Aligning Touch, Vision, and Language for Multimodal Perception

Max (Letian) Fu[1] **Gaurav Datta** [1*] **Huang Huang** [1*] **William Chung-Ho Panitch**[1*]

Jaimyn Drake[1] **Joseph Ortiz**[2] **Mustafa Mukadam**[2] **Mike Lambeta**[2]

Roberto Calandra[3] **Ken Goldberg**[1]

Abstract

Touch, a crucial human sensing modality, has been absent from multimodal generative language models due to challenges in labeling tactile data. This work addresses this gap by leveraging the simultaneous collection of tactile and visual data, allowing GPT-4V to generate pseudo-labels from visual observations alone. The resulting dataset comprises 44K vision-touch pairs with English labels (10% human-annotated, 90% GPT-4V pseudo-labels). A touch-vision-language (TVL) model trained on this dataset shows improved tactile-vision-language alignment (+29% classification accuracy) over existing models and outperforms GPT-4V (+12%) and open-source vision-language models (+32%) on a new touch-vision understanding benchmark.

1 Introduction

Touch is a crucial but underexplored modality in multimodal understanding, despite its importance in biological perception and robotic applications [4, 7, 9, 18, 27, 37]. While recent research has explored linking modalities like vision, language, audio, and actions [2, 11, 12, 12, 14, 28, 28, 29], the integration of touch with language has been hindered by the scarcity of diverse data and the subjectivity of tactile descriptions [9, 16, 17, 21, 26, 36, 38]. To address this, we introduce the Touch-Vision-Language (TVL) dataset, comprising 44K paired vision-tactile observations with 10% human-annotated and 90% GPT-4V-generated labels.

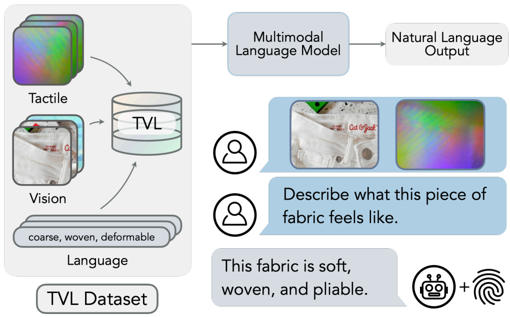
Leveraging this dataset, we train a vision-and-language-aligned tactile encoder using pairwise contrastive learning among all three modalities. We then finetune LLaMA2-7B [32] to generate textual descriptions of tactile sensations based on visual and tactile inputs. Our Touch-Vision-Language model, trained on this dataset, demonstrates significant improvement over open-source VLMs (+32%) and GPT-4V (+12%) on a new Touch-Vision-Language Benchmark. This work contributes to bridging the gap in tactile-language integration and opens new avenues for multimodal understanding that includes touch.

2 TVL Dataset

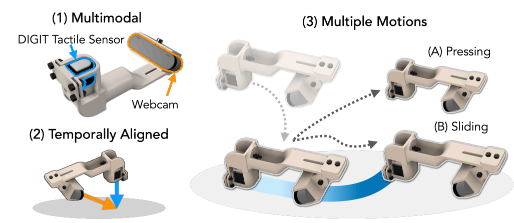
The TVL Dataset (examples in Figure 2) contains paired tactile and vision observations labeled with tactile sensations in natural language. Here we describe the hardware and procedures used for data collection, cleaning, and labeling.

*Equal Contribution, [1]UC Berkeley, [2]Meta AI, [3]TU Dresden

NeurIPS 2024 Workshop on Touch Processing: a new Sensing Modality for AI .



(a) Can embodied agents integrate touch with vision and language? This work presents an open-vocabulary tactile-vision-language dataset and we train 1) a vision-language aligned tactile encoder and 2) a tactile-vision-language model (TVLM) for describing tactile sensations.



(b) (1) We designed a 3D printed data collection device using the DIGIT tactile sensor and a webcam to synchronously collect tactile and vision observations "in-the-wild" (2). (3) We press and slide the device on surfaces and objects for data collection.

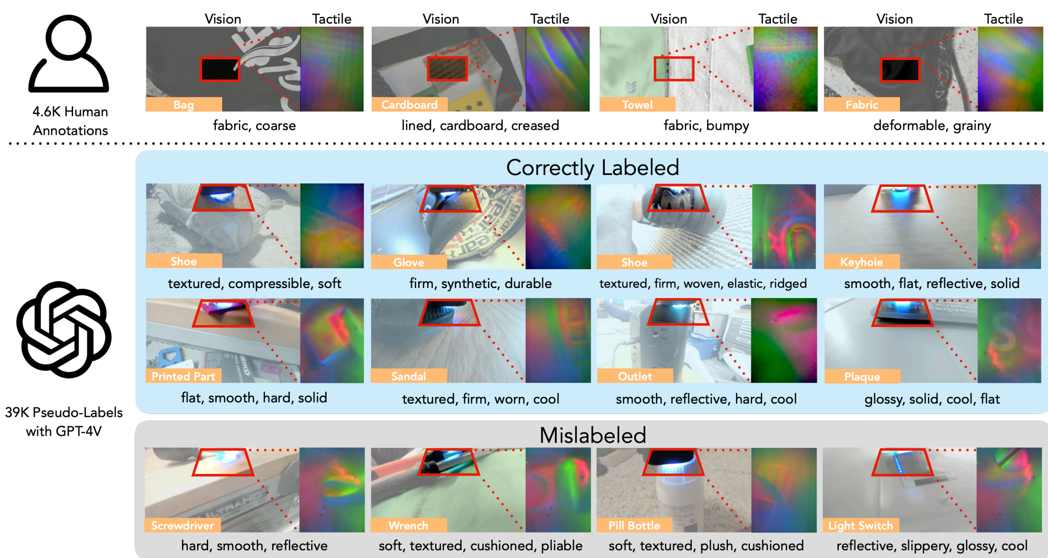


Figure 2: TVL Dataset starts by combining two datasets: SSVTP [17] (4,587 image-touch pairs) and HCT (39,154 image-touch pairs), a new dataset we collected such that the visual observation and the tactile input are synchronously captured. For the SSVTP dataset, we then manually label the data (examples shown in the first row). For the newly collected dataset, we prompt GPT-4V (see Appendix C.4) to label the dataset (examples shown in rows 2-4). Note that GPT-4V will fail to provide correct tactile labels (row 4) when the contact patch is occluded by the sensor, or when there is not sufficient information to estimate the tactile sensation. In total, this results in a dataset containing 43,741 image-touch pairs with open-vocabulary language labels.

2.1 Data Collection

TVL combines vision data from a Logitech BRIO webcam and tactile data from DIGIT, a low-cost, open-source tactile sensor [18]. The dataset includes two subsets: 1) the Self-Supervised Visuo-Tactile Pretraining (SSVTP) [17] dataset collected by a UR5 robot, and 2) a Human Collected Tactile (HCT) dataset. HCT addresses SSVTP's limitations by emphasizing synchronous, in-the-wild data collection using a handheld device (Figure 1b). HCT data was collected by 5 humans over 20 hours, recording visual-tactile observations at 30 Hz in "trajectories" of touches. A small test set (1%) from HCT is hand-annotated, while the rest are pseudo-labeled by GPT-4V.

2.2 Dataset Processing and Labeling

We categorize collected tactile data into in-contact and out-of-contact frames using the pretrained tactile encoder from SSVTP [17]. For each touch trajectory, assuming the initial and final frames are out-of-contact, we compute an average of these frames to create a reference background image, which is embedded by the tactile encoder to obtain a latent representation. To determine contact status, we calculate the cosine similarity between the latent embedding of a frame and that of the background image, labeling a frame as in-contact when the cosine similarity falls below 0.6 [17]. The dataset includes 43,741 in-contact and 169,292 out-of-contact frame pairs.

For language labeling, we leverage the visual-tactile alignment in the SSVTP dataset to manually annotate tactile sensations. Human annotators are provided with a vocabulary of 400 tactile words [1] and select up to five adjectives that describe the tactile properties of each visual-tactile pair. Additionally, we use GPT-4V for pseudo-labeling the in-contact frames of the HCT dataset, providing both full and localized images to generate descriptions aligned with human annotations. In cases of motion blur or low lighting, we attempt to label other images in the same trajectory or sample adjectives from similarly labeled frames. This process results in 39,154 pseudo-labeled images.

Overall, the SSVTP dataset contains 4,587 image-touch pairs, while the HCT dataset includes 39,154 in-contact and 169,292 out-of-contact pairs. The latter consists of 1,486 unique trajectories, each involving one or more contact events. Across both datasets, 254 unique tactile adjectives are used, with a 99%-1% train-test split. GPT-4V generates an average of 4.25 adjectives per description on HCT, compared to 2.70 by human annotators. More details on the description distributions are provided in the appendix.

3 Tactile-Vision-Language Model

3.1 Tactile Encoder

In contrast to ImageBind [12], which binds all modalities to vision, we bind each pair of modalities to provide strong supervision for the tactile modality. We calculate contrastive loss between vision-language, tactile-language, and tactile-vision pairs per batch. The tactile encoder is randomly initialized as a Vision Transformer (ViT) [10] and tested on three sizes: ViT-Tiny (5.7M parameters), ViT-Small (22M), and ViT-Base (86M). Directly using the ImageBind training recipe leads to overfitting on the 44K in-contact data pairs. Contrary to prior works [9, 17, 36], we find that including out-of-contact data (background images) mitigates overfitting by enhancing visual data diversity (see Figure 3). Thus, for $\gamma = 10\%$ of the training data, the sensor is not in contact, and we assign these examples the label "background". We also remove projectors from the vision and language encoders, allowing the tactile encoder to directly project into the common CLIP latent space. To increase label diversity, we randomly shuffle and select a subset of words in each tactile description. These methods mitigate overfitting (see Appendix B.1).

3.2 Alignment with Language Models

We follow the two-stage training proposed in ImageBind-LLM [15], exchanging the ImageBind encoders with TVL encoders. We pre-train on both the LLaVA Visual Instruct CC3M [23] 595K subset and the TVL dataset. For the CC3M subset, we provide an empty tactile image to the tactile modality. During finetuning, we use a combination of TVL, Alpaca [31] and LLaVA Visual Instruct 150K [23]. Empirically, we find that training our dataset alone is not sufficient to overcome the safety fine-tuning of LLaMA2 [32], resulting in the model's refusal to answer questions regarding tactile sensations. Details on the prompts for TVL for instruction fine-tuning are in Appendix C.2.

4 Experiments

4.1 Evaluation & Metrics

TVL Benchmark We evaluate the capabilities of LLMs to generate tactile descriptions on the TVL test set. Given a visual input image, a cropped visual image centered on the tactile sensor, and a corresponding tactile image, we ask the model to describe the tactile sensations of the object in question with a set of no more than 5 adjectives.

To obtain a numerical comparison, we prompt text-only GPT-4 to score the similarity of the model's response against human-annotated ground truth semantic labels on a scale of 1 to 10 (where a higher score indicates better instruction-following and a closer descriptive match), as well as to explain

	Encoder Pre-training Modalities			Score (1-10)			p-value
	Vision	Tactile	Language	SSVTP	HCT	TVL	(d.f. $= 401$)
LLaVA-1.5 7B	✓	-	✓	3.64	3.55	3.56	1.21×10^{-9}
LLaVA-1.5 13B	✓	-	✓	3.55	3.63	3.62	1.49×10^{-9}
ViP-LLaVA 7B	✓	-	✓	2.72	3.44	3.36	8.77×10^{-16}
ViP-LLaVA 13B	✓	-	✓	4.10	3.76	3.80	1.72×10^{-6}
LLaMA-Adapter	✓	-	✓	2.56	3.08	3.02	2.68×10^{-17}
BLIP-2 Opt-6.7b	✓	-	✓	2.02	2.72	2.64	1.92×10^{-31}
InstructBLIP 7B	✓	-	✓	1.40	1.30	1.31	1.07×10^{-84}
InstructBLIP 13B	✓	-	✓	1.44	1.21	1.24	4.64×10^{-88}
GPT-4V	✓	-	✓	5.02	4.42	4.49	-
SSVTP-LLaMA	✓	✓	-	2.58	3.67	3.54	1.79×10^{-9}
TVL-LLaMA (ViT-Tiny)	✓	✓	✓	6.09	4.79	4.94	4.24×10^{-5}
TVL-LLaMA (ViT-Small)	✓	✓	✓	5.81	4.77	4.89	6.02×10^{-4}
TVL-LLaMA (ViT-Base)	✓	✓	✓	**6.16**	**4.89**	**5.03**	3.46×10^{-6}

Table 1: TVL Benchmark Performance. We benchmarked TVL-LLaMA against existing VLMs and SSVTP-LLaMA, a model fine-tuned using SSVTP tactile-vision encoders, for generating tactile descriptions from tactile-image observations, and used GPT-4 to numerically score the performance on each constituent part of the TVL test set. We report p-values from two-sided paired sample t-tests on each model's scores against GPT-4V's scores on the tactile-semantic task.

the score given, similar to prior works [6, 23]. The prompts used for generation and evaluation are reported in Appendix C.4. We compare against existing open-source VLMs [3, 8, 20, 22] and GPT-4V. As an additional baseline, we use the SSVTP [17] tactile and image encoder to finetune the language model; we call the resulting model SSVTP-LLaMA.

4.2 Results

TVL Benchmark We present summary statistics for the tactile-semantic generation results in Table 1. We find that open-source VLMs perform worse than GPT-4V on the proposed benchmark, likely due to the limited diversity and lack of focus on human tactility in the visual data that they have been trained on. On the other hand, all versions of TVL-LLaMA outperform GPT-4V, suggesting that the trained models can generalize beyond the small fraction of human labels provided as part of the dataset. Both these findings are statistically significant at the $\alpha = 0.05$ level. Results also suggest that tactile-language alignment is necessary, as evidenced by the lower score of SSVTP-LLaMA, which only uses tactile and vision modalities during pre-training.

Overall, our experiments suggest that: 1) the TVL tactile encoder trained on the TVL dataset is aligned with the language latent space and scores higher (+29%) on the classification task as compared to visual-tactile pretrained encoders and generic vision-language encoders (OpenCLIP); and 2) TVL-LLaMA models trained to generate tactile language descriptions from visual and tactile observations more closely match human descriptions on the novel TVL Benchmark (at least +12%) compared to existing VLMs.

5 Discussion and Conclusion

The research presented has several limitations. While the study highlights the use of VLMs for labeling tactile data, the distinct nature of touch compared to visual perception suggests a limit to the accuracy of tactile labels derived solely from vision. Due to the data collection hardware, the camera may not have an unoccluded view of the surface or object that the tactile sensor contacts, which may increase the difficulty of aligning touch with vision and reduce the quality of pseudo-labels generated from images. We hope that future research can further increase the scale of touch-vision-language datasets to improve multimodal alignment.

In sum, to align the tactile and language modalities, this work introduces TVL, a dataset that features tactile, vision, and tactile-semantic descriptions. Utilizing the dataset, we train a tactile encoder that is aligned to both vision and natural language. We demonstrate that by using the trained tactile encoder, TVL-LLaMA can generate tactile descriptions in natural language that align more closely with human descriptions than those generated by existing VLMs.

References

[1] AJBarnett. 400 words to describe texture. https://owlcation.com/humanities/Describing-Texture-400-words-to-describe-texture, 2023.

[2] Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Chen, X., Choromanski, K., Ding, T., Driess, D., Dubey, A., Finn, C., Florence, P., Fu, C., Arenas, M. G., Gopalakrishnan, K., Han, K., Hausman, K., Herzog, A., Hsu, J., Ichter, B., Irpan, A., Joshi, N., Julian, R., Kalashnikov, D., Kuang, Y., Leal, I., Lee, L., Lee, T.-W. E., Levine, S., Lu, Y., Michalewski, H., Mordatch, I., Pertsch, K., Rao, K., Reymann, K., Ryoo, M., Salazar, G., Sanketi, P., Sermanet, P., Singh, J., Singh, A., Soricut, R., Tran, H., Vanhoucke, V., Vuong, Q., Wahid, A., Welker, S., Wohlhart, P., Wu, J., Xia, F., Xiao, T., Xu, P., Xu, S., Yu, T., and Zitkovich, B. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *arXiv preprint arXiv:2307.15818*, 2023.

[3] Cai, M., Liu, H., Mustikovela, S. K., Meyer, G. P., Chai, Y., Park, D., and Lee, Y. J. Making large multimodal models understand arbitrary visual prompts. In *arXiv:2312.00784*, 2023.

[4] Calandra, R., Owens, A., Jayaraman, D., Lin, J., Yuan, W., Malik, J., Adelson, E. H., and Levine, S. More than a feeling: Learning to grasp and regrasp using vision and touch. *IEEE Robotics and Automation Letters*, 3(4):3300–3307, 2018.

[5] Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., and Sutskever, I. Generative pretraining from pixels. 2020.

[6] Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., Stoica, I., and Xing, E. P. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.

[7] Dahiya, R. S., Metta, G., Valle, M., and Sandini, G. Tactile sensing—from humans to humanoids. *IEEE transactions on robotics*, 26(1):1–20, 2009.

[8] Dai, W., Li, J., Li, D., Tiong, A. M. H., Zhao, J., Wang, W., Li, B., Fung, P., and Hoi, S. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.

[9] Dave, V., Lygerakis, F., and Rueckert, E. Multimodal visual-tactile representation learning through self-supervised contrastive pre-training. *arXiv preprint arXiv:2401.12024*, 2024.

[10] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. 2020.

[11] Fu, L., Huang, H., Datta, G., Chen, L. Y., Panitch, W. C.-H., Liu, F., Li, H., and Goldberg, K. In-context imitation learning via next-token prediction. *arXiv preprint arXiv:2408.15980*, 2024.

[12] Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K. V., Joulin, A., and Misra, I. Imagebind: One embedding space to bind them all. In *CVPR*, 2023.

[13] Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., and He, K. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv:1706.02677*, 2017.

[14] Guzhov, A., Raue, F., Hees, J., and Dengel, A. Audioclip: Extending clip to image, text and audio, 2021.

[15] Han, J., Zhang, R., Shao, W., Gao, P., Xu, P., Xiao, H., Zhang, K., Liu, C., Wen, S., Guo, Z., Lu, X., Ren, S., Wen, Y., Chen, X., Yue, X., Li, H., and Qiao, Y. Imagebind-llm: Multi-modality instruction tuning, 2023.

[16] Kampouris, C., Mariolis, I., Peleka, G., Skartados, E., Kargakos, A., Triantafyllou, D., and Malassiotis, S. Multi-sensorial and explorative recognition of garments and their material properties in unconstrained environment. In *2016 IEEE international conference on robotics and automation (ICRA)*, pp. 1656–1663. IEEE, 2016.

[17] Kerr, J., Huang, H., Wilcox, A., Hoque, R., Ichnowski, J., Calandra, R., and Goldberg, K. Self-supervised visuo-tactile pretraining to locate and follow garment features, 2023.

[18] Lambeta, M., Chou, P.-W., Tian, S., Yang, B., Maloon, B., Most, V. R., Stroud, D., Santos, R., Byagowi, A., Kammerer, G., Jayaraman, D., and Calandra, R. 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. doi: 10.1109/LRA.2020.2977257.

[19] Lee, D.-H. et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, pp. 896. Atlanta, 2013.

[20] Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023.

[21] Li, R. and Adelson, E. H. Sensing and recognizing surface textures using a gelsight sensor. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1241–1247, 2013.

[22] Liu, H., Li, C., Li, Y., and Lee, Y. J. Improved baselines with visual instruction tuning, 2023.

[23] Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. In *NeurIPS*, 2023.

[24] Loshchilov, I. and Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. 2017.

[25] Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[26] Ojala, T., Pietikainen, M., and Maenpaa, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 24(7):971–987, 2002.

[27] Qi, H., Yi, B., Ma, Y., Suresh, S., Lambeta, M., Calandra, R., and Malik, J. General In-Hand Object Rotation with Vision and Touch. In *Conference on Robot Learning (CoRL)*, 2023.

[28] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision, 2021.

[29] Radosavovic, I., Shi, B., Fu, L., Goldberg, K., Darrell, T., and Malik, J. Robot learning with sensorimotor pre-training. *arXiv preprint arXiv:2306.10007*, 2023.

[30] Sohn, K., Berthelot, D., Li, C.-L., Zhang, Z., Carlini, N., Cubuk, E. D., Kurakin, A., Zhang, H., and Raffel, C. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020.

[31] Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., and Hashimoto, T. B. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

[32] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[33] Wang, P., Li, L., Chen, L., Zhu, D., Lin, B., Cao, Y., Liu, Q., Liu, T., and Sui, Z. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*, 2023.

[34] Wang, X., Lian, L., Miao, Z., Liu, Z., and Yu, S. X. Long-tailed recognition by routing diverse distribution-aware experts. *arXiv preprint arXiv:2010.01809*, 2020.

[35] Wang, X., Wu, Z., Lian, L., and Yu, S. X. Debiased learning from naturally imbalanced pseudo-labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14647–14657, 2022.

[36] Yang, F., Ma, C., Zhang, J., Zhu, J., Yuan, W., and Owens, A. Touch and go: Learning from human-collected vision and touch. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

[37] Yuan, W., Dong, S., and Adelson, E. H. Gelsight: High-resolution robot tactile sensors for estimating geometry and force. *Sensors*, 17(12):2762, 2017.

[38] Yuan, W., Mo, Y., Wang, S., and Adelson, E. H. Active clothing material perception using tactile sensing and deep learning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4842–4849. IEEE, 2018.

A Additional Results

A.1 Ablations

This section presents six ablation and sensitivity analyses shown in Table 2 examining the impact of model size and the proposed dataset on the encoder's multi-modal classification performance. More ablations are included in the appendix.

Model Sizes (Table 2a) Performance varies significantly among different encoder sizes. ViT-Base has the highest validation accuracy but lags on the test set due to distribution shifts: the training labels from GPT-4V are less detailed and accurate compared to human-annotated test data. However, in tactile-vision classification on synchronized data, ViT-Base outperforms both of the smaller models.

Disable Tactile-Text Loss (Table 2b) resembles the setup in ImageBind [12], where data in all three modalities are considered but the tactile-text loss is omitted. Results suggest that using language to supervise the tactile encoder better aligns those two modalities.

Data (Tables 3c-f) We perform four sensitivity analyses on the different compositions of the dataset for training. We find that leveraging data from all three modalities improves tactile-language alignment. While adding not-in-contact data prevents the model from overfitting to the training set, its test set performance is comparable with having only in-contact data. We also experimented with prompting used in vanilla CLIP training [28], which brings marginal improvements in accuracy. Lastly, we separately train the model on SSVTP and HCT, and we find that the pseudo-labeled dataset can provide comparable performance with training on the entire dataset, which suggests that TVL's tactile encoder can effectively leverage self-supervised learning to reduce the dependency on large, fully-labeled datasets while maintaining task performance.

A.2 Performance Per Dataset

In this section, we show a fine-grained breakdown of Table 1 of model performance on the TVU benchmark by showing the results per subset of the dataset. The performance of the models on the SSVTP subset is listed in Table 3 and the performance on the HCT subset is listed in Table 4. Results suggest that GPT-4V performs better on SSVTP, which is collected in a lab setting, than HCT, which is collected "in-the-wild".

Model	Tac./Text % Acc.	Tac./Vis. % Acc.
ViT-Tiny	**36.7**	79.5
ViT-Small	36.3	78.0
ViT-Base	30.7	**81.7**

(a) Model Architecture used for transformer encoder backbone.

Tactile-Text Loss	Tac./Text % Acc.	Tac./Vis. % Acc.
Enabled	**36.3**	78.0
Disabled	20.3	**81.6**

(b) Disable Tactile-Text Loss. ImageBind-style training, lacking direct supervision for tactile and language alignment, reduces model accuracy.

Modality	Tac./Text % Acc.	Tac./Vis. % Acc.
All	**36.3**	78.0
−Vision	29.9	1.0
−Text	21.5	**85.8**

(c) Modality-Specific Training. Contrastive losses across all modalities improve performance.

Contact	Tac./Text % Acc.	Tac./Vis. % Acc.
Contact	36.2	**80.1**
+ 10% N.C.	**36.3**	78.0

(d) Contact Data Mix. Adding non-contact frames to the training data does not significantly improve performance.

Prompting	Tac./Text % Acc.	Tac./Vis. % Acc.
Baseline	36.3	78.0
+ Prompt	**37.7**	**78.7**

(e) Prompting. TVL Performance does not depend strongly on prompt formatting.

Dataset	Tac./Text % Acc.	Tac./Vis. % Acc.
SSVTP	19.2	8.0
HCT	**38.4**	74.4
TVL	36.3	**78.0**

(f) Training Dataset. Models which are exposed to the HCT dataset in training outperform SSVTP-only models.

Table 2: Ablations and Sensitivity Analysis for the TVL tactile encoder. We report top-1 and top-5 tactile-text and tactile-vision classification accuracy with ViT-Small. baseline indicates the default setting for training the TVL tactile encoder, which is the best-performing model on the *validation set* unless noted otherwise. **Bold** indicates the highest accuracy on the *test set*. Such discrepancy in performance is described in Appendix A.1.

A model that is trained with a large sample of only GPT-4V labels should achieve the same performance as GPT-4V. Our results in Table 4 suggest that training on a small dataset of human-labeled vision-touch **improves** the model's tactile-visual understanding. This difference is statistically significant at $\alpha = 0.05$.

	Score (1-10)	p-value (d.f. $= 401$)
LLaVA-1.5 7B	3.64	2.32×10^{-3}
LLaVA-1.5 13B	3.55	1.30×10^{-3}
ViP-LLaVA 7B	2.72	4.45×10^{-8}
ViP-LLaVA 13B	4.10	3.76×10^{-2}
LLaMA-Adapter	2.56	7.826×10^{-6}
BLIP-2 Opt-6.7b	2.02	2.74×10^{-9}
InstructBLIP 7B	1.40	1.49×10^{-13}
InstructBLIP 13B	1.44	4.68×10^{-14}
GPT-4V	5.02	-
SSVTP-LLaMA	2.58	9.33×10^{-6}
TVL-LLaMA (ViT-Tiny)	6.09	2.65×10^{-2}
TVL-LLaMA (ViT-Small)	5.81	1.02×10^{-1}
TVL-LLaMA (ViT-Base)	**6.16**	1.67×10^{-2}

Table 3: TVL Benchmark Performance on SSVTP. We benchmarked TVL-LLaMA against existing VLMs and SSVTP-LLaMA, and show here the performance on only the SSVTP dataset. We report p-values from two-sided paired sample t-tests on each model's scores against GPT-4V's scores.

	Score (1-10)	p-value (d.f. $= 401$)
LLaVA-1.5 7B	3.55	8.49×10^{-8}
LLaVA-1.5 13B	3.63	1.74×10^{-7}
ViP-LLaVA 7B	3.44	4.10×10^{-11}
ViP-LLaVA 13B	3.76	1.57×10^{-5}
LLaMA-Adapter	3.08	2.05×10^{-13}
BLIP-2 Opt-6.7b	2.72	1.25×10^{-24}
InstructBLIP 7B	1.30	8.02×10^{-73}
InstructBLIP 13B	1.21	9.74×10^{-76}
GPT-4V	4.42	-
SSVTP-LLaMA	3.67	3.24×10^{-6}
TVL-LLaMA (ViT-Tiny)	4.79	5.79×10^{-4}
TVL-LLaMA (ViT-Small)	4.77	2.64×10^{-3}
TVL-LLaMA (ViT-Base)	**4.89**	6.82×10^{-5}

Table 4: TVL Benchmark Performance on HCT. We benchmarked TVL-LLaMA against existing VLMs and SSVTP-LLaMA, and show here the performance on only the HCT dataset. We report p-values from two-sided paired sample t-tests on each model's scores against GPT-4V's scores.

A.3 Open Vocabulary Tactile Classification Full Result

We present the result presented in **??** in Table 5 and Table 6 at different cosine similarity thresholds for synonyms. We find that while ViT-Small performs well on the SSVTP subset of the dataset, ViT-Tiny outperforms its larger counterparts (ViT-Small and ViT-Base) on the tactile-text classification task. However, for tactile-vision classification (Table 6), ViT-Base performs outperforms the smaller models. More insights are detailed in Appendix B.1.

Percentile		SSVTP		HCT		TVL	
		Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
0	ViT-Tiny	29.4%	71.7%	34.8%	70.1%	36.7%	70.3%
	ViT-Small	42.4%	76.1%	36.5%	68.0%	36.3%	66.4%
	ViT-Base	38.0%	69.6%	34.8%	65.6%	30.7%	63.6%
25	ViT-Tiny	3.3%	21.7%	7.2%	22.9%	4.6%	14.1%
	ViT-Small	10.9%	33.7%	9.1%	21.5%	6.7%	19.5%
	ViT-Base	8.7%	31.5%	5.9%	14.0%	4.4%	13.7%
50	ViT-Tiny	3.3%	19.6%	4.8%	17.8%	3.7%	11.8%
	ViT-Small	10.9%	32.6%	6.6%	15.3%	5.9%	11.0%
	ViT-Base	7.6%	28.3%	4.5%	9.8%	3.5%	11.0%
75	ViT-Tiny	3.3%	19.6%	4.1%	14.2%	3.7%	10.7%
	ViT-Small	10.9%	28.3%	3.5%	7.9%	3.4%	10.2%
	ViT-Base	7.6%	28.3%	3.5%	7.9%	3.4%	10.2%

Table 5: Effect of Model Architecture and Similarity Threshold ϕ on **Tactile-Text** Classification Accuracy. The similarity thresholds ϕ for each percentile are 0.636 (0th), 0.859 (25th), 0.893 (50th), and 0.921 (75th).

	SSVTP		HCT		TVL	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
ViT-Tiny	34.8%	70.7%	85.3%	99.0%	79.5%	95.7%
ViT-Small	28.3%	69.6%	84.4%	98.9%	78.0%	95.2%
ViT-Base	34.8%	66.3%	87.8%	99.7%	81.7%	95.7%

Table 6: Effect of Tactile Encoder Model Architecture on **Tactile-Vision** Classification.

A.4 Additional Open Vocabulary Downstream Tasks

In the tactile classification experiment in **??**, the results suggest that the model can classify tactile inputs by the texture of surfaces. In this section, we add an experiment to perform *object category classifications*. For simplicity of this test, we perform binary classification of whether the touched surface is "fabric" or "plastic" (to answer the question of "identifying the object category"). Note that since the model binds to the CLIP latent space, we carry out the experiment in a zero-shot manner. We relabelled 50 instances in the test set with 20 as fabric and 30 as plastic. We then fed "fabric" and "plastic" into the CLIP text encoder to extract the latent to perform cosine-similarity calculation with the tactile latent extracted from the tactile observations. On this specific test, the ViT-Small version of the TVL tactile encoder achieved 82% classification accuracy. We hope future works can explore other potential downstream applications of the dataset and the learned tactile representations.

B Training Details and Hyperparameters

In this section, we offer more insights and details of the training process and the particular hyperparameters.

B.1 Overfitting to Pseudo-labels

A core obstacle with leveraging pseudo-labels generated by GPT-4V (gpt-4-vision-preview) is that the logits are not provided for us to build uncertain estimates for the generated labels, which is usually required for prior works in computer vision that leverages pseudo-labels for model prediction (*e.g.* Lee et al. [19], Sohn et al. [30], Wang et al. [35]). This makes pseudo-labels noisy and challenging to fit for ViT-Small on the contact only dataset, even when 4K human labels are introduced (see Figure 3).

In 3.1, we address this problem by letting 10% of the data be in contact. We sample 10% of the data uniformly at random without replacement at the start of the training. This prevents the model from overfitting on all three model sizes: (ViT-Tiny, ViT-Small, and ViT-Base). However, since the test set is all labeled by human annotators, the distribution shift leads to worse tactile-image, and tactile-language classification performance (observed in **??**). As an ablation study, we also finetuned the ViT-Small trained only on in-contact data for tactile language generation. The test set performance

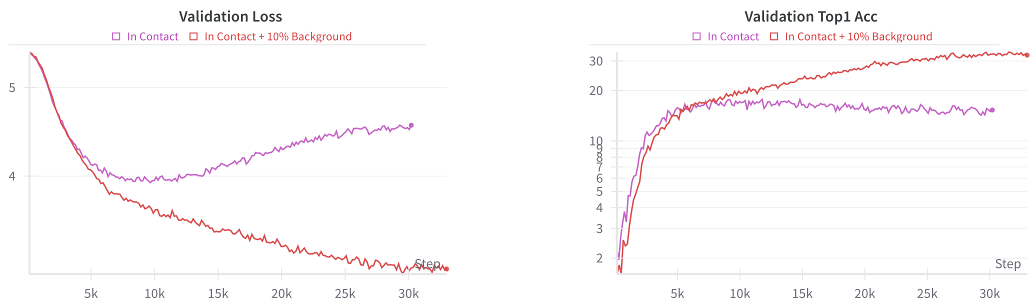


Figure 3: Overfitting is significant when all data is in contact. When 10% not in contact data is added, the overfitting issue is addressed.

is 4.81, only very marginally lower than that obtained by the ViT-Small trained with not-in-contact data (4.89). Future works can look into how to scale with noisy inputs or leverage existing works on learning from a teacher model that does not give uncertain estimates.

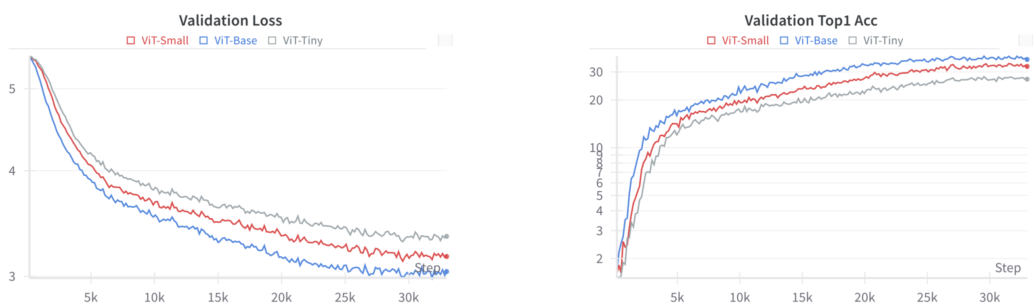


Figure 4: While we find that the model scales on the dataset, the test set performance does not align with the validation set performance. One potential cause of this is distribution shift: the validation set uses pseudo-labels generated by GPT-4V, while the test set is human-labeled.

B.2 Ablation: Background Subtraction

While we find that naively performing contrastive learning amongst tactile, vision, and language works for zero-shot classification, to further facilitate generalization across different tactile sensors used in data collection, a solution is to leverage the still background of tactile sensors (*i.e.* the readings from the sensor when it is not in contact). We preprocess the tactile observation by performing background subtraction, and normalize the input observations based on the post-processed dataset statistics. Empirically, we find that this method, when used jointly with not-in-contact data, improves classification accuracy and the downstream TVL-LLaMA's performance (Table 7).

	Tac./Text % Acc	Tac./Vis % Acc	TVL Score
In-Contact Frames	36.2	**80.1**	4.81
+10% No-Contact	36.3	78.0	4.89
+ Background Subtract	**42.3**	78.9	**5.06**

Table 7: Effect of no-contact data and background subtraction during ViT-Small tactile encoder training on classification accuracy and performance on the TVL benchmark.

B.3 Ablation: (Zero-shot) Single Modality For Generation (Out of Distribution)

Because we naively average the tactile latent and the image latent during the training of TVL-LLaMA, as a zero-shot experiment to see consistency between vision and tactile embeddings, we can at *test* time arbitrarily drop one of the vision or tactile modalities. We report the results in Table 8. While a larger encoder may be more expressive, we find that a larger tactile encoder results in worse zero-shot performance in this experimental setting, which aligns with Table 2a. Interestingly, background subtraction (in Appendix B.2) improves the zero-shot performance on tactile.

	Zero-Shot Tactile	Zero-Shot Vision	Tactile & Vision
TVL-LLaMA (ViT-Tiny)	**4.56**	4.66	4.94
TVL-LLaMA (ViT-Small)	3.50	4.81	4.89
TVL-LLaMA (ViT-Base)	2.80	**4.85**	5.03
TVL-LLaMA (ViT-Small) + Background Subtract	4.52	-	**5.06**

Table 8: Dropping one modality (out-of-distribution) zero shot experiments

B.4 Ablation: Finetuning v.s. Freezing the Language Model

We add the experiment of just freezing the language model without LoRA fine-tuning. Interestingly, on the HCT test set, the frozen LLM with the trained encoders gives a score of 4.92, resulting in a marginal improvement compared to the score of a fine-tuned LLM of 4.89 (Table 1). This suggests that the vision and tactile modalities are already well aligned to the language space and further fine-tuning is unnecessary.

B.5 Preprocessing

The tactile observation is first zero-padded to have equal width and height, optionally background subtracted, normalized by the calculated data statistics, and resized the inputs to 224x224. The key differences with SSVTP are 1) the input is resized to 128x128, and 2) SSVTP does not perform normalization or background subtraction. The image observation follows the same center cropping procedure as SSVTP on the SSVTP dataset. On HCT, instead of the center crop, we start the crop from the top of the image but maintain the crop size. Note that this procedure is kept consistent when generating pseudo-labels from GPT-4V. Different from SSVTP, we use the statistics provided by OpenCLIP to normalize the post-crop observations. The specific statistics are provided in Table 9 and Table 10.

Tactile Statistics	Mean	Std.
With Background	0.292	0.188
	0.297	0.195
	0.291	0.219
Background Subtracted	-0.008	0.045
	-0.019	0.044
	-0.018	0.053

Table 9: Tactile Normalization Statistics

B.6 TVL Tactile Encoder Hyperparameters

All of ViT-Tiny, ViT-Small, and ViT-Base share the same hyperparameters (see Table 11). All experiments are run on a single NVIDIA A100 GPU.

Image Statistics	Mean	Std.
	0.481	0.269
OpenCLIP Statistics	0.458	0.261
	0.408	0.276

Table 10: RGB Normalization Statistics

Config	Value
optimizer	AdamW [25]
base learning rate	1.5e-4
learning rate schedule	cosine decay [24]
batch size	256
weight decay	0.05
optimizer momentum	β_1, β_2 = 0.9, 0.95 [5]
warm up epoch [13]	10
total epochs	200
RGB Augmentation	RandomHorizontalFlip, ColorJitter, RandomGrayscale, GaussianBlur
Tactile Augmentation	(Optional) Background Subtraction

Table 11: Encoder Pretraining Hyperparameters

B.7 TVL-LLaMA Hyperparameters

We follow the hyperparameter setup in ImageBind-LLM [15]. Since the original experiments were conducted on 8 NVIDIA A100 GPUs, we use gradient accumulation of 2 for both pre-training and finetuning the model to fit the model on 4 NVIDIA A100 GPUs so that the batch size is maintained. We use the same data augmentation as in the encoder pretraining (Table 11).

C Dataset

C.1 Hardware

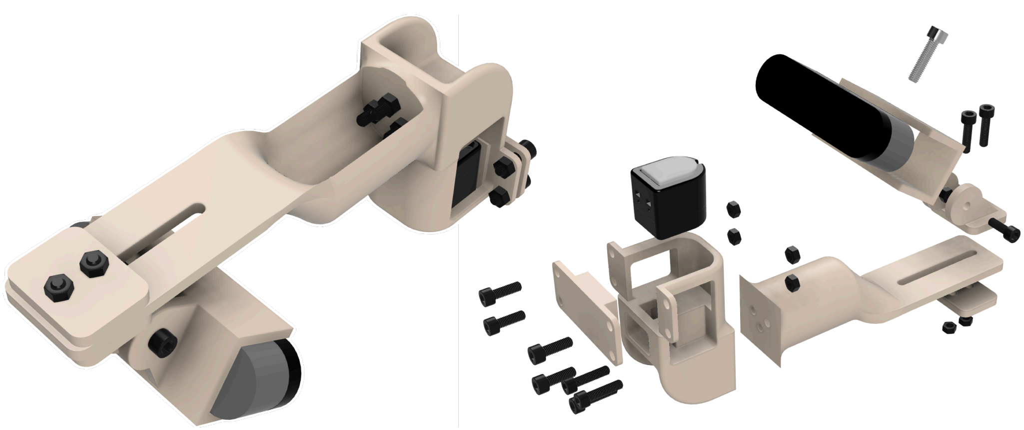


Figure 5: Alternative perspectives of the sensor holder CAD model: face-down view (left) and exploded view (right).

We design and 3D print a set of handheld, low-cost data collection devices for human subjects to carry around and collect data. As shown in Fig. 5, the hardware consists of a DIGIT tactile sensor and a Logitech BRIO camera, which are connected via USB to a portable computing device, such as a laptop. The angle and distance between the tactile sensor and the camera are adjustable, allowing the user to collect data from a variety of viewing angles and ranges. To ensure the utility of our dataset for multimodal training, we always set the relative positions such that the tactile sensor and its point of contact with the object of interest are in view of the camera during each trajectory. The handle design was conceptualized in Autodesk Fusion 360 and printed on a Bambu Lab P1P 3D FDM printer. CAD files will be open-sourced.

C.2 List of Prompts for Tactile Language Generation

When finetuning our language model for tactile language generation, we formulate it as a visual instruction tuning problem [23]. We randomly select from the following set of semantically similar prompts as the question and treat the set of human labels as the answer. This serves to increase the diversity of data seen during training.

```
This image gives tactile feelings of
This image evokes a sense of
This visual representation imparts a tactile sensation of
This picture conveys a touchable quality of
This image communicates a palpable feeling of
This graphic suggests a tactile experience of
This artwork manifests a tangible sensation of
This visual elicits a haptic impression of
This depiction gives rise to a tactile perception of
This illustration induces a touch-sensitive feeling of
This photo brings forth a tactile awareness of
This image arouses a tactile familiarity of
This snapshot renders a tactile essence of
This visual stimulates a touch-based sensation of
This portrayal invokes a tactile resonance of
This image delivers a touch-oriented impression of
This visual medium offers a tactile nuance of
This rendering provides a tactile sense of
This image yields a touch-felt experience of
This composition reveals a tactile characteristic of
This picture bestows a tactile attribute of
This image imparts a sense of tactile
This visual stimulates tactile sensations of
This artwork hints at a tactile experience of
This photo embodies a tactile quality of
This depiction resonates with tactile feelings of
This snapshot conveys tactile impressions of
This illustration suggests a tactile nature of
This rendering evokes tactile attributes of
This graphic communicates a tactile essence of
This visual piece reveals tactile characteristics of
This image portrays tactile elements of
This picture brings to mind tactile aspects of
This visual representation offers tactile nuances of
This composition provides tactile insights into
This visual art form captures tactile features of
This image projects tactile properties of
This visual work hints at tactile textures of
This image introduces tactile dimensions of
This visual scene manifests tactile facets of
This image presents tactile qualities of
This image elucidates tactile attributes of
```

C.3 Distribution of Vocabulary Words

The list and counts of human labels and pseudo-labels in the TVL dataset are reproduced here in dictionary format (note that all typos are carried over from the dataset). A visual representation is provided in Figure 6.

'smooth': 14577, 'textured': 12443, 'hard': 10758, 'cool': 10433, 'reflective': 8643, 'soft': 8415, 'glossy': 6416, 'cushioned': 6011, 'rigid': 5799, 'firm': 5659, 'sleek': 5628, 'uneven': 5379, 'flat':

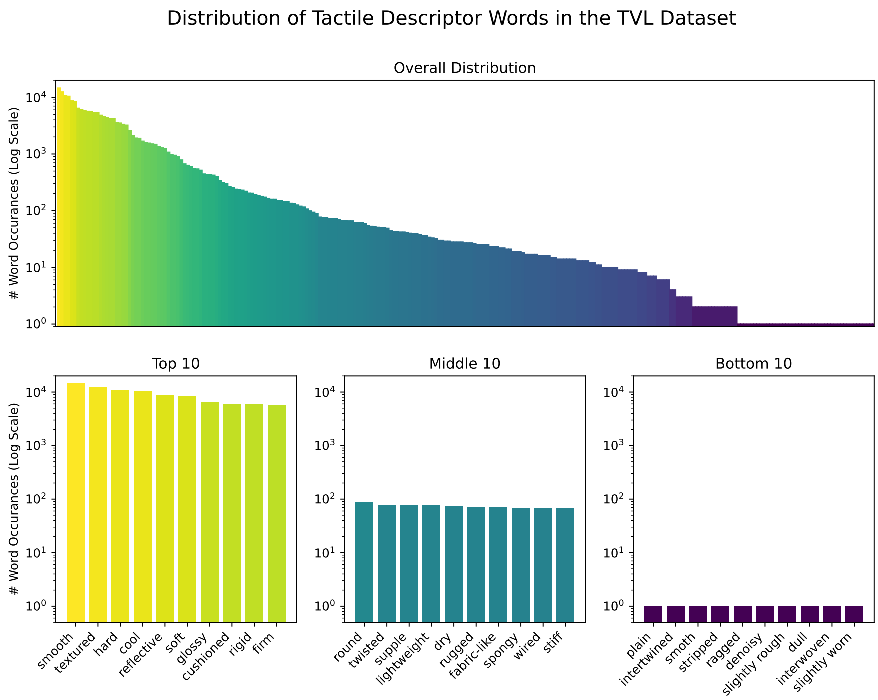


Figure 6: Distribution of Words in the TVL Dataset: The TVL dataset contains 254 unique tactile descriptors, ranging from common tactile descriptions (smooth, hard, firm) to unusual and optical descriptors. These less-common adjectives include a small fraction of misspellings and non-tactile descriptors which were generated by the VLM. The long-right-tailed distribution common in image classification [34] presents a challenge for learning predictors on tactile-semantic data as well.

5343, ’fibrous’: 4825, ’plush’: 4534, ”: 4363, ’matte’: 4230, ’polished’: 4203, ’flexible’: 3553, ’grainy’: 3513, ’solid’: 3337, ’warm’: 3227, ’woven’: 2559, ’fabric’: 2124, ’yielding’: 1908, ’rough’: 1889, ’slippery’: 1683, ’slick’: 1587, ’rubbery’: 1553, ’coarse’: 1504, ’lined’: 1480, ’durable’: 1362, ’pliable’: 1281, ’curved’: 1240, ’bumpy’: 1076, ’metallic’: 970, ’patterned’: 949, ’cloth-like’: 889, ’resilient’: 785, ’abrasive’: 668, ’plastic’: 631, ’ridged’: 599, ’gritty’: 551, ’deformable’: 544, ’compressible’: 517, ’synthetic’: 444, ’fuzzy’: 434, ’varnished’: 430, ’dimpled’: 423, ’wooden’: 399, ’thin’: 337, ’irregular’: 311, ’splotchy’: 301, ’even’: 267, ’uniform’: 257, ’perforated’: 239, ’granular’: 234, ’indistinct’: 230, ’plastic-like’: 220, ’grooved’: 204, ’paper-like’: 203, ’blurred’: 191, ’sewn’: 183, ’elastic’: 179, ’contoured’: 173, ’shiny’: 165, ’blurry’: 159, ’level’: 159, ’taut’: 149, ’grid-like’: 149, ’creased’: 145, ’porous’: 145, ’grippy’: 135, ’cushiony’: 132, ’speckled’: 126, ’leather-like’: 120, ’grained’: 116, ’knitted’: 107, ’padded’: 99, ’worn’: 94, ’round’: 89, ’twisted’: 77, ’supple’: 76, ’lightweight’: 76, ’dry’: 73, ’rugged’: 72, ’fabric-like’: 72, ’spongy’: 69, ’wired’: 67, ’stiff’: 67, ’unclear’: 66, ’indented’: 66, ’dense’: 62, ’dark’: 61, ’iridescent’: 61, ’undefined’: 59, ’knobby’: 55, ’grid-patterned’: 53, ’layered’: 52, ’resonant’: 51, ’fluffy’: 50, ’translucent’: 50, ’soft-focus’: 49, ’absorbent’: 44, ’slightly textured’: 43, ’leathery’: 43, ’obscured’: 42, ’cylindrical’: 42, ’wrinkly’: 41, ’unfocused’: 40, ’ribbed’: 39, ’rippled’: 39, ’thick’: 38, ’sturdy’: 36, ’striated’: 36, ’hairy’: 34, ’hazy’: 33, ’embroidered’: 32, ’raised’: 30, ’cottony’: 30, ’colorful’: 29, ’slightly compressible’: 29, ’straight’: 28, ’silky’: 28, ’braided’: 28, ’straight-edged’: 28, ’overexposed’: 27, ’angular’: 27, ’ethereal’: 27, ’glowing’: 26, ’lettered’: 25, ’tough’: 25, ’edged’: 25, ’rounded’: 25, ’transparent’: 23, ’smeared’: 23, ’carpeted’: 23, ’stretchy’: 22, ’slightly squishy’: 22, ’fleshy’: 21, ’ceramic’: 21, ’engraved’: 19, ’opaque’: 19, ’clothlike’: 19, ’bright’: 18, ’folded’: 17, ’striped’: 17, ’embossed’: 17, ’brushed’: 17, ’mesh’: 16, ’stable’: 16, ’bendable’: 16, ’slightly bendable’: 16, ’frayed’: 15, ’printed’: 15, ’vague’: 14, ’cardboard’: 14, ’clickable’: 14, ’organic’: 14, ’delicate’: 14, ’undulating’: 14, ’clear’: 13, ’stringy’: 13, ’clicky’: 13, ’smooth edges’: 13, ’sticky’: 12, ’out-of-focus’: 12, ’lace’: 11, ’brittle’: 11, ’regular’: 10, ’open’: 10, ’continuous’: 10, ’muted’: 10, ’slightly abrasive’: 10, ’malleable’: 9, ’incised’: 9, ’motion-blurred’: 9, ’slightly warm’: 9, ’intricate’: 9, ’obscure’: 9, ’laced’: 8, ’slightly curved’: 8, ’compliant’: 8, ’metal’: 7, ’sewed’: 7, ’pressed’: 7, 'flimsy': 6, 'sandy': 6, 'insulated': 6, 'convex': 6, 'sharp': 4, 'crinkled': 4, 'springy': 3, 'complex': 3, 'grainy fabric': 3, 'line': 3, 'slightly gritty': 3, 'consistent': 2, 'loose': 2, 'paper': 2, 'fraying': 2, 'lustrous': 2, 'spotty': 2, 'light': 2, 'bristly': 2, 'woolen': 2, 'wrinkled': 2, 'griany': 2, 'precise': 2, 'non-glossy': 2, 'wavy': 2, 'lacey': 1, 'meshed': 1, 'imprinted': 1, 'flat smooth': 1, 'sewn fabric': 1, 'shadow': 1, 'bendy': 1, 'rigit': 1, 'jagged': 1, 'flash': 1, 'frabric': 1, 'patterened': 1, 'floor': 1, 'flawless': 1, 'long': 1, 'spolotchy': 1, 'granulated': 1, 'cloth': 1, 'thready': 1, 'patterend': 1, 'smooth fabric': 1, 'deformalbe': 1, 'smmoth': 1, 'wirey': 1, 'fabric granular': 1, 'graint': 1, 'lined sewn': 1, 'smotth': 1, 'wiry': 1, 'torn': 1, 'vauge': 1, 'facrib': 1, 'gariny': 1, 'plain': 1, 'intertwined': 1, 'smoth': 1, 'stripped': 1, 'ragged': 1, 'denoisy': 1, 'slightly rough': 1, 'dull': 1, 'interwoven': 1, 'slightly worn': 1

C.4 Prompting for Psuedo-Label Generation

We use the following prompt with GPT-4V in order to label the images with tactile descriptions:

```
Surface Type: [Specify the surface type, e.g., "metal," "fabric"]
Images: The first image is from a camera observing the tactile sensor
    (shiny, near the top of the image) and the surface. The second
    image is a cropped version of the first image that focuses on the
    contact patch.
Example: For a smooth and cold surface, the description might be "
    slick, chilly, hard, unyielding, glossy."
Task: Based on these images, describe the possible tactile feelings of
     the contact patch using sensory adjectives. Limit your response
    up to five adjectives, separated by commas.
```

C.5 Prompting GPT-4 for Evaluation

We use the following prompt for TVL Benchmark:

```
[User Question]: {prompt}
[Assistant Response]: {assistant_response}
[Correct Response]: {correct_response}

We would like to request your feedback on the performance of an AI
    assistant in response to the user question displayed above.
The user asks the question on observing an image. The assistant's
    response is followed by the correct response.

Please evaluate the assistant's response based on how closely it
    matches the correct response which describes tactile feelings.
    Please compare only the semantics of the answers. DO NOT consider
    grammatical errors in scoring the assistant. The assistant
    receives an overall score on a scale of 1 to 10, where a higher
    score indicates better overall performance.

Please first output a single line containing only one value indicating
     the score for the assistant.

In the subsequent line, please provide a comprehensive explanation of
    your evaluation, avoiding any potential bias.
```

C.6 Improved Prompting Format

To investigate the effect of the prompting format, we conduct reference-guided grading for evaluation. In addition, to mitigate the position bias mentioned in [33], we randomly shuffle the order of the agent's response and human label on the test set. The prompt is adjusted to the following:

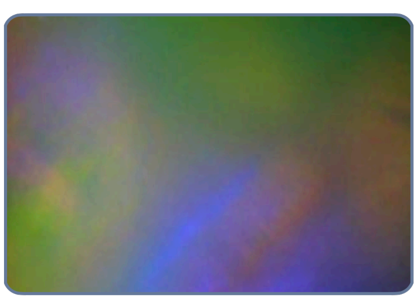
```
[User Question]: {prompt}
{assistant_response or human_label}
{human_label or assistant_response}
We would like to request your feedback on the performance of an AI
    assistant in response to the user question displayed above.
```

```
The user asks the question on observing an image. The assistant's response is followed by the correct response.
Please evaluate the assistant's response based on how closely it matches the correct response which describes tactile feelings. Please compare only the semantics of the answers. DO NOT consider grammatical errors in scoring the assistant. The assistant receives an overall score on a scale of 1 to 10, where a higher score indicates better overall performance. Please first output a single line containing only one value indicating the score for the assistant. In the subsequent line, please provide a comprehensive explanation of your evaluation, avoiding any potential bias.
Example:
[User Question]: This image gives tactile feelings of?
[Assistant Response]: fabric, grainy.
[Correct Response]: coarse, fabric, deformable.
9.5
The assistant's response is very close to the correct response. Fabric and grainy are similar to coarse and fabric. The assistant's response is missing the word deformable, which is a minor error.
Example:
[User Question]: This image gives tactile feelings of?
[Assistant Response]: flat, hard
[Correct Response]: soft, smooth, deformable
1
The assistant's response is not close to the correct response. Hard and flat are opposite to soft and smooth.
```

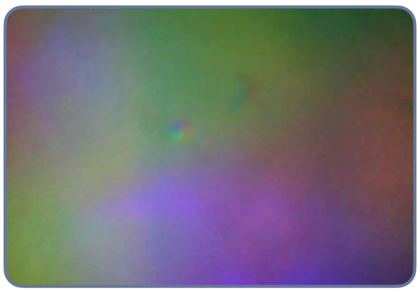
We tested TVL-LLaMA (ViT-B) with the reformed prompt. The score achieved by the model on the prompt above is similar to the prompt mentioned in Appendix C.5 used for Table 1 (5.15 v.s. 5.03) with a slightly smaller p-value (1.08e-8 v.s. 3.46e-6). We encourage future works to further investigate the effect of prompting on multimodal models.

D Generation Examples

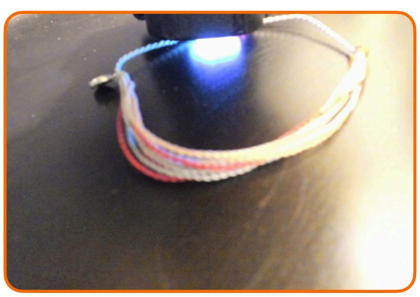
We provide a few positive and negative samples of image-tactile pairs from our dataset and the language descriptions generated for them by our various baseline models.

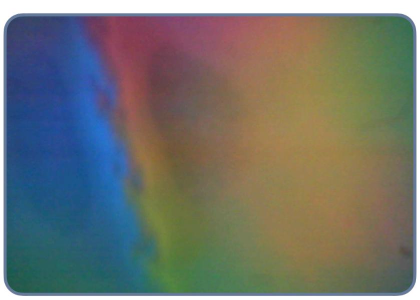


Model	Description	Score (1-10)
Human Labels	lined, fabric, sewn	GT
TVL-LLaMA	fabric, lined	7.5
GPT-4V	textured, woven, soft, flexible, durable	6
SSVTP-LLaMA	smooth, reflective, hard, cool, sleek.	1
LLaVA 1.5 7B:	Soft, grippy, textured, warm, comfortable	4
LLaVA 1.5 13B	Rough, uneven, sticky, warm, slightly damp	2
ViP-LLaVA 7B	Textured, uneven, pitted, grainy, rough###	3
ViP-LLaVA 13B	Faded, worn, creased, soft, cozy.	3
LLaMA-Adapter	1. Slick, hard, unyielding, cold, glossy.	2
BLIP-2 OPT-6.7	There is no tactile surface on this image.	1
InstructBLIP 7B	the close up picture of a blue denim jacket with button on the right side	2
InstructBLIP 13B	light blue jean jacket nothing 3	1

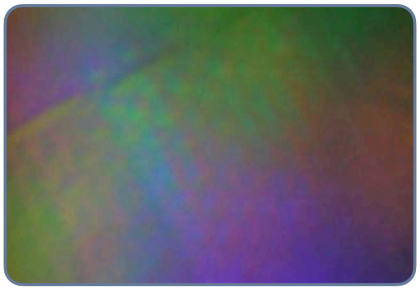


Model	Description	Score (1-10)
Human Labels	hard, flat, splotchy, grainy	GT
TVL-LLaMA	flat, hard.	6.5
GPT-4V	Textured, solid, cool, rough, uneven	4
SSVTP-LLaMA	smooth, reflective, hard, cool, sleek.	2
LLaVA 1.5 7B:	Slick, rough, cold, hard, glassy.	5
LLaVA 1.5 13B	For a white, speckled surface, the description might be "pebble, textured, matte, porous, spotty."	6.5
ViP-LLaVA 7B	Shiny, cool, smooth, hard, unyielding.	2
ViP-LLaVA 13B	Smooth, hard, cool, sturdy, firm.	4
LLaMA-Adapter	1. Slick, hard, unyielding, cold, glossy.	6.5
BLIP-2 OPT-6.7	the patch is soft, velvety, smooth and slippery	1
InstructBLIP 7B	some dice and a blue pair of jeans on the table	1
InstructBLIP 13B	a close up of a person's jeans and a dice 2	1

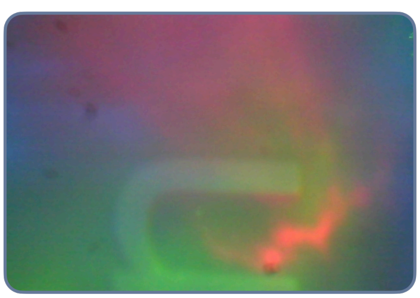




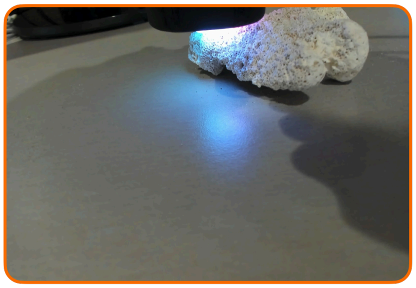
Model	Description	Score (1-10)
Human Labels	fibrous, textured, uneven, pliable, patterned	GT
TVL-LLaMA	woven, flexible, soft, glossy.	6.5
GPT-4V	smooth, reflective, hard, cool, glossy	1
SSVTP-LLaMA	smooth, reflective, hard, cool, sleek.	1
LLaVA 1.5 7B:	Cool, gritty, textured, adhesive, reflective	6.5
LLaVA 1.5 13B	For the patch where the sensor makes contact, the possible feeling could be described as "smooth, soft, textured, slightly warm, and secure."	2
ViP-LLaVA 7B	Slick, smooth, shiny, glossy, reflective	1
ViP-LLaVA 13B	Shiny, reflective, firm, smooth, cool	1
LLaMA-Adapter	1. Slick, hard, unyielding, cold, glossy.	1
BLIP-2 OPT-6.7	The sensor is shiny and there is a smooth surface.	2
InstructBLIP 7B	A multicolored bracelet sitting on top of a black table	1
InstructBLIP 13B	A wrist band with a light is on a table 81-1	1

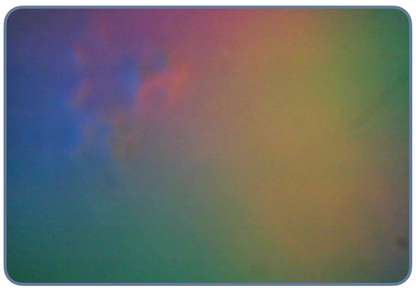


Model	Description	Score (1-10)
Human Labels	sewn, coarse, fabric, deformable	GT
TVL-LLaMA	fabric, grainy.	7.5
GPT-4V	Textured, flexible, woven, soft, uneven	7
SSVTP-LLaMA	smooth, reflective, hard, cool, sleek.	2
LLaVA 1.5 7B:	Torn, frayed, worn, stitched, black and white	6
LLaVA 1.5 13B	For the black and white striped material, the tactile sensor is making contact at its edge, creating a sense of tension or pressure, resulting in a "tight, firm, sturdy, sharp, and precise" feeling.	1
ViP-LLaVA 7B	Shiny, squishy, cloth-like, soft, textured.	4
ViP-LLaVA 13B	Textured, frayed, torn, rough, jagged.	7
LLaMA-Adapter	1. Slick, hard, unyielding, cold, glossy.	3
BLIP-2 OPT-6.7	I'm not exactly sure, but it looks like the patch has markings on it, so I guess it makes the sensor feel rough and hard	4
InstructBLIP 7B	there is a white and black striped object	2
InstructBLIP 13B	a close up of an old football uniform 1950s	1

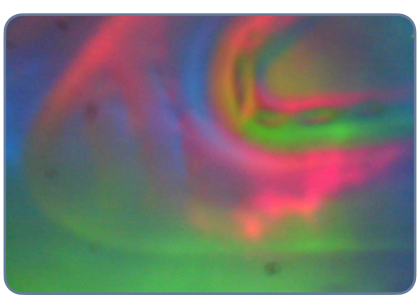


Model	Description	Score (1-10)
Human Labels	smooth, hard, sleek	GT
TVL-LLaMA	reflective, cool, sleek, smooth.	8
GPT-4V	smooth, reflective, hard, cool, sleek	8
SSVTP-LLaMA	smooth, reflective, hard, cool, sleek.	8.5
LLaVA 1.5 7B:	Slick, cold, smooth, hard, reflective.	8.5
LLaVA 1.5 13B	The surface is smooth, cold, hard, unyielding, and glossy.	8.5
ViP-LLaVA 7B	Slick, cold, smooth, shiny, reflective	7.5
ViP-LLaVA 13B	Glossy, hard, unyielding, smooth, shiny.	7.5
LLaMA-Adapter	1. Slick, hard, unyielding, cold, glossy.	8.5
BLIP-2 OPT-6.7	The surface is metallic.	6.5
InstructBLIP 7B	1 a person is using a black tablet on their lap	1
InstructBLIP 13B	a person uses a finger to press the light on a keyboard 3	1





Model	Description	Score (1-10)
Human Labels	uneven, grainy, bumpy, hard	GT
TVL-LLaMA	smooth, matte, firm, cool, flat.	1
GPT-4V	smooth, cool, hard, matte, solid	2
SSVTP-LLaMA	smooth, reflective, hard, cool, sleek.	2
LLaVA 1.5 7B:	Shiny, silky, smooth, slippery, polished.	1
LLaVA 1.5 13B	Glossy, smooth, cool, hard, and non-porous are some of the adjectives that could be used to describe the surface where the tactile sensor makes contact.	2
ViP-LLaVA 7B	Slick, gritty, textured, sharp, rough	5.5
ViP-LLaVA 13B	Slick, gritty, textured, sharp, rough	1
LLaMA-Adapter	1. Slick 2. Cold 3. Hard 4. Unyielding 5. Glossy.	4
BLIP-2 OPT-6.7	The first word that came to my mind was "sharp" because I can see the pointy edge. I also think it's very comfortable because it's	2
InstructBLIP 7B	some rocks are being exposed to some shining light	2
InstructBLIP 13B	a black light is shown shining on a piece of coral 60924	1



Model	Description	Score (1-10)
Human Labels	rough, uneven, firm, patterned, ridged	GT
TVL-LLaMA	textured, flexible, woven, soft, elastic.	4
GPT-4V	Textured, firm, woven, rough, uneven	7
SSVTP-LLaMA	smooth, reflective, hard, cool, sleek.	1
LLaVA 1.5 7B:	Glossy, smooth, and cold.	1
LLaVA 1.5 13B	For the shiny patch in the image, the description would be "shiny, reflective, smooth, bright, glimmering."	2
ViP-LLaVA 7B	Rubbery, clammy, slippery, sticky, and smudged.	2
ViP-LLaVA 13B	Slick, firm, grainy, cushioned, textured	5.5
LLaMA-Adapter	1. For a rough and hot surface, the description might be "scratchy, searing, scorching, jagged, scalding." 2. For a smooth and cold surface, the description might be "slippery, icy, hard, unyielding, glassy." 3. For a rough and hot surface, the description might be "scratchy, scorching, jagged, scalding, searing."	2
BLIP-2 OPT-6.7	"cool"	4
InstructBLIP 7B	a person holding the back of a pair of shoes with a shoe lace tie.	2
InstructBLIP 13B	a person wearing a pair of shoes on a floor 300px	2