A Touch, Vision, and Language Dataset for Multimodal Alignment

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 is an important sensing modality for humans, but it has not yet been incorporated into a multimodal generative language model. This is partially due to the difficulty of obtaining natural language labels for tactile data and the complexity of aligning tactile readings with both visual observations and language descriptions. As a step towards bridging that gap, this work introduces a new dataset of 44K in-the-wild vision-touch pairs, with English language labels annotated by humans (10%) and textual pseudo-labels from GPT-4V (90%). We use this dataset to train a vision-language-aligned tactile encoder for open-vocabulary classification and a touch-vision-language (TVL) model for text generation using the trained encoder. Results suggest that by incorporating touch, the TVL model improves (+29% classification accuracy) tactile-vision-language alignment over existing models trained on any pair of those modalities. Although only a small fraction of the dataset is human labeled, the TVL model demonstrates improved visual-tactile understanding over GPT-4V (+12%) and open-source vision-language models (+32%) on a new touch-vision understanding benchmark. Code, checkpoints and data are available on https://tactile-vlm.github.io.

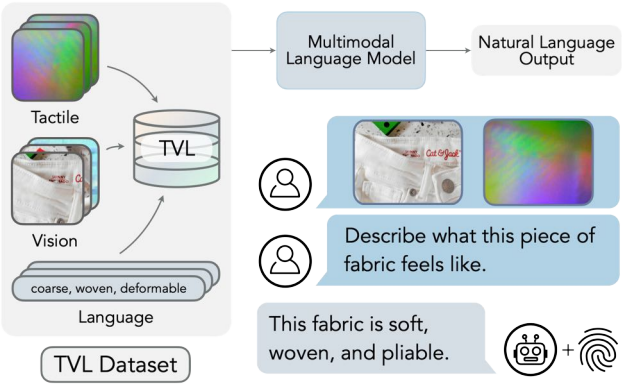


Figure 1: Can embodied agents integrate touch with vision and language? To the best of our knowledge, this work presents the first open-vocabulary tactile-vision-language dataset and we train 1) a vision-language aligned tactile encoder and 2) a tactile-vision-language model (TVLM) for describing tactile sensations.

1. Introduction

Almost all biological perception is inherently multimodal (Bertelson & De Gelder, 2004; Turk, 2014; Bruck et al., 2022), enabling agents to reason and make decisions based on multiple streams of information. Recent research in artificial multimodal representation learning has explored linking modalities such as vision, language, audio, temperature, and robot actions (Radford et al., 2021; Girdhar et al., 2023; Guzhov et al., 2021; Brohan et al., 2023; Radosavovic et al., 2023). However, the tactile modality remains underexplored in multimodal understanding. Touch enables humans to distinguish surface textures, object materials, dimensions, and contact forces (Johansson & Flanagan, 2009; Dahiya et al., 2009; Klatzky & Lederman, 2003). Tactile perception has also proven useful in robotic applications, particularly for contact-rich manipulation tasks (Lambeta et al., 2020; Dahiya et al., 2009; Calandra et al., 2018; Yuan et al., 2017; Dave et al., 2024; Qi et al., 2023).

Many works also explore visual tactile association, build cross-modal generators, and leverage cross-modal pertaining for material property, surface texture, and cloth classification on a closed set of vocabularies (Yang et al., 2022; Dave et al., 2024; Li & Adelson, 2013; Ojala et al., 2002; Kampouris et al., 2016; Yuan et al., 2018; Kerr et al., 2023).

However, human tactile perception captures *more* than tactile-visual associations; the tactile modality captures diverse semantic information and demonstrates deep integration with language (Schmidt et al., 2019; Speed et al., 2021; Miller et al., 2018; AJBarnett, 2023). One major obstacle to the integration of touch and language is the scarcity of diverse data. While recent work has collected both datasets of paired tactile and visual observations and human-labeled

[*]Equal contribution [1]UC Berkeley [2]Meta AI [3]TU Dresden. Correspondence to: Letian Fu <max.fu.letian@berkeley.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

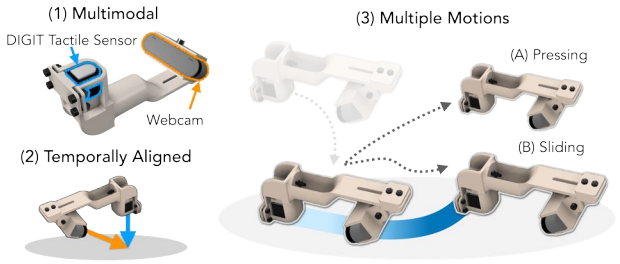


Figure 2: (1) We designed a 3D printed data collection device using the DIGIT tactile sensor and a webcam to synchronously collect tactile and vision observations "in-the-wild" (2). (3) We press and slide the device on surfaces and objects for data collection.

datasets for tactile-based texture or material classification, we are not aware of any tactile dataset that contains open vocabulary language labels. Therefore, we develop a custom hand-held device (Figure 2) for synchronized "in-the-wild" touch-vision data collection, outside of a controlled laboratory setting. This setup allows us to capture *close-up* visual observations and tactile readings while pressing and sliding on various foreground surfaces and objects with diverse backgrounds. Another challenge is that human labeling can be costly and language descriptions of tactile experiences are subjective and vary between individuals. To address these challenges, we draw inspiration from prior works on training large language models (LLMs) and vision language models (VLMs) (Taori et al., 2023; Wang et al., 2022b; Liu et al., 2023b; Chen et al., 2023b), which demonstrate vision language understanding by training on data synthesized by themselves or existing LLMs. We generate tactile descriptions from visual observations using an off-the-shelf LLM (GPT-4V (OpenAI et al., 2023)) and hypothesize that it can serve as an effective captioner to mitigate the scarcity of labeled tactile-language data.

In this work, we present the **T**ouch-**V**ision-**L**anguage (**TVL**) dataset, a novel dataset consisting of 44K paired vision-tactile observations, where 10% of the data are annotated by humans while the rest are labeled by GPT-4V. Instead of binding all modalities to vision (Girdhar et al., 2023), we train a tactile encoder on this dataset by performing pairwise contrastive learning among all three modalities. We leverage existing vision and language encoders from OpenCLIP (Ilharco et al., 2021) to train a tactile encoder that is aligned with both the textual and visual modalities. We evaluate alignment using the encoder's capability for touch-vision and touch-language classification.

Leveraging the dataset and the trained tactile encoder, we subsequently finetune LLaMA2 7B (Touvron et al., 2023) to generate textual descriptions of tactile images based on visual and tactile observations (Figure 1). To evaluate this model, we propose a Touch-Vision-Language Benchmark in which we query multimodal models to generate tactile descriptions and use an LLM to rate their consistency with ground truth human annotations.

The proposed touch-vision-language model, trained on only a small amount of human-labeled data, demonstrates statistically significant improvement in performance on the TVL Benchmark when compared to open-source VLMs (+32% improvement) and GPT-4V (+12% improvement), the label-generating model.

This paper makes the following contributions:

1. **TVL**, a new dataset containing 44K paired tactile-visual observations annotated with either human or VLM generated tactile descriptions, addressing the shortage of language-annotated tactile data;
2. **A Vision-and-Language-Aligned Tactile Encoder** trained on the TVL dataset via pairwise contrastive learning between all three modalities and **a Touch-Vision-Language Model**, a multimodal model capable of generating tactile descriptions from both visual and tactile inputs;
3. Experiments on the TVL Benchmark suggesting that a mix of human annotations and VLM pseudo-labels improves model performance in touch-vision-language understanding, surpassing existing VLMs by at least 12%.

2. Related Work

2.1. Learning Multimodal Encoders

Pretraining multi-modal encoders is a necessary step towards multi-task learning, as it can naturally structure the latent space to perform zero-shot cross-modal reasoning. CLIP (Radford et al., 2021; Ilharco et al., 2021) is among the first to utilize internet-scale data to perform contrastive pretraining to learn a joint embedding space between vision and text. Guzhov et al. (2021) and Zhang et al. (2021); Guo et al. (2023) extend CLIP to include audio and point clouds. ImageBind (Girdhar et al., 2023) contrastively trains encoders for six modalities using only image-paired data. Many works also explored masking as an alternative strategy for multimodal pretraining (Bachmann et al., 2022; Li et al., 2023b; Geng et al., 2022). In this work, we align the tactile modality with the CLIP latent space to capture its relationship with image observations and natural language descriptions of human tactility.

2.2. Tactile Perception

Integrating tactile sensation with vision, inspired by the concurrent use of sight and touch in human perception (Bresciani et al., 2006; Ittyerah & Marks, 2007; Jones et al., 2005; Camponogara & Volcic, 2021; Stone & Gonzalez, 2015), is an active area of research in both robotics and embodied AI (Goldberg & Bajcsy, 1984; Pacchierotti et al., 2017). Work in this field is facilitated by low-cost, vision-based tactile sensors (Chorley et al., 2009; Yamaguchi &

Atkeson, 2016; Yuan et al., 2017; Lambeta et al., 2020; Sferrazza & D'Andrea, 2019; Shimonomura, 2019). Several recent works find that leveraging a combination of vision and touch helps with force and sensor pose estimation (Suresh et al., 2022), cross-modal image generation and prediction (Higuera et al., 2023; Zhong et al., 2022; Yang et al., 2022; Li et al., 2019), dexterous manipulation (Calandra et al., 2018; Fu et al., 2023; Zhang & Demiris, 2023; Chen et al., 2022; Qi et al., 2023; Kerr et al., 2023), and have produced datasets that include tactile, vision, and audio data (Gao et al., 2021; 2022).

Many works study the use of tactile sensing for classifying surface textures, object material, and clothes. Li & Adelson (2013) classify 40 material properties from tactile observations using a non-learning-based texture classification method (Ojala et al., 2002); subsequent works use learning-based methods for garment classification (Kampouris et al., 2016; Yuan et al., 2018). By collecting data "in-the-wild", Yang et al. (2022) expanded the tactile observation diversity and trained a material classifier. All of these works use closed-vocabulary *human* annotations of the entire dataset, whereas we use a vision-language model to label a dataset collected "in-the-wild," and test on open-vocabulary tasks. Concurrent with this work, Yang et al. (2024) binds touch to the vision modality, conducts open-vocabulary classification across tactile, vision, and language modalities, and aligns tactile inputs with language models for text generation without finetuning ImageBind-LLM (Han et al., 2023).

2.3. Multimodal Alignment in LLMs

Pretrained multimodal encoders, when aligned with language models, enable language models to reason with non-text modalities. Based on the capabilities of Large Language Models (LLMs), Unified-IO 2 (Lu et al., 2023), Generalist Agent (Reed et al., 2022), Robot Transformer 2 (Brohan et al., 2023), and PaLM-E (Driess et al., 2023) end-to-end finetune language models with internet and visual data from multiple domains. Recent work attempts to make alignment faster and more parameter efficient (Zhu et al., 2023; Moon et al., 2023; Dai et al., 2023; Lin et al., 2023; Chen et al., 2023a; Cai et al., 2023a; Bai et al., 2023; Hu et al., 2022). Analogous to how open source language models train on GPT generated data (Taori et al., 2023), many vision-language models (Liu et al., 2023b;a; Zhang et al., 2023; Gao et al., 2023; Chen et al., 2023b) finetune the model on language-image instruction-following data generated by GPT-4 (OpenAI et al., 2023) and show general visual reasoning capabilities. ImageBind-LLM (Han et al., 2023) and PandaGPT (Su et al., 2023) introduce multimodal reasoning capability using ImageBind encoders. More recent work aligns pretrained LLMs, encoders, and decoders to finetune a model that can understand and generate multimodal data (Wu et al., 2023; Tang et al., 2023; Sun et al., 2023). Similar to Imagebind-LLM, this work aligns the multimodal encoder with a pretrained LLaMA-2 (Touvron et al., 2023).

2.4. Training from Pseudo-labels

The effectiveness of supervised learning is often limited by the availability of labeled data. Teacher models trained on a small set of labeled data can provide an inexpensive source of supervision in the form of pseudo-labels. A student model then learns from pseudo-labels generated by the teacher model on a large volume of unlabeled data (Sohn et al., 2020; Lee et al., 2013; Wang et al., 2022a; Rosenberg et al., 2005; McLachlan, 1975). While previous works leverage training teacher models on labeled datasets, recent works in both vision and language literature leverage large-scale pretrained models. CutLER (Wang et al., 2023b) uses DINO (Caron et al., 2021) features to generate bounding boxes, enabling unsupervised training of object detection and segmentation models. InstructPix2Pix and InstructNeRF2NeRF (Brooks et al., 2023; Haque et al., 2023) use GPT (Brown et al., 2020) and Stable Diffusion (Rombach et al., 2022) to generate a dataset of image editing examples and subsequently train a diffusion model based on these examples. Recent LLMs and VLMs (Wang et al., 2022b; Taori et al., 2023; Liu et al., 2023b;a) are trained using pseudo-labels generated by GPT models (Brown et al., 2020; OpenAI et al., 2023). However, in these works the teacher and student models share the same input and output modalities. Similar to the framework proposed by Burnel et al. (2023), we use a vision-only multi-modal model to generate textual labels from vision data, which in turn to match with tactile data to train the language-aligned tactile encoder and the TVL model. The teacher we use (GPT-4V) is more general than a specialist model trained on only the student task.

3. TVL Dataset

The TVL Dataset (examples in Figure 3) 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.

3.1. Data Collection

TVL uses vision data from a Logitech BRIO webcam and tactile data from DIGIT, a low-cost, compact, and open-source tactile sensor that provides high-resolution tactile observations in the form of RGB images of an internal deformable surface (Lambeta et al., 2020). The raw vision-tactile dataset amalgamates two distinct subsets: 1) the **S**elf-**S**upervised **V**isuo-**T**actile **P**retraining (SSVTP) (Kerr et al., 2023) dataset and 2) a **H**uman **C**ollected **T**actile (HCT) dataset. The SSVTP dataset (4,587 image-touch pairs) is collected by a UR5 robot, which first captures top-down images from above a work surface on which a set of ob-

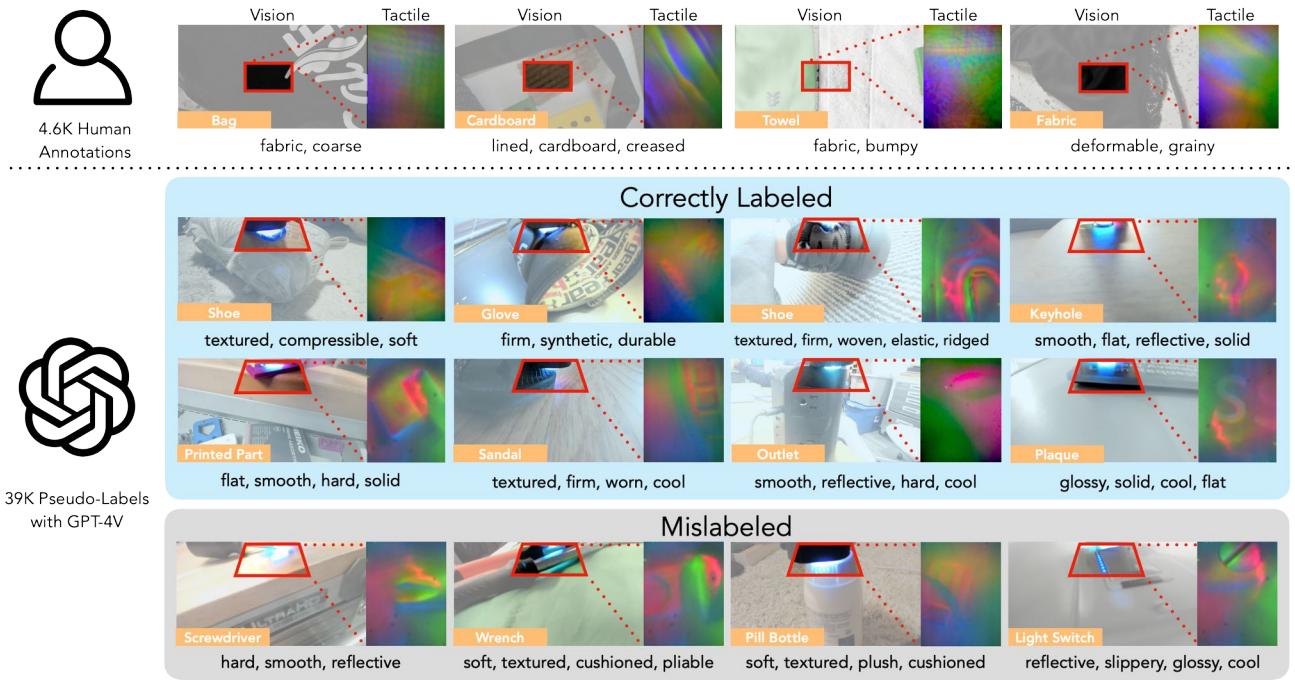


Figure 3: TVL Dataset starts by combining two datasets: SSVTP (Kerr et al., 2023) (4,587 image-touch pairs) and HCT (39,154 image-touch pairs), a new dataset we collected such that the visual observation and the tactile input are synchronously captured. For the SSVTP dataset, we then manually label the 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.

jects is prearranged, then subsequently presses the DIGIT sensor onto the corresponding location in the workspace. Nonetheless, the SSVTP dataset faces two limitations: 1) its collection in a laboratory environment restricts the diversity of objects, and 2) the asynchronous capture of tactile and visual data can result in misalignments, especially if the object is inadvertently moved by the robot during data acquisition. To address these issues, HCT emphasizes the synchronous acquisition of tactile and visual data to ensure alignment in the captured sensory information.

HCT consists of in-the-wild data visual-tactile data examples collected by 5 humans over 20 total hours using the handheld, 3D-printed data collection device featured in Figure 2. The device records both visual and tactile observations at 30 Hz. Data frames are collected in "trajectories" of touches: each trajectory consists of the human approaching, contacting, sliding, and withdrawing from an object with the tactile sensor. We categorize the touch-vision pairs as either in- or out-of-contact with the surface. The visual data are collected at an oblique angle such that the tactile sensor and point of contact are always within the field of view of the camera to preserve vision-touch synchronicity. To improve variety within this dataset, human collectors were instructed to search for interesting and novel real-world tactile examples, such as textures and edges. A small held-out test set (1% of pairs) from the HCT is hand-annotated, while the rest are pseudo-labeled by GPT-4V, as described in Section 3.3.

3.2. Cleaning Candidate Tactile Images

We categorize the collected data into in-contact and out-of-contact frames using the pretrained tactile encoder from SSVTP (Kerr et al., 2023). For every touch trajectory, under the assumption that the initial and final frames are out-of-contact, we compute an average of these frames to create a reference background image. This image is then embedded by the pretrained tactile encoder to obtain a latent representation. To determine whether a frame in a touch trajectory is in-contact, we calculate the cosine similarity between its tactile latent embedding and that of the estimated background frame. We consider a tactile frame to be in contact when the cosine similarity falls below 0.6 (Kerr et al., 2023). The collected data contains 43,741 pairs of in-contact frames and 169,292 pairs of out-of-contact frames.

3.3. Language Labeling

Human Labeling Since the SSVTP dataset demonstrates strong visual-tactile alignment, we use it as the basis for aligning touch and language as well; we manually annotate the dataset with natural language descriptions of the tactile sensations captured by each data point. We pro-

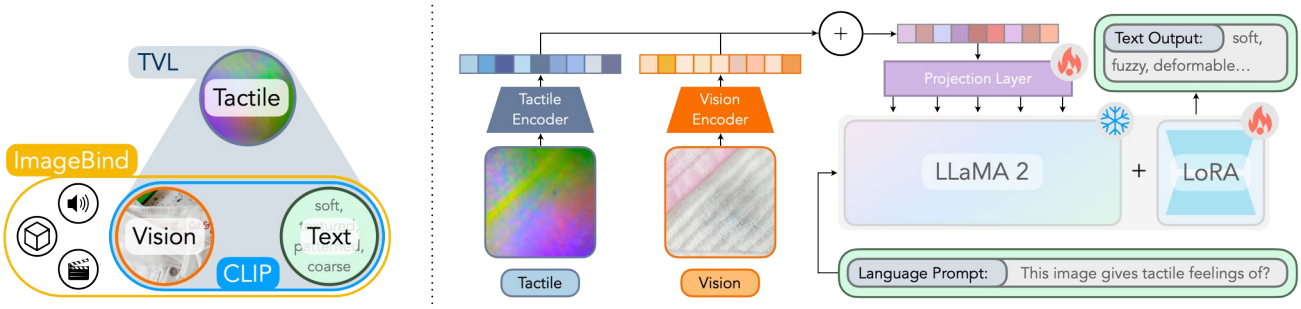


Figure 4: Method. (Left) TVL is different from ImageBind (Girdhar et al., 2023) as ImageBind only considers the loss between the vision modality and every other modality. TVL calculates loss between every pair of modalities, including that between the new modality (tactile) and language. Empirically, we show that including such loss can improve the model's capability to capture tactile semantics. (Right) Following Han et al. (2023), we average the latent from the tactile and vision modality and finetune the language model.

vide human annotators with a tactile vocabulary list of 400 words (AJBarnett, 2023) from which to generate language descriptions of the material properties and tactile feelings of pairs in the SSVTP dataset. These annotators are instructed to choose up to five applicable adjectives that most accurately describe the tactile patterns displayed in each visual-tactile pair.

Pseudo-Label Generation with GPT-4V We perform pseudo-labeling on the portion of the HCT dataset that is in contact, using GPT-4V to generate language labels describing tactile feelings. We empirically find that providing both the full image and a localized version that is cropped around the point of contact encourages GPT-4V to generate textual labels that are aligned with those of humans, as the full images may contain numerous distractors and out-of-contact objects (see success and failure cases in Figure 3). The specific prompt provided to GPT-4V for pseudo-label generation is reported in Appendix C.4.

Occasionally, GPT-4V fails or refuses to generate tactile labels for motion blurred or low lighting images. In such cases, we first attempt to generate labels for other images in the same trajectory, then populate the missing labels by randomly sampling from the set of words applied to other in-contact images within the same trajectory. If *no* image in the trajectory can successfully be labeled, that trajectory is excluded from the training portion of the dataset. After this process, we are left with 39,154 pseudo-labeled images.

3.4. Dataset Statistics

The SSVTP component contains 4,587 independent image-touch pairs. The HCT component consists of 39,154 newly-collected corresponding in-contact image-tactile frame pairs and 169,292 out-of-contact data pairs. The former dataset contains a unique touch trajectory for each data point, while the latter are collected as 1,486 unique continuous trajectories, each of which consists of one or more contact events with an object of interest. Across both the human- and GPT-4V-labeled portions of the dataset, annotators use 254 unique tactile adjectives. We perform a 99%-1% train-test split across both dataset components, with human annotators manually labeling the test set (402 image-touch pairs) for both datasets. On average, GPT-4V uses 4.25 adjectives to describe the tactile sensation on HCT, while human annotators average 2.70 adjectives. A more detailed breakdown of the descriptions is shown in Appendix C.3.

4. Tactile-Vision-Language Model

We first revisit the formulation of ImageBind and ImageBind-LLM. We then describe our pairwise contrastive approach for tactile encoder training, and finally discuss the training recipe of our aligned TVL Model.

4.1. Preliminary

ImageBind (Girdhar et al., 2023) is a multimodal model that learns a joint embedding across six different modalities: images, text, audio, depth, thermal, and IMU data. It utilizes data pairs consisting of vision and one of the other modalities, so that all are "bound" to vision. The vision and language encoders are initialized from OpenCLIP (Ilharco et al., 2021) and remain frozen, while the encoders for the other modalities are randomly initialized. Each encoder uses a small, trainable adapter network at the end to project inputs onto a latent space of the same dimension. Encoders are jointly trained through contrastive learning on the normalized latent embeddings using the InfoNCE loss.

LLaMA-Adapter (Zhang et al., 2023) and **ImageBind-LLM** (Han et al., 2023) provide efficient instruction finetuning approaches for VLMs, leveraging pretrained multimodal models to encode new modalities. The efficiency of these methods comes from (1) averaging multimodal observations in a single token and (2) a zero-initialized gate that adaptively fuses the multimodal token with the language model. LLaMA-Adapter first pretrains the zero-initialized gate and the projector from the encoder to the language model, then

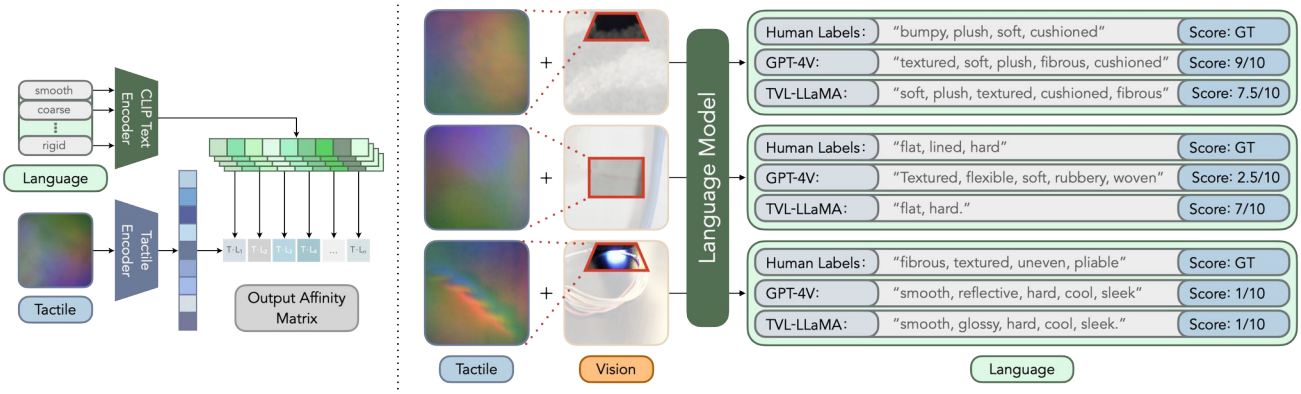


Figure 5: *Left:* We measure the cosine similarity between tactile and language on the entire test set containing 402 tactile, image, and language triplets. However, because different tactile observations may have synonymous language descriptions, in 5.1 we update top-1 and top-5 accuracy calculations to take this into account. *Right:* GPT-4V and TVL-LLaMA generations with scores rated by GPT-4 based on the human labels. GPT-4V may be distracted by objects that are not in contact as it does not take tactile into account, and we empirically found there is no improvement when including tactile observation when prompting it because the observation is out-of-distribution. As TVL-LLaMA is trained on GPT-4V pseudo-labels, it suffers from the same failure mode.

finetunes the language model with LoRA (Hu et al., 2022).

4.2. Tactile Encoder

In contrast to ImageBind, which independently 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 for each data batch. We randomly initialize the tactile encoder as a Vision Transformer (ViT) (Dosovitskiy et al., 2020) and test on three model sizes: ViT-Tiny (5.7M paraeters), ViT-Small (22M), and ViT-Base (86M). We notice that directly adopting the ImageBind training recipe leads to overfitting the relatively small training dataset of 44K pairs of in-contact data. Contrary to prior works (Kerr et al., 2023; Yang et al., 2022; Dave et al., 2024), we find that leveraging data in which the tactile sensor is not in contact with a surface (background images) can mitigate this overfitting problem and enhance tactile representation learning by improving visual data diversity (see Figure 6 in appendix). Therefore, we ensure that for a fraction $\gamma = 10\%$ of the training data, the sensor is not in contact, and we assign these examples a text label of "background". In addition, we remove the projectors from the vision and language encoders, so that the tactile encoder directly projects to the common latent space of the original CLIP. Finally, to increase the diversity of language labels, we randomly shuffle and select a subset of the words in the tactile description for each image. Together, these methods help to mitigate overfitting (refer to Appendix B.1).

4.3. Alignment with Language Models

We follow the two-stage training proposed in ImageBind-LLM (Han et al., 2023), exchanging the ImageBind encoders with TVL encoders. We pretrain on both the LLaVA Visual Instruct CC3M (Liu et al., 2023b) 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 (Taori et al., 2023) and LLaVA Visual Instruct 150K (Liu et al., 2023b). Empirically, we find that training our dataset alone is not sufficient to overcome the safety fine-tuning of LLaMA2 (Touvron et al., 2023), resulting in the model's refusal to answer questions regarding tactile sensations. Details on the prompts for TVL for instruction fine-tuning is in Appendix C.2.

	Tactile-Text		Tactile-Vision		Vision-Text	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
CLIP	-	-	-	-	28.4%	64.9%
SSVTP	-	-	0.2%	0.3%	-	-
TVL	36.7%	70.3%	79.5%	95.7%	28.4%	64.9%

Table 1: Top-1 and Top-5 Accuracy across different modality pairs. We find that the trained TVL encoder (ViT-Tiny) shows better tactile-language alignment than OpenCLIP's vision-language alignment, suggesting that vanilla CLIP may not capture tactile semantics well. Because SSVTP is trained on a subset of the TVL dataset, it does not generalize well across the entire TVL dataset, motivating the need to scale tactile-vision datasets.

5. Experiments

We quantitatively assess the multimodal capabilities of the TVL model in two experimental settings: a cross-modal classification task and a tactile-semantic description task.

5.1. Evaluation & Metrics

Open Vocabulary Tactile Classification We cast the human-labeled TVL test set as a 402-way classification problem and evaluate the tactile encoder's performance by

	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 2: 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.

measuring the top-1 and top-5 accuracy for both tactile-vision and tactile-language classification. Since many tactile observations can be described in multiple semantically similar ways (*e.g.* rigid is synonymous with stiff) and CLIP language embedding is not permutation invariant (*e.g.* "soft, smooth" and "smooth, soft" have different embeddings), we propose an alternative method to calculate the ground truth labels for tactile-language classification.

We first prompt GPT-4 to generate a set of 5 (the average length of tactile pseudo-labels) synonyms for each word in the set of descriptors used by the human annotators of the SSVTP dataset, resulting in 799 distinct adjectives describing tactile sensations. We obtain the CLIP language embedding for these adjectives and calculate the cosine similarities of each original descriptor with each of its generated synonyms. We consider the minimum ϕ of these cosine similarities to be a threshold for semantically similar vocabulary. For each tactile image, we define the set of correct language labels as all labels in the test set whose cosine similarity with the image's original language label exceeds ϕ. Using these labels, we calculate the top-1 and top-5 accuracy. Empirically, we find $\phi = 0.636$. We also report top-1 and top-5 accuracy using the 25th, 50th, and 75th percentile of the cosine similarities as the threshold in Table 6.

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 the score given, similar to prior works (Liu et al., 2023b; Chiang et al., 2023). A sample of model outputs is provided in Figure 5, and prompts used for generation and evaluation are reported in Appendix C.4. We compare against existing open-source VLMs (Liu et al., 2023a; Cai et al., 2023b; Li et al., 2023a; Dai et al., 2023) and GPT-4V. As an additional baseline, we use the SSVTP (Kerr et al., 2023) tactile and image encoder to finetune the language model; we call the resulting model SSVTP-LLaMA.

5.2. Results

Classification We summarize the tactile classification task results in Table 1. Because we use OpenCLIP to encode image and language observations, the TVL encoder shares its vision-language accuracy scores with OpenCLIP. We compare the tactile-vision accuracy of our encoder against Kerr et al. (2023); because they train on a small dataset collected in a lab setup, their model performs well on the SSVTP dataset, but does not generalize well to the new "in-the-wild" dataset. Since the tactile encoder is aligned to the language description of tactility, it shows better tactile-text alignment than OpenCLIP's vision-text alignment.

TVL Benchmark We present summary statistics for the tactile-semantic generation results in Table 2. We find that open-source VLMs perform worse than GPT-4V on the proposed benchmark, likely due to the limited diversity and

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 3: 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 Section 5.3.

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.

5.3. Ablations

This section presents six ablation and sensitivity analyses shown in Table 3 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 3a) 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 3b) resembles the setup in ImageBind (Girdhar et al., 2023), 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 (Radford et al., 2021), 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.

6. 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.

Acknowledgments

This research was supported as a BAIR Open Research Common Project with Meta. This research was performed at the AUTOLAB at UC Berkeley in affiliation with the Berkeley AI Research (BAIR) Lab, and the CITRIS "People and Robots" (CPAR) Initiative. In their academic roles at UC Berkeley, Letian Fu, Gaurav Datta, Huang Huang, William Chung-Ho Panitch, Jaimyn Drake, and Ken Goldberg are supported in part by donations from Meta, Google, Autodesk, Siemens, Toyota Research Institute, Bosch, and by equipment grants from PhotoNeo, Nvidia, and Intuitive Surgical. Roberto Calandra is funded by the German Research Foundation (DFG, Deutsche Forschungsgemeinschaft) as part of Germany's Excellence Strategy – EXC 2050/1 – Project ID 390696704 – Cluster of Excellence "Centre for Tactile Internet with Human-in-the-Loop" (CeTI) of Technische Universität Dresden, and by Bundesministerium für Bildung und Forschung (BMBF) and German Academic Exchange Service (DAAD) in project 57616814 (SECAI, School of Embedded and Composite AI). We thank Justin Kerr, Chung Min Kim, Ryan Hoque, and Xudong Wang for their helpful discussions and feedback.

Impact Statement

The data present in this paper is anonymized. This work could benefit future large generative models also considering touch as a sensing modality and can be useful for researchers studying pseudo-label-based learning methods. At the same time, the model introduced will contribute to achieving a better digitalization of touch and the use of touch in robotics. This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal benefits of our work, none of which we feel must be specifically highlighted here.

References

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

Bachmann, R., Mizrahi, D., Atanov, A., and Zamir, A. Multimae: Multi-modal multi-task masked autoencoders. *arXiv:2204.01678*, 2022.

Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., and Zhou, J. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023.

Bertelson, P. and De Gelder, B. The psychology of multimodal perception. *Crossmodal space and crossmodal attention*, pp. 141–177, 2004.

Bresciani, J.-P., Dammeier, F., and Ernst, M. O. Vision and touch are automatically integrated for the perception of sequences of events. *Journal of vision*, 6(5):2–2, 2006.

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.

Brooks, T., Holynski, A., and Efros, A. A. Instructpix2pix: Learning to follow image editing instructions, 2023.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners, 2020.

Bruck, J. N., Walmsley, S. F., and Janik, V. M. Cross-modal perception of identity by sound and taste in bottlenose dolphins. *Science Advances*, 8(20):eabm7684, 2022.

Burnel, J.-C., Courtrai, L., and Lefèvre, S. Less labels, more modalities: A self-training framework to reuse pretrained networks. In Rousseau, J.-J. and Kapralos, B. (eds.), *Pattern Recognition, Computer Vision, and Image Processing. ICPR 2022 International Workshops and Challenges*, pp. 287–302, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-37731-0.

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, 2023a.

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*, 2023b.

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.

Camponogara, I. and Volcic, R. Integration of haptics and vision in human multisensory grasping. *Cortex*, 135: 173–185, 2021.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging properties in self-supervised vision transformers, 2021.

Chen, K., Zhang, Z., Zeng, W., Zhang, R., Zhu, F., and Zhao, R. Shikra: Unleashing multimodal llm's referential dialogue magic, 2023a.

Chen, L., Li, J., Dong, X., Zhang, P., He, C., Wang, J., Zhao, F., and Lin, D. Sharegpt4v: Improving large multi-modal models with better captions, 2023b.

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

Chen, Y., Sipos, A., der Merwe, M. V., and Fazeli, N. Visuo-tactile transformers for manipulation, 2022.

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/.

Chorley, C., Melhuish, C., Pipe, T., and Rossiter, J. Development of a tactile sensor based on biologically inspired edge encoding. In *2009 International Conference on Advanced Robotics*, pp. 1–6. IEEE, 2009.

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.

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.

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

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.

Driess, D., Xia, F., Sajjadi, M. S. M., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., Huang, W., Chebotar, Y., Sermanet, P., Duckworth, D., Levine, S., Vanhoucke, V., Hausman, K., Toussaint, M., Greff, K., Zeng, A., Mordatch, I., and Florence, P. Palm-e: An embodied multimodal language model. In *arXiv preprint arXiv:2303.03378*, 2023.

Fu, L., Huang, H., Berscheid, L., Li, H., Goldberg, K., and Chitta, S. Safe self-supervised learning in real of visuo-tactile feedback policies for industrial insertion, 2023.

Gao, P., Han, J., Zhang, R., Lin, Z., Geng, S., Zhou, A., Zhang, W., Lu, P., He, C., Yue, X., Li, H., and Qiao, Y. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023.

Gao, R., Chang, Y.-Y., Mall, S., Fei-Fei, L., and Wu, J. Objectfolder: A dataset of objects with implicit visual, auditory, and tactile representations. In *Conference on Robot Learning*, 2021.

Gao, R., Si, Z., Chang, Y.-Y., Clarke, S., Bohg, J., Fei-Fei, L., Yuan, W., and Wu, J. Objectfolder 2.0: A multisensory object dataset for sim2real transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10598–10608, June 2022.

Geng, X., Liu, H., Lee, L., Schuurams, D., Levine, S., and Abbeel, P. Multimodal masked autoencoders learn transferable representations. *arXiv preprint arXiv:2205.14204*, 2022.

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.

Goldberg, K. Y. and Bajcsy, R. Active touch and robot perception. *Cognition and Brain Theory*, 7(2):199–214, 1984.

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.

Guo, Z., Zhang, R., Zhu, X., Tang, Y., Ma, X., Han, J., Chen, K., Gao, P., Li, X., Li, H., and Heng, P.-A. Point-bind and

point-llm: Aligning point cloud with multi-modality for 3d understanding, generation, and instruction following, 2023.

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

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.

Haque, A., Tancik, M., Efros, A., Holynski, A., and Kanazawa, A. Instruct-nerf2nerf: Editing 3d scenes with instructions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Higuera, C., Boots, B., and Mukadam, M. Learning to read braille: Bridging the tactile reality gap with diffusion models. 2023.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., and Schmidt, L. Openclip, July 2021. URL https://doi.org/10.5281/zenodo.5143773. If you use this software, please cite it as below.

Ittyerah, M. and Marks, L. E. Memory for curvature of objects: Haptic touch vs. vision. *British Journal of Psychology*, 98(4):589–610, 2007.

Johansson, R. S. and Flanagan, J. R. Coding and use of tactile signals from the fingertips in object manipulation tasks. *Nature Reviews Neuroscience*, 10(5):345–359, 2009.

Jones, M. G., Bokinsky, A., Tretter, T., and Negishi, A. A comparison of learning with haptic and visual modalities. 2005.

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.

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.

Klatzky, R. L. and Lederman, S. J. The skin and its receptors 148 pathways to cortex and major cortical areas. *Handbook of psychology, experimental psychology*, 4:147, 2003.

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.

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.

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

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.

Li, Y., Zhu, J.-Y., Tedrake, R., and Torralba, A. Connecting touch and vision via cross-modal prediction, 2019.

Li, Y., Fan, H., Hu, R., Feichtenhofer, C., and He, K. Scaling language-image pre-training via masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23390–23400, 2023b.

Lin, Z., Liu, C., Zhang, R., Gao, P., Qiu, L., Xiao, H., Qiu, H., Lin, C., Shao, W., Chen, K., Han, J., Huang, S., Zhang, Y., He, X., Li, H., and Qiao, Y. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models, 2023.

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

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

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

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

Lu, J., Clark, C., Lee, S., Zhang, Z., Khosla, S., Marten, R., Hoiem, D., and Kembhavi, A. Unified-io 2: Scaling autoregressive multimodal models with vision, language, audio, and action, 2023.

McLachlan, G. J. Iterative reclassification procedure for constructing an asymptotically optimal rule of allocation in discriminant analysis. *Journal of the American Statistical Association*, 70(350):365–369, 1975.

Miller, T. M., Schmidt, T. T., Blankenburg, F., and Pulvermüller, F. Verbal labels facilitate tactile perception. *Cognition*, 171:172–179, 2018.

Moon, S., Madotto, A., Lin, Z., Nagarajan, T., Smith, M., Jain, S., Yeh, C.-F., Murugesan, P., Heidari, P., Liu, Y., Srinet, K., Damavandi, B., and Kumar, A. Anymal: An efficient and scalable any-modality augmented language model, 2023.

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.

OpenAI, :, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.-L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H. W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S. P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S. S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Łukasz Kaiser, Kamali, A., Kanitscheider, I., Keskar, N. S., Khan, T., Kilpatrick, L., Kim, J. W., Kim, C., Kim, Y., Kirchner, H., Kiros, J., Knight, M., Kokotajlo, D., Łukasz Kondraciuk, Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C. M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S. M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O'Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., de Avila Belbute Peres, F., Petrov, M., de Oliveira Pinto, H. P., Michael, Pokorny, Pokrass, M., Pong, V., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F. P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J. F. C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J. J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., and Zoph, B. Gpt-4 technical report, 2023.

Pacchierotti, C., Sinclair, S., Solazzi, M., Frisoli, A., Hayward, V., and Prattichizzo, D. Wearable haptic systems for the fingertip and the hand: Taxonomy, review, and perspectives. *IEEE Transactions on Haptics*, 10(4):580–600, 2017. doi: 10.1109/TOH.2017.2689006.

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.

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.

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.

Reed, S., Zolna, K., Parisotto, E., Colmenarejo, S. G., Novikov, A., Barth-Maron, G., Gimenez, M., Sulsky, Y., Kay, J., Springenberg, J. T., Eccles, T., Bruce, J., Razavi, A., Edwards, A., Heess, N., Chen, Y., Hadsell, R., Vinyals, O., Bordbar, M., and de Freitas, N. A generalist agent, 2022.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models, 2022.

Rosenberg, C., Hebert, M., and Schneiderman, H. Semi-supervised self-training of object detection models. 2005.

Schmidt, T. T., Miller, T. M., Blankenburg, F., and Pulvermüller, F. Neuronal correlates of label facilitated tactile perception. *Scientific Reports*, 9(1):1606, 2019.

Sferrazza, C. and D'Andrea, R. Design, motivation and evaluation of a full-resolution optical tactile sensor. *Sensors*, 19(4):928, 2019.

Shimonomura, K. Tactile image sensors employing camera: A review. *Sensors*, 19(18):3933, 2019.

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.

Speed, L. J., Croijmans, I., Dolscheid, S., and Majid, A. Crossmodal associations with olfactory, auditory, and tactile stimuli in children and adults. *i-Perception*, 12(6): 20416695211048513, 2021.

Stone, K. D. and Gonzalez, C. L. The contributions of vision and haptics to reaching and grasping. *Frontiers in psychology*, 6:1403, 2015.

Su, Y., Lan, T., Li, H., Xu, J., Wang, Y., and Cai, D. Pandagpt: One model to instruction-follow them all, 2023.

Sun, Q., Cui, Y., Zhang, X., Zhang, F., Yu, Q., Luo, Z., Wang, Y., Rao, Y., Liu, J., Huang, T., and Wang, X. Generative multimodal models are in-context learners, 2023.

Suresh, S., Si, Z., Anderson, S., Kaess, M., and Mukadam, M. Midastouch: Monte-carlo inference over distributions across sliding touch. In *6th Annual Conference on Robot Learning*, 2022. URL https://openreview.net/forum?id=JWROnOf4w-K.

Tang, Z., Yang, Z., Khademi, M., Liu, Y., Zhu, C., and Bansal, M. Codi-2: In-context, interleaved, and interactive any-to-any generation, 2023.

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.

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.

Turk, M. Multimodal interaction: A review. *Pattern recognition letters*, 36:189–195, 2014.

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*, 2023a.

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.

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, 2022a.

Wang, X., Girdhar, R., Yu, S. X., and Misra, I. Cut and learn for unsupervised object detection and instance segmentation, 2023b.

Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Khashabi, D., and Hajishirzi, H. Self-instruct: Aligning language model with self generated instructions, 2022b.

Wu, S., Fei, H., Qu, L., Ji, W., and Chua, T.-S. Next-gpt: Any-to-any multimodal llm. *CoRR*, abs/2309.05519, 2023.

Yamaguchi, A. and Atkeson, C. G. Combining finger vision and optical tactile sensing: Reducing and handling errors while cutting vegetables. In *2016 IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids)*, pp. 1045–1051, 2016. doi: 10.1109/HUMANOIDS.2016.7803400.

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.

Yang, F., Feng, C., Chen, Z., Park, H., Wang, D., Dou, Y., Zeng, Z., Chen, X., Gangopadhyay, R., Owens, A., et al. Binding touch to everything: Learning unified multimodal tactile representations. *arXiv preprint arXiv:2401.18084*, 2024.

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

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.

Zhang, F. and Demiris, Y. Visual-tactile learning of garment unfolding for robot-assisted dressing. *IEEE Robotics and Automation Letters*, 8(9):5512–5519, 2023. doi: 10.1109/LRA.2023.3296371.

Zhang, R., Guo, Z., Zhang, W., Li, K., Miao, X., Cui, B., Qiao, Y., Gao, P., and Li, H. Pointclip: Point cloud understanding by clip, 2021.

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

Zhong, S., Albini, A., Jones, O. P., Maiolino, P., and Posner, I. Touching a neRF: Leveraging neural radiance fields for tactile sensory data generation. In *6th Annual Conference on Robot Learning*, 2022. URL https://openreview.net/forum?id=No3mbanRlZJ.

Zhu, D., Chen, J., Shen, X., Li, X., and Elhoseiny, M. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

A. Additional Results

A.1. Performance Per Dataset

In this section, we show a fine-grained breakdown of Table 2 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 4 and the performance on the HCT subset is listed in Table 5. Results suggest that GPT-4V performs better on SSVTP, which is collected in a lab setting, than HCT, which is collected "in-the-wild".

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 5 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 4: 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.

A.2. Open Vocabulary Tactile Classification Full Result

We present the result presented in Table 1 in Table 6 and Table 7 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 7), ViT-Base performs outperforms the smaller models. More insights are detailed in Appendix B.1.

	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 5: 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. Additional Open Vocabulary Downstream Tasks

In the tactile classification experiment in Table 1, 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.*

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 6: 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 7: Effect of Tactile Encoder Model Architecture on **Tactile-Vision** Classification.

Sohn et al. (2020); Lee et al. (2013); Wang et al. (2022a)). 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 6).

In 4.2, 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 Table 1). As an ablation study, we also finetuned the ViT-Small trained only on in-contact data for tactile language generation. The test set performance 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.

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 8).

	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 8: 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 9. 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 3a. Interestingly, background subtraction (in Ap-

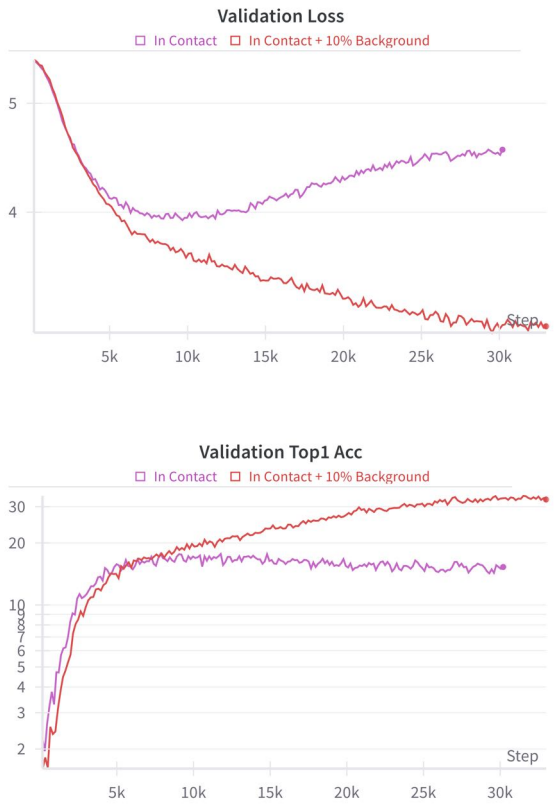


Figure 6: Overfitting is significant when all data is in contact. When 10% not in contact data is added, the overfitting issue is addressed.

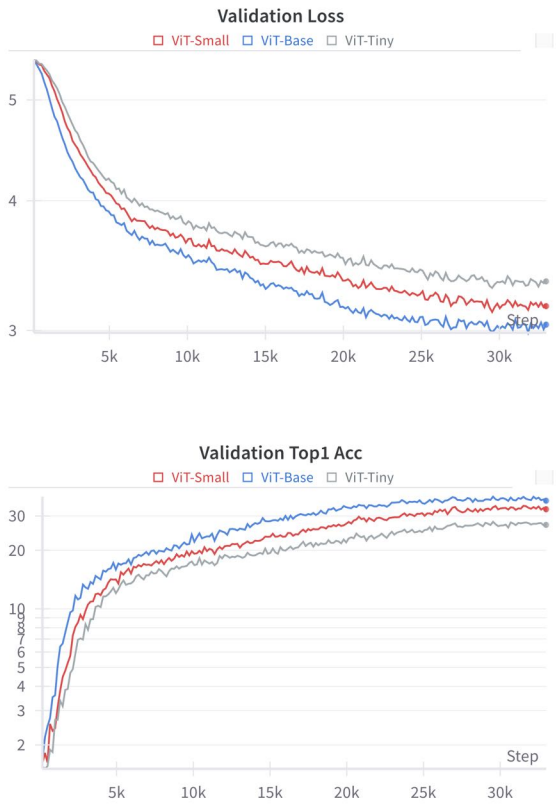


Figure 7: While we find that the model scales on the dataset, the test set performance does not align with the validation set performance. One potential cause of this is distribution shift: the validation set uses pseudo-labels generated by GPT-4V, while the test set is human-labeled.

pendix 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 9: 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 2). 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 10 and Table 11.

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 10: Tactile Normalization Statistics

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

Table 11: RGB Normalization Statistics

B.6. TVL Tactile Encoder Hyperparameters

All of ViT-Tiny, ViT-Small, and ViT-Base share the same hyperparameters (see Table 12). All experiments are run on a single NVIDIA A100 GPU.

B.7. TVL-LLaMA Hyperparameters

We follow the hyperparameter setup in ImageBind-LLM (Han et al., 2023). 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 12).

C. Dataset

C.1. Hardware

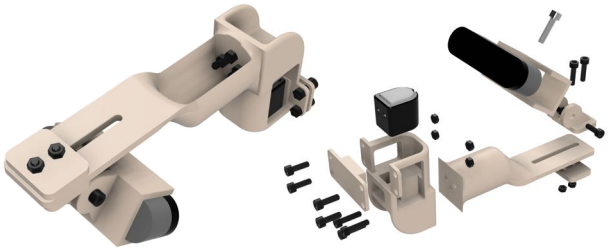


Figure 8: Alternative perspectives of the sensor holder CAD model: face-down view (left) and exploded view (right).

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. 8, 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 (Liu et al., 2023b). 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
```

Config	Value
optimizer	AdamW (Loshchilov & Hutter, 2017b)
base learning rate	1.5e-4
learning rate schedule	cosine decay (Loshchilov & Hutter, 2017a)
batch size	256
weight decay	0.05
optimizer momentum	β_1, β_2 = 0.9, 0.95 (Chen et al., 2020)
warm up epoch (Goyal et al., 2017)	10
total epochs	200
RGB Augmentation	RandomHorizontalFlip, ColorJitter, RandomGrayscale, GaussianBlur
Tactile Augmentation	(Optional) Background Subtraction

Table 12: Encoder Pretraining Hyperparameters

```
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 9.

'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': 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, 'clothlike': 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,

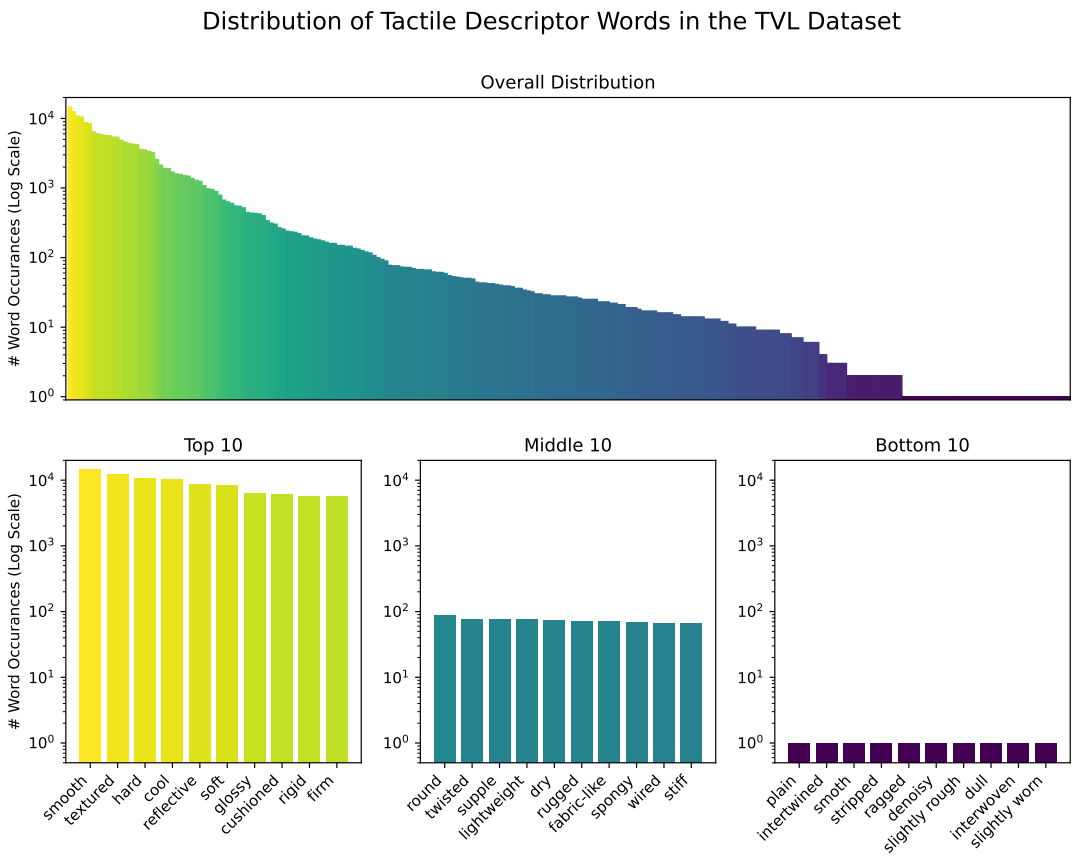


Figure 9: Distribution of Words in the TVL Dataset: The TVL dataset contains 254 unique tactile descriptors, ranging from common tactile descriptions (smooth, hard, firm) to unusual and optical descriptors. These less-common adjectives include a small fraction of misspellings and non-tactile descriptors which were generated by the VLM. The long-right-tailed distribution common in image classification (Wang et al., 2020) presents a challenge for learning predictors on tactile-semantic data as well.

'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 (Wang et al., 2023a), we randomly shuffle the order of the agent's response and human label on the test set. The prompt is adjusted to the following:

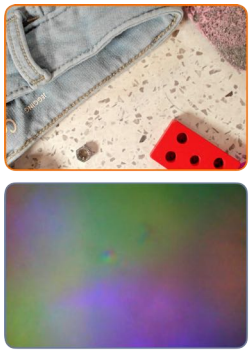
```
[User Question]: {prompt}
{assistant_response or human_label}
{human_label or assistant_response}
We would like to request your feedback on the performance of an AI assistant in response to the user question displayed above.
The user asks the question on observing an image. The assistant's response 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 2 (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

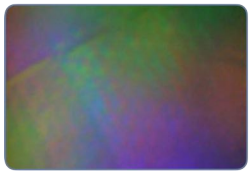
We provide a few positive and negative samples of image-tactile pairs from our dataset and the language descriptions generated for them by our various baseline models.

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

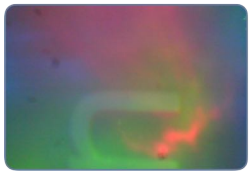


Model	Description	Score (1-10)
Human Labels	hard, flat, splotchy, grainy	GT
TVL-LLaMA	flat, hard.	6.5
GPT-4V	Textured, solid, cool, rough, uneven	4
SSVTP-LLaMA	smooth, reflective, hard, cool, sleek.	2
LLaVA 1.5 7B:	Slick, rough, cold, hard, glassy.	5
LLaVA 1.5 13B	For a white, speckled surface, the description 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

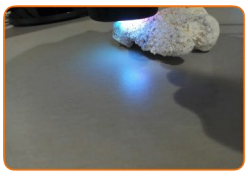
Model	Description	Score (1-10)
Human Labels	fibrous, textured, uneven, pliable, patterned	GT
TVL-LLaMA	woven, flexible, soft, glossy.	6.5
GPT-4V	smooth, reflective, hard, cool, glossy	1
SSVTP-LLaMA	smooth, reflective, hard, cool, sleek.	1
LLaVA 1.5 7B:	Cool, gritty, textured, adhesive, reflective	6.5
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



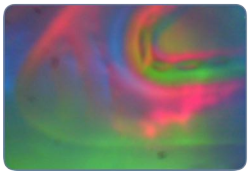
Model	Description	Score (1-10)
Human Labels	sewn, coarse, fabric, deformable	GT
TVL-LLaMA	fabric, grainy.	7.5
GPT-4V	Textured, flexible, woven, soft, uneven	7
SSVTP-LLaMA	smooth, reflective, hard, cool, sleek.	2
LLaVA 1.5 7B:	Torn, frayed, worn, stitched, black and white	6
LLaVA 1.5 13B	For the black 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