning for robotics," *arXiv preprint arXiv:2302.12766*, 2023. 1

[4] Y. J. Ma, W. Liang, V. Som, V. Kumar, A. Zhang, O. Bastani, and D. Jayaraman, "Liv: Language-image representations and rewards for robotic control," *arXiv preprint arXiv:2306.00958*, 2023. 1, 5

[5] S. Nair, A. Rajeswaran, V. Kumar, C. Finn, and A. Gupta, "R3m: A universal visual representation for robot manipulation," in *Conference on Robot Learning*. PMLR, 2023, pp. 892–909. 1

[6] A. Wilson, H. Jiang, W. Lian, and W. Yuan, "Cable routing and assembly using tactile-driven motion primitives," *arXiv preprint arXiv:2303.11765*, 2023. 1

[7] Y. She, S. Wang, S. Dong, N. Sunil, A. Rodriguez, and E. Adelson, "Cable manipulation with a tactile-reactive gripper," *The International Journal of Robotics Research*, vol. 40, no. 12-14, pp. 1385–1401, 2021. 1

[8] A. Alspach, K. Hashimoto, N. Kuppuswamy, and R. Tedrake, "Soft-bubble: A highly compliant dense geometry tactile sensor for robot manipulation," in *2019 2nd IEEE International Conference on Soft Robotics (RoboSoft)*. IEEE, 2019, pp. 597–604. 1

[9] S. Joonhigh, N. Kuppuswamy, A. Beaulieu, A. Alspach, and R. Tedrake, "Variable compliance and geometry regulation of soft-bubble grippers with active pressure control," in *2021 IEEE 4th International Conference on Soft Robotics (RoboSoft)*. IEEE, 2021, pp. 169–175. 1

[10] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra, "Imagebind: One embedding space to bind them all," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15 180–15 190. 1, 2, 4, 5

[11] Y. Zhang, K. Gong, K. Zhang, H. Li, Y. Qiao, W. Ouyang, and X. Yue, "Meta-transformer: A unified framework for multimodal learning," *arXiv preprint arXiv:2307.10802*, 2023. 1, 2

[12] F. Yang, C. Ma, J. Zhang, J. Zhu, W. Yuan, and A. Owens, "Touch and go: Learning from human-collected vision and touch," *Advances in Neural Information Processing Systems*, vol. 35, pp. 8081–8103, 2022. 1, 2, 3

[13] W. Yuan, S. Dong, and E. H. Adelson, "Gelsight: High-resolution robot tactile sensors for estimating geometry and force," *Sensors*, vol. 17, no. 12, p. 2762, 2017. 1, 3

[14] R. Calandra, A. Owens, M. Upadhyaya, W. Yuan, J. Lin, E. H. Adelson, and S. Levine, "The feeling of success: Does touch sensing help predict grasp outcomes?" in *Conference on Robot Learning*. PMLR, 2017, pp. 314–323. 2, 3

[15] R. Calandra, A. Owens, D. Jayaraman, J. Lin, W. Yuan, J. Malik, E. H. Adelson, and S. Levine, "More than a feeling: Learning to grasp and regrasp using vision and touch," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3300–3307, 2018. 2, 3

[16] Y. Li, J.-Y. Zhu, R. Tedrake, and A. Torralba, "Connecting touch and vision via cross-modal prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 609–10 618. 2, 3, 4, 7

[17] R. Gao, Y. Dou, H. Li, T. Agarwal, J. Bohg, Y. Li, L. Fei-Fei, and J. Wu, "The objectfolder benchmark: Multisensory learning with neural and real objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17 276–17 286. 2, 3, 5

[18] R. Gao, Z. Si, Y.-Y. Chang, S. Clarke, J. Bohg, L. Fei-Fei, W. Yuan, and J. Wu, "Objectfolder 2.0: A multisensory object dataset for sim2real transfer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 598–10 608. 2

[19] R. Gao, Y.-Y. Chang, S. Mall, L. Fei-Fei, and J. Wu, "Objectfolder: A dataset of objects with implicit visual, auditory, and tactile representations," in *CoRL*, 2021. 2

[20] S. Kanitkar, H. Jiang, and W. Yuanl, "Poseit: A visual-tactile dataset of holding poses for grasp stability analysis," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 71–78. 2

[21] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763. 2, 4

[22] T. Lüddecke and A. Ecker, "Image segmentation using text and image prompts," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7086–7096. 2

[23] A. Guzhov, F. Raue, J. Hees, and A. Dengel, "Audioclip: Extending clip to image, text and audio," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 976–980. 2

[24] R. Zhang, Z. Guo, W. Zhang, K. Li, X. Miao, B. Cui, Y. Qiao, P. Gao, and H. Li, "Pointclip: Point cloud understanding by clip," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8552–8562. 2

[25] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020. 2

[26] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023. 2

[27] R. Anil, A. M. Dai, O. Firat, M. Johnson, D. Lepikhin, A. Passos, S. Shakeri, E. Taropa, P. Bailey, Z. Chen, *et al.*, "Palm 2 technical report," *arXiv preprint arXiv:2305.10403*, 2023. 2

[28] C. Zimmermann, D. Ceylan, J. Yang, B. Russell, M. Argus, and T. Brox, "Freihand: A dataset for markerless capture of hand pose and shape from single rgb images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 813–822. 3

[29] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018. 4

[30] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2020. 4, 5

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, cite arxiv:1412.6980 Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015. [Online]. Available: http://arxiv.org/abs/1412.6980 4

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778. 5

[33] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014. 5

[34] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255. 5

[35] Y. J. Ma, S. Sodhani, D. Jayaraman, O. Bastani, V. Kumar, and A. Zhang, "Vip: Towards universal visual reward and representation via value-implicit pre-training," *arXiv preprint arXiv:2210.00030*, 2022. 5

[36] L. H. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J.-N. Hwang, *et al.*, "Grounded language-image pre-training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 965–10 975. 7

[37] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=gZ9hCDWe6ke 7

[38] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding yolo series in 2021," *arXiv preprint arXiv:2107.08430*, 2021. 7

[39] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *ECCV*, 2020. 7

[40] Y. Cui, C. Jiang, L. Wang, and G. Wu, "Mixformer: End-to-end tracking with iterative mixed attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 608–13 618. 7

[41] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," *arXiv preprint arXiv:2301.12597*, 2023. 7