

Binding Touch to Everything: Learning Unified Multimodal Tactile Representations

Fengyu Yang^{1*} Chao Feng^{2*} Ziyang Chen^{2*} Hyungseob Park¹ Daniel Wang¹ Yiming Dou²
Ziyao Zeng¹ Xien Chen¹ Rit Gangopadhyay¹ Andrew Owens² Alex Wong¹

¹Yale University ²University of Michigan

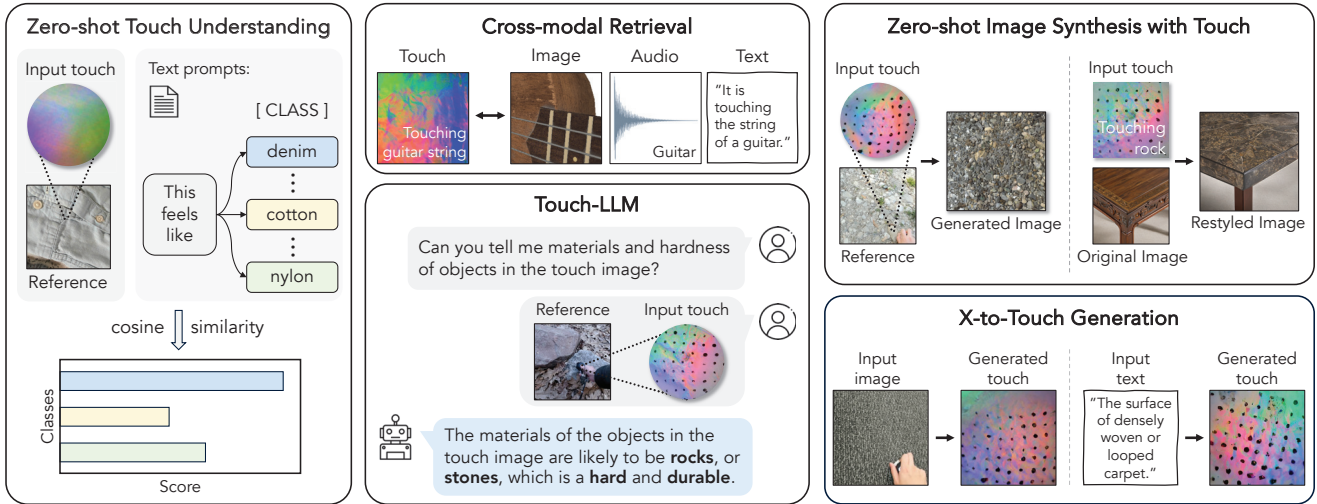


Figure 1. **Putting touch “in touch” with other modalities.** We show that a variety of tactile sensing tasks, ranging from tactile image understanding to image synthesis with touch, can be solved zero-shot by aligning touch to pretrained multimodal models, extending previous approaches on work on other modalities [41]. Our learned model can be applied to various vision-based tactile sensors and simulators (e.g., GelSight, DIGIT, Taxim, and Tacto). For visualization purposes, we show the corresponding visual signal (labeled “reference”) for each touch signal, even though it is not used by the model.

Abstract

Touch provides crucial information about the physical properties of the objects around us. Creating models that capture cross-modal associations between touch and other modalities, however, remains a challenging problem, due to wide variety of touch sensors and the intensive effort required to collect tactile data. We propose UniTouch, a unified model for vision-based touch sensors that connects their tactile signals to other modalities, including vision, language, and sound. We achieve this by aligning our tactile embeddings to pretrained image embeddings already associated with a variety of other modalities. We further propose learnable sensor-specific tokens, allowing the model to learn from a set of heterogeneous tactile sensors, all at the same time. UniTouch is capable of conducting various touch sensing tasks in a zero-shot setting, from robot grasping prediction to touch-based question answering. To the best of our knowledge, UniTouch is the first model to demonstrate these capabilities. Project

* Indicates equal contribution.

Page: <https://cfengl6.github.io/UniTouch/>.

1. Introduction

Amongst our five main senses, touch sensing is perhaps the most crucial to human survival, due to its role in perceiving physical contact — rivaling even vision in its overall importance [53, 83, 91]. Our ability to form cross-modal associations between touch and our other senses [106] thus underlies a great deal of our physical capabilities. For example, we predict from vision how a surface will feel before we touch it, and we predict from touch how an object will sound before we strike it. These cross-modal associations are also a key component of computational systems, such as for robotic manipulation [8, 73, 85, 89, 90, 95, 99, 103, 124, 132, 134], material and geometry estimation [10, 44, 111, 129, 137], assistive technology [49], and texture recognition [57, 88, 136].

Despite their importance, cross-modal associations be-

tween touch and other modalities have received considerably less attention from the multimodal research community than those of other modalities, such as vision, language, and sound. Touch is expensive to acquire [36, 38, 129] as it requires actively probing objects with touch sensors, limiting the scale of data collected for training tactile “foundation” models. Moreover, touch sensors are not fully standardized, and thus there are large differences between outputs of different sensors [37, 139]. Even amongst the commonly used vision-based sensors, the difference in mechanical design and elastomeric material will lead to divergent artifacts, limiting generalization (Fig. 2). As a result, existing tactile representations are typically constrained to a single sensor.

An emerging line of work has addressed the challenges of learning from other low-resource modalities, like sound, point clouds, and depth, by aligning examples with pre-trained vision-language embeddings [41, 72, 127]. In our paper, we show that this approach can be adapted to tactile sensing. We align tactile signals to visual signals, thereby linking touch to a variety of other modalities, such as language and sound. Then we can use the representations within off-the-shelf models trained on other modalities (*e.g.*, CLIP [101]), to solve different tactile sensing tasks. To deal with the large variations in different touch sensors, we train a single model with multiple tactile signals at once, and introduce learnable tokens to model sensor-specific properties, such as the calibration and intensity profiles in the touch signal.

Our trained model, which we call **UniTouch**, is a general-purpose interface for multiple vision-based tactile sensors. Our model unifies many previously studied tactile sensing tasks “zero shot” and greatly expands the range of tasks that touch sensing can be applied, as shown in Fig. 1: (i) We apply it to zero-shot touch understanding tasks like material recognition and robotic grasp stability prediction. (ii) We obtain strong performance in cross-modal retrieval with touch by aligning touch with other modalities in a shared latent space. (iii) The learned representation can also support image synthesis tasks, including touch-to-image generation [79, 130] and tactile-driven image stylization [129, 130], by using it within off-the-shelf text-to-image diffusion models. (iv) We combine touch with large language models (LLM), allowing us to perform tasks such as tactile question answering in a variety of tactile domains, including contact localization, grasping stability prediction, and *etc.* (v) Finally, we perform “X-to-touch” generation, producing touch images from vision, text, and audio. Our experiments suggest our zero-shot model achieves competitive (or even better) performance than previously proposed approaches on multiple tasks.

2. Related Work

Tactile sensing. Early tactile sensors were chiefly engineered to register fundamental, low-dimensional sensory outputs such as force, pressure, vibration, and tempera-

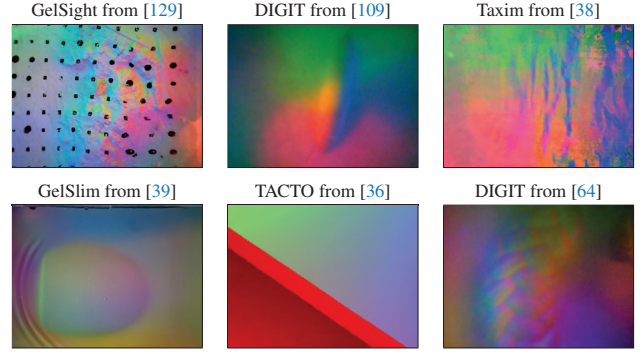


Figure 2. **Tactile images of different sensors and datasets.** In contrast to many other modalities, signals from different touch sensing hardware exhibit large amounts of variation.

ture [20, 63, 69, 70]. Lately, there has been a growing focus on vision-based tactile sensors. GelSight [61, 135] as one of the representative sensors, features an elastomeric gel with an embedded camera and illumination system. The gel deforms upon contact with an object and creates a high-resolution height map using photometric stereo [62], which provides detailed information about the shape and physical properties of touch [74, 113]. One variant, DIGIT [66], has a specially designed silicone-based elastomer gel with a harder surface and a different illumination system. Another variant GelSlim [113] contains a stretchy, loose-weave fabric gel surface. Recent work also turns into the simulation of tactile sensors [1, 18, 42, 60, 105, 117]. Taxim [105] simulates the optical response of a GelSight sensor and TACTO [117] calculates the local contact geometry and the corresponding rendering. We focus on vision-based sensors as they are widely available in visuo-tactile datasets [16, 27, 33, 109, 116, 125, 136], are commonly used in various applications [9, 11, 12, 22, 48, 52, 58, 67, 75, 76, 80, 98, 110, 133, 148], and all adopt image as the output format. While these vision-based tactile sensors and simulators share similar imaging patterns, the difference in design and calibration results in a significant domain gap (Fig. 2). Hence, researchers typically study each sensor separately. In our work, we introduce a novel approach to understanding multiple sensors through our unified touch encoder.

Representation learning with touch. The initial efforts learn tactile representations for specific tasks [35, 71, 82, 112, 136]. Lee *et al.* [71] undertook a collaborative training of Convolutional Neural Networks (CNN) for an RGB camera and a force sensor to facilitate contact-rich manipulation tasks. Similarly, Yuan *et al.* [136] employed a comparable methodology to establish a shared latent space between visual and tactile modalities using the Gelsight touch sensor, aimed at precise fabric classification. Recently, researchers have learned general representations of touch through self-supervision. Yang *et al.* [129] learned tactile representations for Gelsight sensors with visuo-tactile contrastive multiview coding [114] and Kerr *et al.* [64] proposed a contrastive pre-

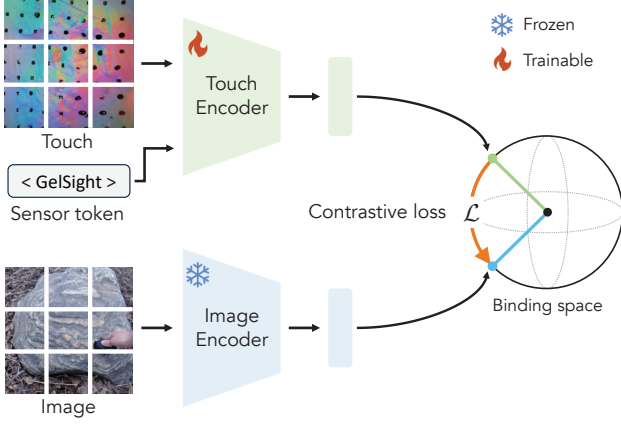


Figure 3. **Method overview.** We align our touch embedding with a pre-trained image embedding derived from large-scale vision language data, using sensor-specific tokens for multi-sensor training.

training method for the DIGIT sensor. Other works adopted BYOL framework [45] or contrastive predictive coding [138] to learn representations for non vision-based tactile sensors like BioTac. Some work [59] applies masked autoencoders to learn tactile representations directly from tactile inputs. Unlike methods concentrated solely on visuo-tactile learning for a single sensor, our approach aims to learn touch representations that can be applied across various sensors and interconnected with multiple modalities.

Multimodal representation learning. The success of vision-language pretraining [25, 87, 96, 100, 102, 123, 147] has demonstrated the ability to bridge the gap between visual content, such as images or videos, and textual descriptions [32, 55, 56, 77, 81, 126]. Furthermore, some researchers have extended the multimodal learning into the 3D domain [43, 68, 97, 120, 140–143, 149]. Some works learn shared audio-visual representation [2, 13, 28, 31, 51, 92, 94, 108, 122] by leveraging natural correspondence with videos. Some works also study shared audio-language representation [29, 46, 119]. Bender *et al.* [4] crafted an embedding space for the flavors of wines by leveraging both image and text annotations. Chen *et al.* [15] learned shared spatial information from binaural sound and vision. Some works learned the association between vision and metadata [14, 118, 146]. Imagebind [41] proposed to learn a joint embedding for six diverse modalities solely through image alignment and emerge zero-shot cross-modal capabilities. In our work, we extend this concept to the sense of touch and bind it to other modalities including text and audio by aligning tactile data with images, encouraging a more comprehensive understanding of cross-modal touch interactions without paired data.

3. Method

We aim to learn a unified tactile representation for different touch sensors that captures relationships between touch and different modalities, *e.g.* vision, text, and audio. First, we

present our contrastive visuo-tactile pretraining, inspired by [41], that can emerge interconnections of touch and other modalities. We then introduce our touch encoder design and data sampling strategy that can be used for different tactile sensors at once. Finally, we show how our learned representation can be applied to various downstream tasks.

3.1. Binding touch with images

We learn a multimodal tactile representation from touch and vision solely, without the need for paired text and audio data for touch. We achieve that by aligning our touch embedding to a pre-trained image embedding using contrastive learning as shown in Fig. 3, where the image embedding is already aligned with modalities like language and audio training from large-scale image-paired datasets [41].

We denote Ω_v as the visual image domain and Ω_t as the tactile image domain. Thus, given B visual and touch pairs in a batch, $\{(\mathbf{v}_i, \mathbf{t}_i)\}_{i=1}^B$, where $\mathbf{v}_i : \Omega_v \subset \mathbb{R}^2 \rightarrow \mathbb{R}^3$ and $\mathbf{t}_i : \Omega_t \subset \mathbb{R}^2 \rightarrow \mathbb{R}^3$, we align a tactile embedding $\mathcal{F}_T(\mathbf{t}_i) \in \mathbb{R}^C$ with the pretrained visual embedding $\mathcal{F}_V(\mathbf{v}_i) \in \mathbb{R}^C$ from [41] by maximizing the cosine similarity between corresponding visuo-tactile pairs. We optimize this objective using InfoNCE loss [93] to match touches to correct images:

$$\mathcal{L}_{T \rightarrow V} = -\frac{1}{B} \sum_{i=1}^B \log \frac{\exp(\mathcal{F}_T(\mathbf{t}_i) \cdot \mathcal{F}_V(\mathbf{v}_i) / \tau)}{\sum_{j=1}^B \exp(\mathcal{F}_T(\mathbf{t}_i) \cdot \mathcal{F}_V(\mathbf{v}_j) / \tau)}, \quad (1)$$

where τ is a temperature hyperparameter [121] and C is feature dimension. Analogously, we can also match from image $\mathbf{v}_i$ to touch $\mathbf{t}_i$ using the loss $\mathcal{L}_{V \rightarrow T}$. Thus, we minimize the overall loss:

$$\mathcal{L} = \mathcal{L}_{T \rightarrow V} + \mathcal{L}_{V \rightarrow T}. \quad (2)$$

Naturally, minimizing the contrastive objective [31, 114, 128, 146] will “pull” a visuo-tactile pair close together and “push” it away from other pairs, achieving the alignment between touch and visual embedding. As the visual embedding comes from a learned joint space that has already aligned with different modalities, touch that is bound with images will bridge a connection to other modalities, yielding a multi-modal unified tactile representation.

3.2. Learning from multiple sensors at once

We want to learn a generalizable tactile representation that will be suitable for different tactile sensors. Therefore, we designed our touch encoder $\mathcal{F}_T$ to bridge the domain gap among various vision-based tactile sensors caused by the difference in sensor designs.

Specifically, we introduce a set of learnable sensor-specific tokens $\{\mathbf{s}_k\}_{k=1}^K$, where $\mathbf{s}_k \in \mathbb{R}^{L \times D}$, to capture specific details for each sensor, *e.g.*, calibration and background color in touch images, so that the remaining model capacity can be used to learn common knowledge across

different type of touch sensors, such as texture and geometry. Here, K represents the number of sensors we train on, L is the number of sensor-specific tokens for each sensor, and D is the token dimension. For the given touch image $\mathbf{t}_i$, and its corresponding tactile sensor tokens $\mathbf{s}_{t_i}$, we append these sensor-specific tokens as prefixes to touch image patch tokens and then encode them with our touch encoder resulting in the final embedding $\mathcal{F}_T(\mathbf{t}_i, \mathbf{s}_{t_i})$ (Fig. 3). For our contrastive vision-touch pretraining, we optimize:

$$\mathcal{L}_{T \rightarrow V} = -\frac{1}{B} \sum_{i=1}^B \log \frac{\exp(\mathcal{F}_T(\mathbf{t}_i, \mathbf{s}_{t_i}) \cdot \mathcal{F}_V(\mathbf{v}_i)/\tau)}{\sum_{j=1}^B \exp(\mathcal{F}_T(\mathbf{t}_i, \mathbf{s}_{t_i}) \cdot \mathcal{F}_V(\mathbf{v}_j)/\tau)}, \quad (3)$$

as well as $\mathcal{L}_{V \rightarrow T}$ from the other direction.

In-batch data sampling. We found that batch sampling strategy [19] plays an important role when we train with data, acquired by multiple touch sensors, using contrastive learning. The model will under-perform if we randomly sample from each data source [131] which results in a surplus of easy negatives due to the domain gap between different sensors. Therefore, we design a batch sampling strategy to guarantee that σ percent of training examples in a batch are sampled from the same datasets. Given that our dataset $\mathcal{D}$ is the union over N datasets collected with diverse tactile sensors $\mathcal{D} = \bigcup_{n \in \{1, 2, \dots, N\}} \mathcal{D}_n$, the probability of selecting a given dataset $\mathcal{D}_n$ to sample from is defined as:

$$p_n = \frac{\|\mathcal{D}_n\|}{\sum_{m=1}^N \|\mathcal{D}_m\|}, \quad (4)$$

where $\|\cdot\|$ denotes cardinality. $\mathcal{D}_\sigma$ denotes the selected dataset from which we perform uniform random sampling to yield $\sigma \cdot B$ examples; the rest $(1 - \sigma) \cdot B$ examples are uniformly sampled from other datasets, *i.e.*, $\mathcal{D} \setminus \mathcal{D}_\sigma$, where σ is a hyperparameter range from 0 to 1 representing the portion of the batch. This batch sampling strategy significantly benefits our training as it allows the model to mostly focus on intra-sensor hard negatives but still be exposed to different sensors to enhance inter-sensor discrimination.

Inference. To generalize our learned representation to unseen types of sensors during the inference, we retrieve the nearest neighbor sensor-specific tokens from the learned sensor set $\{\mathbf{s}_k\}_{k=1}^N$. Specifically, we first compute a prototype for each sensor, a 1D vector that averages all the raw pixels belonging to the tactile images collected by this sensor, and store these prototypes after training. Then, during the inference stage, we compute the L1 distance between an input tactile image and all the sensor prototypes and retrieve the sensor with minimum distance.

3.3. Applications

By aligning our touch embedding to the joint latent space, we establish a link between touch and other modalities. These alignments allow us to perform various zero-shot and cross-modal applications without any further training.

	Dataset	Sensor	# data	Material cls.	Robot grasp
Train & Eval	Touch and Go [129]	GelSight	120k	✓	
	The Feeling of Success [6]	GelSight	9.3k		✓
	YCB-Slide [109]	DIGIT	183k	✓	
	Object Folder 2.0 [38]	Taxim	180k	✓	✓
Eval.	Object Folder Real [39]	GelSlim	20k	✓	
	Object Folder 1.0 [36]	TACTO	20k	✓	✓
	SSVTP [64]	DIGIT	4.6k	✓	

Table 1. Datasets for training and evaluation.

Zero-shot touch understanding. Emergent alignment of touch and text enables zero-shot touch understanding, *e.g.*, material classification and grasp stability prediction. Following CLIP [102], we encode the touch images and text prompts with templates and class names. We compute their similarity score and rank them to achieve the zero-shot classification.

Touch-LLM. Using an existing vision-language model [34, 144] with the image embedding [41] that we align our touch embedding with, we can create our touch-language model by switching to our touch encoder. Given the touch image and language inputs, we can obtain a more comprehensive understanding via question-answering.

Image synthesis with touch. Binding touch with text also opens up more potential abilities for image synthesis with touch. We leverage the pretrained text-to-image diffusion model [104] and use our touch features to condition the denoising process, achieving zero-shot touch-to-image generation [79, 130] and tactile-driven image stylization.

X-to-touch generation. We also connect other modalities to touch using the diffusion model so that we can achieve x-to-touch generation, where we imagine the touch by seeing, describing, or listening. We train an image-to-touch diffusion model [130] using the pretrained joint image embedding and then we can generate touch from text and audio as well.

4. Experiments

We evaluate our model on extensive tasks spanning various application domains, including zero-shot touch understanding, cross-modal retrieval, zero-shot image synthesis with touch, Touch-LLM, and X-to-touch generation.

Implementations. We base our model on ImageBind [41]. We use the AdamW optimizer [65, 86] with the base learning rate of 1×10^{-5} and cosine decay learning rate scheduler. We train our model with a batch size of 48 on each of the 4 NVIDIA A40 GPUs for 150 epochs. We set the temperature parameter $\tau = 0.07$. We adopt Vision Transformer (ViT) [26] as the backbone for our touch encoder, which contains 24 multi-head attention blocks with 16 heads on each. The feature dimension C is 1024. We use $L = 5$ learnable tokens for each sensor type in our pretraining datasets with

	Method	Pretrain Data	In domain Datasets			Out-of-domain Datasets		
			<i>Touch and Go</i>	<i>ObjectFolder 2.0</i>	<i>YCB-Slide</i>	<i>ObjectFolder 1.0</i>	<i>ObjectFolder Real</i>	<i>SSVTP</i>
	Chance	–	5.0	14.2	10.0	14.2	14.2	16.6
Linear Probing	Supervised	ImageNet	47.1	70.3	72.3	37.5	54.8	73.4
	VT CMC [129]	Single	56.5	74.3	75.2	–	–	–
	SSVTP [64]	Single	47.6	69.8	74.8	–	–	–
	VT CMC [129]	All	49.2	70.3	69.5	33.8	48.1	68.5
	SSVTP [64]	All	43.8	68.9	67.4	35.1	49.7	66.8
	Ours	All	61.3	85.4	78.1	41.3	61.2	77.4
Zero-Shot	Ours	All	52.7	43.5	66.4	32.7	33.2	60.9

Table 2. **Tactile material classification.** We compare our touch features with other methods and ImageNet pretraining. We also report our zero-shot classification performance. The metric is accuracy (%).

Method		Pretrain Data	In domain		Out-of-domain
			Feeling	OF 2.0	OF 1.0
Linear Probing	Chance	-	52.3	52.0	50.7
	Supervised	ImageNet	75.9	70.1	68.9
	VT CMC [129]	Single	80.1	74.8	-
	SSVTP [64]	Single	80.3	74.0	-
	VT CMC [129]	All	66.1	65.8	67.2
	SSVTP [64]	All	65.8	64.2	65.3
	Ours	All	82.3	78.1	75.8
Zero-Shot	Ours	All	65.5	64.3	64.7

Table 3. **Robotics grasping stability prediction.** We compare our touch features with other methods and ImageNet pretraining on grasping stability prediction task. We report our zero-shot results. The metric is accuracy (%).

$K = 3$ different sensors. For the in-batch sampling, we set $\sigma = 0.75$, meaning that 75% of the data comes from the same dataset, with the remainder sourced from others.

Datasets. We train and evaluate our model on four visuo-tactile datasets collected by three different vision-based tactile sensors (Tab. 1). These include the real-world dataset Touch and Go [129], the robotic dataset Feeling of Success [6], the YCB-Slide [109] dataset featuring DIGIT sensor interactions, and the multimodal dataset ObjectFolder 2.0 [38] which contains simulated visual, tactile, and audio data of daily objects using Taxim tactile simulators. We train our model solely on the naturally paired image and touch data via self-supervision. To test the generalization ability of our model, we also evaluate it with three out-of-domain datasets with two unseen sensors, including ObjectFolder Real [39], ObjectFolder 1.0 [36] and SSVTP [64]. We specifically select objects 101-1000 from ObjectFolder 2.0 to avoid overlap with ObjectFolder 1.0. Also, ObjectFolder Real contains objects distinct from those in ObjectFolder 1.0 and 2.0. Please see the supp. for more details.

4.1. UniTouch representation

First, we evaluate the quality of our learned touch features for downstream tasks: material classification and grasping stability prediction via linear probing. We freeze the learned touch embeddings and train a linear classifier on the downstream tasks for specific datasets.

Baselines. We compare our model with two recent visuo-tactile self-supervised methods for vision-based tactile sensors: VT CMC [129] and SSVTP [64]. We also adopt them to our multi-dataset setup. We use the same architectures to ensure a fair comparison. We also compare with the supervised ImageNet [24] features, which are commonly used to represent tactile images [6, 7, 137]. Following [6, 39, 129], we evaluate models’ performance via accuracy metric for both downstream tasks.

Material classification. We evaluate the touch material classification task on three in-domain datasets Touch and Go, ObjectFolder 2.0, and YCB-Slide, and three out-of-domain datasets ObjectFolder 1.0, ObjectFolder Real, and SSVTP. It is worth noting that ObjectFolder Real and ObjectFolder 1.0 contain sensors never seen during the training.

Tab. 2 shows results on linear probing. UniTouch outperforms all the baselines by a large margin, implying that our tactile representations benefit from the alignment to a well-structured embedding space trained on large-scale datasets. In addition, the consistent improvements across all datasets and sensors validate our proposed sensor-specific tokens and in-batch sampling strategy during training – resulting in insignificant generalization gains across different sensors.

Grasping stability prediction. We follow the setting of [6, 39] to predict, from tactile input, whether a robotic gripper can successfully grasp and stably hold an object before it is lifted. Failures occur when the grasped object slips by more than 3cm. We evaluate UniTouch on three datasets: Feeling of Success, ObjectFolder 2.0, and ObjectFolder 1.0, where ObjectFolder 1.0 is an out-of-domain dataset.

The linear probing results are shown in Tab. 3. Our performance consistently outperforms existing baselines by a large margin. Thus, we further demonstrate that our model design and training paradigm are useful not only in computer vision but also can be generalized to robotics tasks.

4.2. Zero-shot touch understanding

We further evaluate UniTouch with zero-shot classification tasks, enabled by the emergent alignment with text during pretraining. We perform material classification and grasping prediction tasks by computing the cosine similarity between the embeddings of touch and corresponding text prompts.

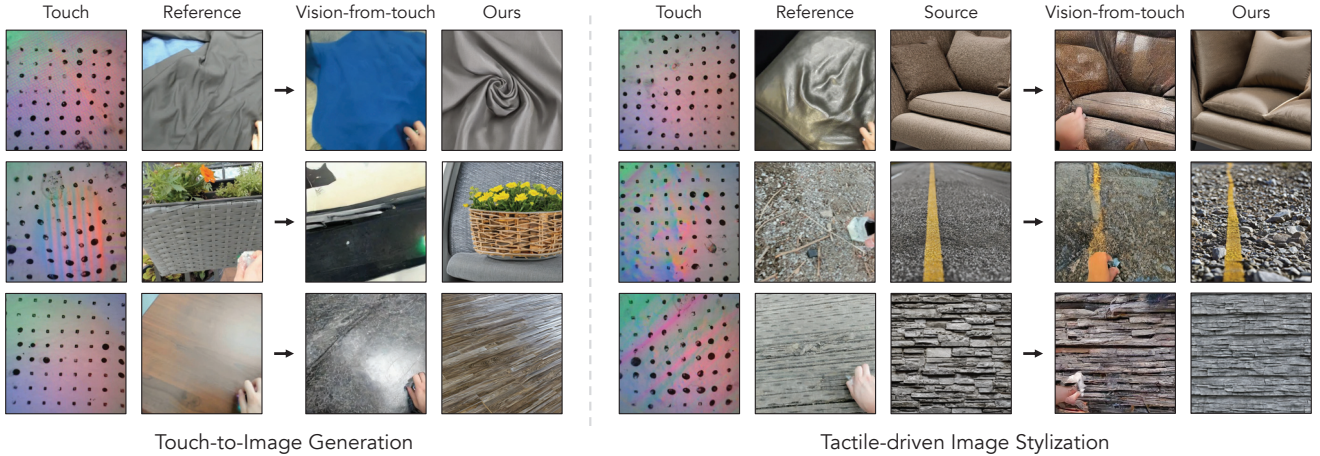


Figure 4. **Zero-shot image synthesis with touch.** (Left) We generate an image of a scene given a tactile signal. (Right) We perform tactile-driven image stylization to manipulate an image to match a given touch signal. We compare our method to the state-of-the-art supervised diffusion method [130] trained on Touch and Go. We denote “reference” as visual images paired with the input touch in the dataset, which are not seen by the model but *only* shown for demonstration purposes. See the supplement for more examples.

Method	Retrieved Modality		
	Touch → Vision	Touch → Audio	Touch → Text
Chance	1.0	1.0	1.0
CCA [†]	8.50	6.18	-
PLSCA [†]	6.25	7.11	-
DSCMR [†]	4.92	6.15	-
Fully supervised			
DAR [†]	8.80	7.77	-
CCA	17.8	15.7	16.8
PLSCA	16.8	15.9	18.2
DSCMR	26.5	19.6	22.7
DAR	32.3	27.8	31.9
Zero-shot			
Ours	41.9	37.9	38.0

Table 4. **Cross-modal retrieval from touch.** We evaluate the performance using mean Average Precision (mAP) on ObjectFolder 2.0. [†] denotes results from [39].

Class predictions are chosen based on highest scores, without training on labeled data. To the best of our knowledge, there are no other baselines that can perform zero-shot touch understanding in our manner.

Material classification. We conduct zero-shot material classification by prompting the model with “This feels like [CLS]”, where [CLS] is the name of the material. We show our zero-shot performance in the last row of Tab. 2. Our zero-shot method shows a comparable performance against several supervised methods, which not only indicates a strong tactile representation that is well-aligned with the text but also shows that off-the-shelf models trained for other modalities can be used to successfully solve touch sensing tasks.

Grasping stability prediction. Similarly, we perform the zero-shot grasping stability prediction task by using text prompts like “the object is lifted in the air” and “the object is falling on the ground”. Tab. 3 shows that we are comparable to some of the supervised methods, demonstrating the capabilities of aligning touch and text can be extended to robotics tasks, which may be out of the training scope of the vision language model like CLIP with appropriate prompting. This may come from the fact that we link the

Method	Evaluation Metrics		
	CVTP (↑)	Material (↑)	FID (↓)
Supervised			
Pix2Pix [54]	0.09	0.15	136.4
VisGel [79]	0.10	0.20	128.3
Vision-from-touch [130]	0.35	0.25	81.2
Zero-shot			
Ours	0.56	0.31	103.11

Table 5. **Zero-shot touch-to-image generation on Touch and Go.**

touch of the successful grasps to the robot’s action of lifting objects while failed grasps as those falling. We found consistent performance in both in and out-of-distribution datasets, demonstrating the generalization capability of this link.

4.3. Cross-modal retrieval with touch

We conduct cross-modal retrieval to evaluate the alignment of our touch embeddings to those of other modalities. Given a touch image, we aim to identify the corresponding vision, text, and audio describing the same point of contact.

Experimental setup. We evaluate on ObjectFolder 2.0 cross-sensory retrieval benchmark [39]. Following [39], we treat points from the same object as positive samples and evaluate using mAP. To evaluate touch-to-text retrieval, we annotated text descriptions that depict the contact point of the object from its visual input, serving as paired ground-truth text. We obtain the retrieval result by ranking the cosine similarity between an input touch and other modalities. Given that our method is not trained with paired audio or text data, we consider its performance in these two modalities as a demonstration of zero-shot learning.

Baselines. We compare our method with several established baselines, including Canonical Correlation Analysis (CCA) [50], Partial Least Squares (PLSCA) [23], Deep Aligned Representations (DAR) [3], and Deep Supervised Cross-Modal Retrieval (DSCMR) [145].

Results. UniTouch achieves state-of-the-art performance on all three modalities and outperforms those supervised

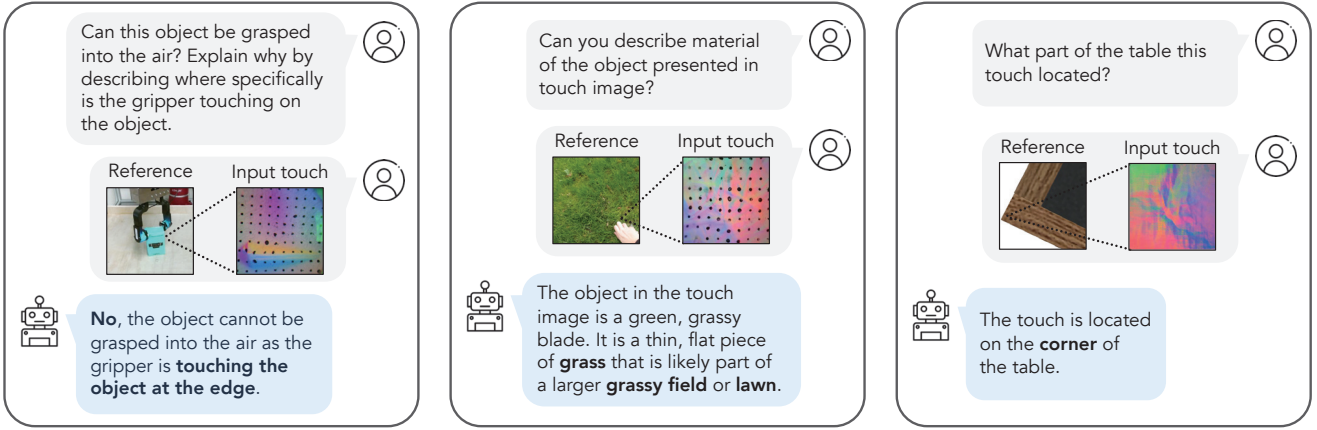


Figure 5. **Touch-LLM**. Our Touch-LLM can conduct a series of tactile question-answer tasks such as robot grasping stability prediction, contact localization, and touch image captioning. We also show “reference” visual images paired with the input touch, for better demonstration. See the supplement for more examples.

Method	LLM	Eval
		GPT-4 Rating ($\uparrow$)
BLIP-2 [78]	Vicuna [17]	1.01
InstructBLIP [21]	Vicuna [17]	1.93
LLaVA-1.5 [84]	Vicuna [17]	2.33
ImageBind-LLM [47]	LLaMA [115]	1.89
Touch-LLM (ours)	LLaMA [115]	3.54

Table 6. **Touch image caption evaluation**. We evaluate our Touch-LLM and four baselines on our test cases from Touch and Go [129]. Each model’s response is rated by GPT-4 on a scale from 1 to 5.

methods that are trained with paired modalities by a large margin (Tab. 4). This demonstrates our strong cross-modal ability to align touch with other modalities without the need for explicit paired training data or additional supervision.

4.4. Image synthesis with touch

In this part, we demonstrate that we can combine our touch embedding with an off-the-shelf image synthesis model easily to perform the image synthesis tasks conditioning touch images in a zero-shot manner. We perform two tasks: touch-to-image generation [30, 40, 79, 130] and tactile-driven image stylization [129, 130]. Following [129, 130], we use three evaluation metrics: Frechet Inception Distance (FID), Contrastive Visuo-Tactile Pre-Training (CVTP), and material classification consistency. See the supplement for details.

Touch-to-image generation. We aim to generate images solely from touch. We use a pretrained text-to-image diffusion model [104], conditioning on our touch features, and guiding the denoising process. Compared to the state-of-the-art visuo-tactile diffusion-based model [130], our method generates more realistic objects that have not been previously seen in the dataset (see Fig. 4 (left)). While the images generated by [130] not only include the sensor and the arm holding it but also closely resemble the visual images in the training set. Tab. 5 shows quantitative results, where we compare with Vision-from-touch [130], VisGel [79] and Pix2Pix [54] on Touch and Go [129]. Despite a slightly lower FID score

Prompt	Datasets	
	Touch and Go	OF 2
This is an image of [CLS]	40.7	34.3
This is a touch image of [CLS]	43.8	36.8
This looks like [CLS]	49.3	41.7
This feels like [CLS]	52.7	43.5
Image of [CLS]	48.8	40.3
Touch of [CLS]	51.2	40.9

Table 7. **Prompt analysis for touch**. We evaluate our prompt designs for zero-shot material classification on Touch and Go and ObjectFolder 2.0 datasets.

compared to [130], our method outperforms on the CVTP and material consistency metrics. This suggests that while our generated images are out of the distribution of Touch and Go, our approach effectively bridges vision and touch.

Tactile-driven image stylization. We also manipulate an image to align with a given touch signal [129, 130] zero shot. We achieve this by mixing the input image embedding with our conditioned touch embedding and feeding it into the pretrained diffusion model. We show qualitative results in Fig. 4 (right), where the input image is out of the distribution of Touch and Go [129]. We observe the supervised state-of-the-art method [130] fails to change the visual style according to the touch images even though these are seen during the training stage. See the supp. for more details.

4.5. Touch-LLM

Interpreting vision-based touch images, crucial for delicate tasks in fields like robotics, is challenging due to human perceptual limitations. To address this, we integrate UniTouch embedding into a large language model (LLM), leveraging its robust understanding and reasoning capabilities for touch image interpretation, and name it as Touch-LLM. Touch-LLM is capable of a series of tactile tasks such as grasping stability prediction, touch image interpretation, tactile contact localization and *etc.*, most of which are non-trivial to humans, demonstrating the usefulness of combining touch

with LLMs. We show some example tasks in Fig. 5.

Quantitatively, we compare our model with four open-source vision-language models (VLMs): BLIP-2 [78], InstructBLIP [21], LLaVA-1.5 [84], and ImageBind-LLM [47] in the touch image captioning task by feeding them the same touch images and text prompts. We manually create captions for 400 randomly sampled RGB images from Touch and Go [129] as the ground truth. Following [5], we use GPT-4 to perform automatic evaluation by instructing GPT-4 to rate each model’s generations on a scale of 1 to 5 given the reference response. As shown in Tab. 6, our Touch-LLM outperforms other VLMs by a large margin, indicating that our Touch-LLM has much better understanding capabilities for touch images. See the supp. for more details.

4.6. X-to-touch generation

We conduct X-to-touch generation to synthesize realistic tactile images corresponding to the input modality of vision, language, and audio. Fig. 1 shows plausible and consistent tactile images generated from both the visual input and its text captioning. Quantitatively, we evaluate our model on Touch and Go [129], where we measure material classification consistency between touch images generated from vision and its corresponding language captions. Our model achieves 55.3% consistency, demonstrating the reliability of our results. See the supp. for more examples and details.

4.7. Ablation study

Learning from multiple sensors. Tab. 8 ablates the importance of each module design on the zero-shot material classification task with the Touch and Go dataset. The baseline, a vanilla transformer model aligning touch embedding to a fixed vision encoder, drops performance significantly when applied to multiple sensors and datasets, *i.e.*, from 43.1% to 21.4%, indicating the difficulty of the sensor domain gap. We improve the performance by 17% by adding the sensor-specific tokens to it. Similarly, we found a 19% by adding our sampling strategy. With our proposed batch sampling strategy and sensor-specific tokens, our model can achieve strong performance, surpassing the model trained on a single dataset, which emphasizes the significance of our proposed methods for learning a better touch representation from multiple sensors. We argue that this is because sensor-specific embeddings help distinguish hard samples from different sensors while sampling strategy helps identify hard negatives within the same sensor in the training. Combining these, we can tackle inter-sensor and intra-sensor hard samples thus obtaining the performance boost.

Language prompting for touch. We explore how language prompting can help with understanding touch, the first endeavor in this domain. Given that vision captures more global and semantic information, and touch focuses on material properties, texture, and microgeometry, directly adopting

Method	Pretrain Data	Eval
		<i>Touch and Go</i>
Chance	–	16.7
Baseline	<i>Touch and Go</i>	43.1
Baseline	All	21.4
Baseline + sensor token	All	38.1
Baseline + sample	All	40.3
Baseline + sensor token + sample	All	52.7

Table 8. **Ablation study.** We ablate the effectiveness of each of our proposed contributions via the zero-shot material classification.

prompts from vision-language works may not yield satisfactory results. We design touch-specific prompt templates by adopting the common prompts from vision-language works and replacing with words related to haptics, *i.e.*, changing “image” to “touch image” and “look like” to “feel like” (see Tab. 7). We evaluate them using the zero-shot material classification task on Touch and Go and ObjectFolder 2.0. We empirically found that our prompts can significantly improve the performance, indicating that language can indeed understand touch. We suspect this phenomenon may be due to the design of visuo-tactile datasets, which feature human or robotic touch actions, thus enabling the model to associate tactile images with these actions.

5. Discussion

We introduced *UniTouch*, a unified multimodal tactile representation for vision-based tactile sensors. To achieve this, we align our touch embedding to a shared multimodal embedding space using contrastive learning. We further introduce sensor-specific tokens that enables learning from different sensors all at once. UniTouch unifies many existing tactile sensing tasks and significantly expands the range of tasks for touch sensing. Nonetheless, the field of multimodal (foundational) model is admittedly still young. Agents, like ourselves, leverage complementary strengths of multi-sensory observations, incorporating all five senses in everyday tasks. With that goal in mind, we see our work as a concrete step towards that direction, opening new avenues for multimodal touch experience beyond vision and touch and integrating tactile sensing into multimodal foundation models.

Limitations. As the full range of tactile sensors exhibits differing output formats (*e.g.*, image, barometric signals, force), we limit our scope to vision-based tactile sensors. Scaling up our training strategy is key to further integrate emerging tactile sensors in the future. In addition, like other multimodal foundational models, our representation is “black-box”, which does not easily for interpretability in the space, where one may benefit from explainability.

Acknowledgements. We thank Jiacheng Zhang, Shaokai Wu and Chenyang Ma for the helpful discussions and feedback on our manuscript. This work is supported by NSF 2112562 Athena AI Institute and Sony Research.

References

- [1] Arpit Agarwal, Tim Man, and Wenzhen Yuan. Simulation of vision-based tactile sensors using physics based rendering. *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–7, 2020. 2
- [2] Yuki Asano, Mandela Patrick, Christian Rupprecht, and Andrea Vedaldi. Labelling unlabelled videos from scratch with multi-modal self-supervision. *Advances in Neural Information Processing Systems*, 33:4660–4671, 2020. 3
- [3] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. See, hear, and read: Deep aligned representations. *ArXiv*, abs/1706.00932, 2017. 6
- [4] Thoranna Bender, Simon Møe Sørensen, Alireza Kashani, K Eldjarn Hjørleifsson, Grethe Hyldig, Søren Hauberg, Serge Belongie, and Frederik Warburg. Learning to taste: A multimodal wine dataset. *arXiv preprint arXiv:2308.16900*, 2023. 3
- [5] Yonatan Bitton, Hritik Bansal, Jack Hessel, Rulin Shao, Wanrong Zhu, Anas Awadalla, Josh Gardner, Rohan Taori, and Ludwig Schimdt. Visit-bench: A benchmark for vision-language instruction following inspired by real-world use. *arXiv preprint arXiv:2308.06595*, 2023. 8, 2
- [6] Roberto Calandra, Andrew Owens, Manu Upadhyaya, Wenzhen Yuan, Justin Lin, Edward H Adelson, and Sergey Levine. The feeling of success: Does touch sensing help predict grasp outcomes? *Conference on Robot Learning (CoRL)*, 2017. 4, 5, 1
- [7] Roberto Calandra, Andrew Owens, Dinesh Jayaraman, Justin Lin, Wenzhen Yuan, Jitendra Malik, Edward H. Adelson, and Sergey Levine. More than a feeling: Learning to grasp and regrasp using vision and touch. *IEEE Robotics and Automation Letters*, 3:3300–3307, 2018. 5
- [8] Guanqun Cao and Shan Luo. Multimodal perception for dexterous manipulation. *ArXiv*, abs/2112.14298, 2021. 1
- [9] Guanqun Cao, Yi Zhou, Danushka Bollegala, and Shan Luo. Spatio-temporal attention model for tactile texture recognition. *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9896–9902, 2020. 2
- [10] Guanqun Cao, Jiaqi Jiang, Chen Lu, Daniel Fernandes Gomes, and Shan Luo. Touchroller: A rolling optical tactile sensor for rapid assessment of large surfaces. *ArXiv*, abs/2103.00595, 2021. 1
- [11] Guanqun Cao, Jiaqi Jiang, Ningtao Mao, Danushka Bollegala, Min Li, and Shan Luo. Vis2hap: Vision-based haptic rendering by cross-modal generation. *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 12443–12449, 2023. 2
- [12] Arkadeep Narayan Chaudhury, Tim Man, Wenzhen Yuan, and Christopher G. Atkeson. Using collocated vision and tactile sensors for visual servoing and localization. *IEEE Robotics and Automation Letters*, 7:3427–3434, 2022. 2
- [13] Jiaben Chen, Renrui Zhang, Dongze Lian, Jiaqi Yang, Ziyao Zeng, and Jianbo Shi. iquery: Instruments as queries for audio-visual sound separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14675–14686, 2023. 3
- [14] Shixing Chen, Chun-Hao Liu, Xiang Hao, Xiaohan Nie, Maxim Arap, and Raffay Hamid. Movies2scenes: Using movie metadata to learn scene representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6535–6544, 2023. 3
- [15] Ziyang Chen, Shengyi Qian, and Andrew Owens. Sound localization from motion: Jointly learning sound direction and camera rotation. In *International Conference on Computer Vision (ICCV)*, 2023. 3
- [16] Ning Cheng, You Li, Jing Gao, Bin Fang, Jinan Xu, and Wenjuan Han. Towards comprehensive multimodal perception: Introducing the touch-language-vision dataset. 2024. 2
- [17] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. 7
- [18] Alex Church, John Lloyd, Raia Hadsell, and Nathan F. Lepora. Tactile sim-to-real policy transfer via real-to-sim image translation. In *Conference on Robot Learning*, 2021. 2
- [19] Quan Cui, Boyan Zhou, Yu Guo, Weidong Yin, Hao Wu, Osamu Yoshie, and Yubo Chen. Contrastive vision-language pre-training with limited resources. In *European Conference on Computer Vision*, 2022. 4
- [20] Mark R. Cutkosky, Robert D. Howe, and William R. Provancher. Force and tactile sensors. In *Springer Handbook of Robotics*, 2008. 2
- [21] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Albert Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *ArXiv*, abs/2305.06500, 2023. 7, 8
- [22] Vedant Dave, Fotios Lygerakis, and Elmar Rueckert. Multimodal visual-tactile representation learning through self-supervised contrastive pre-training. *ArXiv*, abs/2401.12024, 2024. 2
- [23] Sijmen de Jong, Barry M. Wise, and N. L. Ricker. Canonical partial least squares and continuum power regression. *Journal of Chemometrics*, 15, 2001. 6
- [24] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 5
- [25] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11162–11173, 2021. 3
- [26] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 4
- [27] Yiming Dou, Fengyu Yang, Yi Liu, Antonio Loquercio, and Andrew Owens. Tactile-augmented radiance fields. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024. 2

- [28] Yuexi Du, Ziyang Chen, Justin Salamon, Bryan Russell, and Andrew Owens. Conditional generation of audio from video via foley analogies. In *Conference on Computer Vision and Pattern Recognition 2023*, 2023. 3
- [29] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 3
- [30] Yu Fang, Xuehe Zhang, Wenqiang Xu, Gangfeng Liu, and Jie Zhao. Bidirectional visual-tactile cross-modal generation using latent feature space flow model. *Neural networks : the official journal of the International Neural Network Society*, 172:106088, 2023. 7
- [31] Chao Feng, Ziyang Chen, and Andrew Owens. Self-supervised video forensics by audio-visual anomaly detection. *Computer Vision and Pattern Recognition (CVPR)*, 2023. 3
- [32] Chao Feng, Xinyu Zhang, and Zichu Fei. Knowledge solver: Teaching llms to search for domain knowledge from knowledge graphs. *ArXiv*, abs/2309.03118, 2023. 3
- [33] Letian Fu, Gaurav Datta, Huang Huang, William Chung-Ho Panitch, Jaimyn Drake, Joseph Ortiz, Mustafa Mukadam, Mike Lambeta, Roberto Calandra, and Ken Goldberg. A touch, vision, and language dataset for multimodal alignment. *arXiv preprint arXiv:2402.13232*, 2024. 2
- [34] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023. 4, 2
- [35] Ruihan Gao, Tasbolat Taunyazov, Zhiping Lin, and Y. Wu. Supervised autoencoder joint learning on heterogeneous tactile sensory data: Improving material classification performance. *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10907–10913, 2020. 2
- [36] Ruohan Gao, Yen-Yu Chang, Shivani Mall, Li Fei-Fei, and Jiajun Wu. Objectfolder: A dataset of objects with implicit visual, auditory, and tactile representations. In *CoRL*, 2021. 2, 4, 5, 1
- [37] Ruihan Gao, Tian Tian, Zhiping Lin, and Y. Wu. On explainability and sensor-adaptability of a robot tactile texture representation using a two-stage recurrent networks. *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1296–1303, 2021. 2
- [38] Ruohan Gao, Zilin Si, Yen-Yu Chang, Samuel Clarke, Jeanette Bohg, Li Fei-Fei, Wenzhen Yuan, and Jiajun Wu. Objectfolder 2.0: A multisensory object dataset for sim2real transfer. In *CVPR*, 2022. 2, 4, 5, 1, 7
- [39] Ruohan Gao, Yiming Dou, Hao Li, Tanmay Agarwal, Jeanette Bohg, Yunzhu Li, Li Fei-Fei, and Jiajun Wu. The objectfolder benchmark: Multisensory learning with neural and real objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17276–17286, 2023. 2, 4, 5, 6, 1
- [40] Ruihan Gao, Wenzhen Yuan, and Jun-Yan Zhu. Controllable visual-tactile synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7040–7052, 2023. 7
- [41] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15180–15190, 2023. 1, 2, 3, 4
- [42] Daniel Fernandes Gomes, Paolo Paoletti, and Shan Luo. Beyond flat gelsight sensors: Simulation of optical tactile sensors of complex morphologies for sim2real learning. *ArXiv*, abs/2305.12605, 2023. 2
- [43] Ziyu Guo, Renrui Zhang, Xiangyang Zhu, Yiwen Tang, Xianzheng Ma, Jiaming Han, Kexin Chen, Peng Gao, Xianzhi Li, Hongsheng Li, et al. Point-bind & point-llm: Aligning point cloud with multi-modality for 3d understanding, generation, and instruction following. *arXiv preprint arXiv:2309.00615*, 2023. 3
- [44] Anupam K. Gupta, Laurence Aitchison, and Nathan F. Lepora. Tactile image-to-image disentanglement of contact geometry from motion-induced shear. In *5th Annual Conference on Robot Learning*, 2021. 1
- [45] Irmak Guzey, Ben Evans, Soumith Chintala, and Lerrel Pinto. Dexterity from touch: Self-supervised pre-training of tactile representations with robotic play, 2023. 3
- [46] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 976–980. IEEE, 2022. 3
- [47] Jiaming Han, Renrui Zhang, Wenqi Shao, Peng Gao, Peng Xu, Han Xiao, Kaipeng Zhang, Chris Liu, Song Wen, Ziyu Guo, et al. Imagebind-llm: Multi-modality instruction tuning. *arXiv preprint arXiv:2309.03905*, 2023. 7, 8
- [48] Negin Heravi, Wenzhen Yuan, Allison M. Okamura, and Jeanette Bohg. Learning an action-conditional model for haptic texture generation. *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11088–11095, 2019. 2
- [49] Carolina Higuera, Byron Boots, and Mustafa Mukadam. Learning to read braille: Bridging the tactile reality gap with diffusion models. *arXiv*, 2023. 1
- [50] Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28:321–377, 1936. 6
- [51] Xixi Hu, Ziyang Chen, and Andrew Owens. Mix and localize: Localizing sound sources in mixtures. *Computer Vision and Pattern Recognition (CVPR)*, 2022. 3
- [52] Hung-Jui Huang, Xiaofeng Guo, and Wenzhen Yuan. Understanding dynamic tactile sensing for liquid property estimation. *ArXiv*, abs/2205.08771, 2022. 2
- [53] Fabian Huttmacher. Why is there so much more research on vision than on any other sensory modality? *Frontiers in psychology*, 10:2246, 2019. 1
- [54] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017. 6, 7

- [55] Wei Ji, Long Chen, Yinwei Wei, Yiming Wu, and Tat-Seng Chua. Mrtnet: Multi-resolution temporal network for video sentence grounding. *ICASSP*, 2023. 3
- [56] Wei Ji, Xiangyan Liu, An Zhang, Yinwei Wei, and Xiang Wang. Online distillation-enhanced multi-modal transformer for sequential recommendation. In *Proceedings of the 31th ACM international conference on Multimedia*, 2023. 3
- [57] Jiaqi Jiang and Shan Luo. Robotic perception of object properties using tactile sensing. *ArXiv*, abs/2112.14119, 2021. 1
- [58] Jiaqi Jiang, Guanqun Cao, Daniel Fernandes Gomes, and Shan Luo. Vision-guided active tactile perception for crack detection and reconstruction. *2021 29th Mediterranean Conference on Control and Automation (MED)*, pages 930–936, 2021. 2
- [59] Jiaqi Jiang, Danushka Bollegala, Shan Luo, et al. Learn from incomplete tactile data: Tactile representation learning with masked autoencoders. In *Proceedings of the 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023. 3
- [60] Tudor Jianu, Daniel Fernandes Gomes, and Shan Luo. Reducing tactile sim2real domain gaps via deep texture generation networks. *2022 International Conference on Robotics and Automation (ICRA)*, pages 8305–8311, 2021. 2
- [61] Micah K Johnson and Edward H Adelson. Retrographic sensing for the measurement of surface texture and shape. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1070–1077. IEEE, 2009. 2
- [62] Micah K. Johnson, Forrester Cole, Alvin Raj, and Edward H. Adelson. Microgeometry capture using an elastomeric sensor. *ACM SIGGRAPH 2011 papers*, 2011. 2
- [63] Zhanat Kappasov, Juan Antonio Corrales, and Véronique Perdereau. Tactile sensing in dexterous robot hands - review. *Robotics Auton. Syst.*, 74:195–220, 2015. 2
- [64] Justin Kerr, Huang Huang, Albert Wilcox, Ryan Hoque, Jeffrey Ichnowski, Roberto Calandra, and Ken Goldberg. Self-supervised visuo-tactile pretraining to locate and follow garment features. In *Robotics: Science and Systems*, 2023. 2, 4, 5, 1
- [65] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representation*, 2015. 4
- [66] Mike Lambeta, Po wei Chou, Stephen Tian, Brian Yang, Benjamin Maloon, Victoria Rose Most, Dave Stroud, Raymond Santos, Ahmad Byagowi, Gregg Kammerer, Dinesh Jayaraman, and Roberto Calandra. Digit: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation. *IEEE Robotics and Automation Letters*, 5:3838–3845, 2020. 2
- [67] Mike Lambeta, Huazhe Xu, Jingwei Xu, Po wei Chou, Shaoxiong Wang, Trevor Darrell, and Roberto Calandra. Pytouch: A machine learning library for touch processing. *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13208–13214, 2021. 2
- [68] Dong Lao, Fengyu Yang, Daniel Wang, Hyounseob Park, Samuel Lu, Alex Wong, and Stefano Soatto. On the viability of monocular depth pre-training for semantic segmentation. *arXiv preprint arXiv:2203.13987*, 2022. 3
- [69] Susan J. Lederman and Roberta L. Klatzky. Hand movements: A window into haptic object recognition. *Cognitive Psychology*, 19:342–368, 1987. 2
- [70] Susan J. Lederman and R. L. Klatzky. Tutorial review haptic perception: A tutorial. 2009. 2
- [71] Michelle A. Lee, Yuke Zhu, Peter Zachares, Matthew Tan, Krishna Parasuram Srinivasan, Silvio Savarese, Fei-Fei Li, Animesh Garg, and Jeannette Bohg. Making sense of vision and touch: Learning multimodal representations for contact-rich tasks. *IEEE Transactions on Robotics*, 36:582–596, 2019. 2
- [72] Seung Hyun Lee, Wonseok Roh, Wonmin Byeon, Sang Ho Yoon, Chanyoung Kim, Jinkyu Kim, and Sangpil Kim. Sound-guided semantic image manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3377–3386, 2022. 2
- [73] Marion Lepert, Chaoyi Pan, Shenli Yuan, Rika Antonova, and Jeannette Bohg. In-hand manipulation of unknown objects with tactile sensing for insertion. In *Embracing Contacts - Workshop at ICRA 2023*, 2023. 1
- [74] Nathan F. Lepora, Yijiong Lin, Ben Money-Coomes, and John Lloyd. Digitac: A digit-tactip hybrid tactile sensor for comparing low-cost high-resolution robot touch. *IEEE Robotics and Automation Letters*, 7:9382–9388, 2022. 2
- [75] Hao Li, Yizhi Zhang, Junzhe Zhu, Shaoxiong Wang, Michelle A. Lee, Huazhe Xu, Edward H. Adelson, Li Fei-Fei, Ruohan Gao, and Jiajun Wu. See, hear, and feel: Smart sensory fusion for robotic manipulation. In *Conference on Robot Learning*, 2022. 2
- [76] Hongyu Li, Snehal Dikhale, Soshi Iba, and Nawid Jamali. Vihope: Visuotactile in-hand object 6d pose estimation with shape completion. *IEEE Robotics and Automation Letters*, 8(11):6963–6970, 2023. 2
- [77] Hangfei Li, Yiming Wu, and Fangfang Wang. Dynamic network for language-based fashion retrieval. In *Proceedings of the 1st International Workshop on Deep Multimodal Learning for Information Retrieval*, pages 49–57, 2023. 3
- [78] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 7, 8
- [79] Yunzhu Li, Jun-Yan Zhu, Russ Tedrake, and Antonio Torralba. Connecting touch and vision via cross-modal prediction. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10601–10610, 2019. 2, 4, 6, 7
- [80] Yichen Li, Yilun Du, Chao Liu, Chao Liu, Francis Williams, Michael Foshey, Benjamin Eckart, Jan Kautz, Joshua B. Tenenbaum, Antonio Torralba, and Wojciech Matusik. Learning to jointly understand visual and tactile signals. In *The Twelfth International Conference on Learning Representations*, 2024. 2
- [81] Yong-Lu Li, Xiaoqian Wu, Xinpeng Liu, Yiming Dou, Yikun Ji, Junyi Zhang, Yixing Li, Jingru Tan, Xudong Lu, and Cewu Lu. From isolated islands to pangea: Unifying semantic space for human action understanding. *arXiv preprint arXiv:2304.00553*, 2023. 3

- [82] Justin Lin, Roberto Calandra, and Sergey Levine. Learning to identify object instances by touch: Tactile recognition via multimodal matching. *2019 International Conference on Robotics and Automation (ICRA)*, pages 3644–3650, 2019. 2
- [83] David J Linden. *Touch: The science of the hand, heart, and mind*. Penguin Books, 2016. 1
- [84] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 7, 8
- [85] John Lloyd and Nathan F. Lepora. Goal-driven robotic pushing using tactile and proprioceptive feedback. *IEEE Transactions on Robotics*, 38:1201–1212, 2020. 1
- [86] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 4
- [87] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022. 3
- [88] Shan Luo, Wenzhen Yuan, Edward H. Adelson, Anthony G. Cohn, and Raul Fuentes. Vitac: Feature sharing between vision and tactile sensing for cloth texture recognition. *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2722–2727, 2018. 1
- [89] Fotios Lygerakis, Vedant Dave, and Elmar Rueckert. M2curl: Sample-efficient multimodal reinforcement learning via self-supervised representation learning for robotic manipulation. *ArXiv*, abs/2401.17032, 2024. 1
- [90] Chenyang Ma, Kai Lu, Ta-Ying Cheng, Niki Trigoni, and Andrew Markham. See, imagine, plan: Discovering and hallucinating tasks from a single image. 2024. 1
- [91] Paul R Manske. The sense of touch. *Journal of Hand Surgery*, 24(2):213–214, 1999. 1
- [92] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12475–12486, 2021. 3
- [93] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3
- [94] Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba. Learning sight from sound: Ambient sound provides supervision for visual learning. 2018. 3
- [95] Chaoyi Pan, Marion Lepert, Shenli Yuan, Rika Antonova, and Jeannette Bohg. In-hand manipulation of unknown objects with tactile sensing for insertion. 2022. 1
- [96] Zixuan Pan, Zihao Wei, and Andrew Owens. Efficient vision-language pre-training by cluster masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3
- [97] Hyoungeob Park, Anjali Gupta, and Alex Wong. Test-time adaptation for depth completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3
- [98] Leszek Pecyna, Siyuan Dong, and Shan Luo. Visual-tactile multimodality for following deformable linear objects using reinforcement learning. *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3987–3994, 2022. 2
- [99] Haozhi Qi, Brent Yi, Sudharshan Suresh, Mike Lambeta, Y. Ma, Roberto Calandra, and Jitendra Malik. General in-hand object rotation with vision and touch. *ArXiv*, abs/2309.09979, 2023. 1
- [100] Longtian Qiu, Renrui Zhang, Ziyu Guo, Ziyao Zeng, Yafeng Li, and Guangnan Zhang. Vt-clip: Enhancing vision-language models with visual-guided texts. *arXiv preprint arXiv:2112.02399*, 2021. 3
- [101] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. 2
- [102] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3, 4
- [103] Adrian Rofer, Nick Heppert, Abdallah Ayman, Eugenio Chisari, and Abhinav Valada. Pseudotouch: Efficiently imaging the surface feel of objects for robotic manipulation. 2024. 1
- [104] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 4, 7, 1, 3
- [105] Zilin Si and Wenzhen Yuan. Taxim: An example-based simulation model for gelsight tactile sensors. *IEEE Robotics and Automation Letters*, PP:1–1, 2021. 2
- [106] Linda Smith and Michael Gasser. The development of embodied cognition: Six lessons from babies. *Artificial life*, 2005. 1
- [107] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1, 2
- [108] Kim Sung-Bin, Arda Senocak, Hyunwoo Ha, Andrew Owens, and Tae-Hyun Oh. Sound to visual scene generation by audio-to-visual latent alignment. *Computer Vision and Pattern Recognition (CVPR)*, 2023. 3
- [109] Sudharshan Suresh, Zilin Si, Stuart Anderson, Michael Kaess, and Mustafa Mukadam. MidasTouch: Monte-Carlo inference over distributions across sliding touch. In *Proc. Conf. on Robot Learning, CoRL*, Auckland, NZ, 2022. 2, 4, 5, 1
- [110] S. Suresh, Z. Si, J. Mangelson, W. Yuan, and M. Kaess. ShapeMap 3-D: Efficient shape mapping through dense touch and vision. In *Proc. IEEE Intl. Conf. on Robotics and Automation, ICRA*, Philadelphia, PA, USA, 2022. 2

- [111] Aiden Swann, Matthew Strong, Won Kyung Do, Gadiel Sznaier Camps, Mac Schwager, and Monroe Kennedy. Touchgs: Visual-tactile supervised 3d gaussian splatting. 2024. 1
- [112] Tasbolat Taunyazov, Yansong Chua, Ruihan Gao, Harold Soh, and Y. Wu. Fast texture classification using tactile neural coding and spiking neural network. *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9890–9895, 2020. 2
- [113] Ian Taylor, Siyuan Dong, and Alberto Rodriguez. Gelslim 3.0: High-resolution measurement of shape, force and slip in a compact tactile-sensing finger. *2022 International Conference on Robotics and Automation (ICRA)*, pages 10781–10787, 2021. 2
- [114] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European conference on computer vision*, pages 776–794. Springer, 2020. 2, 3
- [115] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 7, 2
- [116] Ruoyu Wang, Shiheng Wang, Songyu Du, Erdong Xiao, Wenzhen Yuan, and Chen Feng. Real-time soft body 3d proprioception via deep vision-based sensing. *IEEE Robotics and Automation Letters*, 5:3382–3389, 2019. 2
- [117] Shaoxiong Wang, Mike Lambeta, Po wei Chou, and Roberto Calandra. Tacto: A fast, flexible, and open-source simulator for high-resolution vision-based tactile sensors. *IEEE Robotics and Automation Letters*, 7:3930–3937, 2020. 2
- [118] Yiming Wu, Xintian Wu, Xi Li, and Jian Tian. Mgh: Meta-data guided hypergraph modeling for unsupervised person re-identification. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1571–1580, 2021. 3
- [119] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 3
- [120] Yangchao Wu, Tian Yu Liu, Hyungseob Park, Stefano Soatto, Dong Lao, and Alex Wong. Augundo: Scaling up augmentations for unsupervised depth completion. *arXiv preprint arXiv:2310.09739*, 2023. 3
- [121] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018. 3
- [122] Eric Zhongcong Xu, Zeyang Song, Satoshi Tsutsui, Chao Feng, Mang Ye, and Mike Zheng Shou. Ava-avd: Audio-visual speaker diarization in the wild. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3838–3847, 2022. 3
- [123] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021. 3
- [124] Huazhe Xu, Yuping Luo, Shaoxiong Wang, Trevor Darrell, and Roberto Calandra. Towards learning to play piano with dexterous hands and touch. *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10410–10416, 2021. 1
- [125] Wenqiang Xu, Zhenjun Yu, Han Xue, Ruolin Ye, Siqiong Yao, and Cewu Lu. Visual-tactile sensing for in-hand object reconstruction. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8803–8812, 2023. 2
- [126] Zhiyang Xu, Chao Feng, Rulin Shao, Trevor Ashby, Ying Shen, dingnan jin, Yu Cheng, Qifan Wang, and Lifu Huang. Vision-flan: Scaling human-labeled tasks in visual instruction tuning. *ArXiv*, abs/2402.11690, 2024. 3
- [127] Le Xue, Mingfei Gao, Chen Xing, Roberto Mart’in-Mart’in, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Nieves, and Silvio Savarese. Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1179–1189, 2022. 2
- [128] Fengyu Yang and Chenyang Ma. Sparse and complete latent organization for geospatial semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1809–1818, 2022. 3
- [129] Fengyu Yang, Chenyang Ma, Jiacheng Zhang, Jing Zhu, Wenzhen Yuan, and Andrew Owens. Touch and go: Learning from human-collected vision and touch. *Neural Information Processing Systems (NeurIPS) - Datasets and Benchmarks Track*, 2022. 1, 2, 4, 5, 7, 8, 3
- [130] Fengyu Yang, Jiacheng Zhang, and Andrew Owens. Generating visual scenes from touch. *International Conference on Computer Vision (ICCV)*, 2023. 2, 4, 6, 7
- [131] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Ce Liu, Lu Yuan, and Jianfeng Gao. Unified contrastive learning in image-text-label space. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19141–19151, 2022. 4
- [132] Zhao-Heng Yin, Binghao Huang, Yuzhe Qin, Qifeng Chen, and Xiaolong Wang. Rotating without seeing: Towards in-hand dexterity through touch. *Robotics: Science and Systems*, 2023. 1
- [133] Kelin Yu, Yunhai Han, Matthew Zhu, and Ye Zhao. Mimic-touch: Learning human’s control strategy with multi-modal tactile feedback. *ArXiv*, abs/2310.16917, 2023. 2
- [134] Xihang Yu, Sangli Teng, Theodor Chakhachiro, Wenzhe Tong, Tingjun Li, Tzu-Yuan Lin, Sarah Koehler, Manuel Ahumada, Jeffrey M Walls, and Maani Ghaffari. Fully proprioceptive slip-velocity-aware state estimation for mobile robots via invariant kalman filtering and disturbance observer. *arXiv preprint arXiv:2209.15140*, 2022. 1
- [135] Wenzhen Yuan, Siyuan Dong, and Edward H. Adelson. Gel-sight: High-resolution robot tactile sensors for estimating geometry and force. *Sensors (Basel, Switzerland)*, 17, 2017. 2

- [136] Wenzhen Yuan, Shaoxiong Wang, Siyuan Dong, and Edward H. Adelson. Connecting look and feel: Associating the visual and tactile properties of physical materials. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4494–4502, 2017. 1, 2
- [137] Wenzhen Yuan, Chenhuo Zhu, Andrew Owens, Man-dayam A Srinivasan, and Edward H Adelson. Shape-independent hardness estimation using deep learning and a gelsight tactile sensor. In *International Conference on Robotics and Automation (ICRA)*, 2017. 1, 5
- [138] Martina Zambelli, Yusuf Aytar, Francesco Visin, Yuxiang Zhou, and Raia Hadsell. Learning rich touch representations through cross-modal self-supervision. In *Conference on Robot Learning*, 2021. 3
- [139] Ben Zandonati, Ruohan Wang, Ruihan Gao, and Y. Wu. Investigating vision foundational models for tactile representation learning. *ArXiv*, abs/2305.00596, 2023. 2
- [140] Ziyao Zeng, Daniel Wang, Fengyu Yang, Hyungseob Park, Stefano Soatto, Dong Lao, and Alex Wong. Worddepth: Variational language prior for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3
- [141] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. *arXiv preprint arXiv:2112.02413*, 2021.
- [142] Renrui Zhang, Ziyao Zeng, Ziyu Guo, Xinben Gao, Kexue Fu, and Jianbo Shi. Dspoint: Dual-scale point cloud recognition with high-frequency fusion. *arXiv preprint arXiv:2111.10332*, 2021.
- [143] Renrui Zhang, Ziyao Zeng, Ziyu Guo, and Yafeng Li. Can language understand depth? In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6868–6874, 2022. 3
- [144] Renrui Zhang, Jiaming Han, AoJun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023. 4, 2
- [145] Liangli Zhen, Peng Hu, Xu Wang, and Dezhong Peng. Deep supervised cross-modal retrieval. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10386–10395, 2019. 6
- [146] Chenhao Zheng, Ayush Shrivastava, and Andrew Owens. Exif as language: Learning cross-modal associations between images and camera metadata. *Computer Vision and Pattern Recognition (CVPR)*, 2023. 3
- [147] Chenhao Zheng, Jieyu Zhang, Aniruddha Kembhavi, and Ranjay Krishna. Iterated learning improves compositionality in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3
- [148] Shaohong Zhong, Alessandro Albin, Oiwi Parker Jones, Perla Maiolino, and Ingmar Posner. Touching a nerf: Leveraging neural radiance fields for tactile sensory data generation. In *Conference on Robot Learning*, pages 1618–1628. PMLR, 2023. 2
- [149] Xiangyang Zhu, Renrui Zhang, Bowei He, Ziyu Guo, Ziyao Zeng, Zipeng Qin, Shanghang Zhang, and Peng Gao. Point-clip v2: Prompting clip and gpt for powerful 3d open-world learning. *ICCV 2023*, 2022. 3

Binding Touch to Everything: Learning Unified Multimodal Tactile Representations

Supplementary Material

A.1. Datasets and Metrics

We provide more details of datasets used in our paper, all of which are publicly available.

Touch and Go [129]. The Touch and Go dataset is a recent, real-world visuo-tactile dataset featuring human interactions with various objects in both indoor and outdoor environments using a GelSight tactile sensor. It comprises 13,900 instances of touch across approximately 4,000 distinct object instances and 20 types of materials. Since it is the only real-world in-the-wild dataset, we apply it to multiple tasks including material classification, image synthesis with touch, Touch LLM, and X-to-touch generation. We use the official train/test split of [129] where the dataset is split by touches, not by frames to avoid similar touch images between the train and test set. For Touch-LLM and X-to-touch applications, we label 400 visual images by asking turkers to provide their captioning to describe the object, touch feeling, and texture from it.

The feeling of success [6]. The Feeling of Success is a robot-collected visuo-tactile dataset of robots grasping objects on a tabletop. The tactile images are all captured by GelSight tactile sensors. It contains 9.3k paired vision and touch images. We apply this dataset to robotic grasping stability predictions. As there is no official split of train/val/test, following [39, 129], we split the dataset by objects in the ratio of 8:1:1.

YCB-Slide [109]. The YCB-Slide dataset comprises DIGIT sliding interactions on YCB objects. The dataset is in the video format where we take all 180k frames for our experiments. The dataset contains 10 YCB objects including a sugar box, a tomato soup can, a mustard bottle, a bleach cleanser, a mug, a power drill, scissors, an adjustable wrench, a hammer, and a baseball. While the tactile images are collected via sliding interaction, the visual input is generated by simulation of the YCB objects. In our experiment, we treat each of the objects as an individual material and our goal is to classify 10 classes. We apply this dataset to material classification.

ObjectFolder 1.0 [36]. The ObjectFolder 1.0 dataset is a simulation dataset containing 3D models of 100 objects from online repositories. The touch images are simulated by TACTO simulators. As the raw dataset is a 3D model with infinite points, we randomly sample 200 points for each object. We apply this dataset to material classification and grasping stability prediction experiments. It is worth noting that for grasping stability prediction experiments, we select 6 objects

suitable for grasping following their setting and achieve relatively balanced successful and failure outcomes for grasping. Following [36], all materials can be categorized into 7 material categories including wood, steel, polycarbonate, plastic, iron, ceramic, and glass. These categories are also applied to ObjectFolder 2.0 and ObjectFolder Real datasets.

ObjectFolder 2.0 [38]. The ObjectFolder 2.0 dataset extends [36] to 1000 objects and improves the acoustic and tactile simulation pipelines to render more realistic multisensory data. For the tactile simulation, it utilizes the Taxim simulator instead of TACTO. Similar to the preprocessing of ObjectFolder 1.0, we sample 200 points for each object. To avoid overlapping with [36], we only take the 101-1000 objects. We apply this dataset to material classification, cross-modal retrieval, robot grasping stability prediction, and Touch-LLM. For cross-modal retrieval and Touch-LLM tasks, we annotate text descriptions that depict the contact point of the object from its visual input, *e.g.* "The corner of a wooden table."

ObjectFolder Real [39]. ObjectFolder Real is an object-centric multimodal dataset containing 100 real-world household objects. The touch images are captured by the GelSlim tactile sensor. Similarly, we sample 200 points for each object thus containing in total of 20k visuo-tactile pairs. We apply this dataset to a material classification task, which is considered an out-of-domain dataset.

SSVTP [64]. SSVTP dataset is a recent human-collected visuo-tactile dataset containing 4.9k paired visuo-tactile images. The touch images are collected via the DIGIT tactile sensor. The objects in this dataset are mainly from garments but also contain materials of metal. We apply this dataset to material classification. As the dataset does not contain material labels, we annotate material labels from the visual images. In total, we classify all images into 6 material categories including cotton, metal, denim fabric, plastic, wood, and nylon.

A.2. Implementation Details

We show more implementation details in this section.

Image synthesis with touch. We used a pretrained stable diffusion-2.1 unclip [104] to perform zero-shot touch-to-image generation by replacing the text condition with our aligned UniTouch embedding. Specifically, we keep the simple text "high quality" as the condition while using our touch embedding as an additional condition. We use DDIM sampler [107] with a guidance scale of 9 and denoising steps of 50. Additionally, we set an embedding strength

of 0.75 for our touch embedding condition. Synthesized images are at the resolution of 768×768 .

As for tactile-driven image stylization, similarly, we still keep the simple text "high quality" as the condition. However, we use both touch and image embeddings as extra conditions to conduct image stylization. We perform a linear combination of touch and image embeddings, the weights for touch and image are set to 0.3 and 0.7 respectively. We use DDIM sampler [107] with a guidance scale of 9 and denoising steps of 50. The strength for linear combination embedding is set to 1 and edited images are at the resolution of 768×768 .

Touch-LLM. We adapt our model from [34, 144], which leverages an adapter to connect our touch encoder and an open-source large language model LLaMA [115]. We replace RGB image embedding with our aligned UniTouch embedding. Concretely, we denote the global touch feature encoded by our touch encoder as $F_T \in \mathbb{R}^{1 \times C_T}$, where C_T is the dimension of the touch embedding. Inspired by prior work [34, 144], we use a projector f , which encodes F_T to have the same dimension as the token embedding in LLaMA [115]:

$$F'_T = f(F_T). \quad (5)$$

Then we repeat F'_T and add it to all text tokens across all layers in language model LLaMA [115] with a zero-initialized learnable gate function:

$$T_j^q = h_{\text{zero}} \cdot F'_T + T_j^q, \quad (6)$$

where j and q denotes the layer and sequence index respectively, T_j^q is the text token embedding, and h_{zero} is the zero-initialized learnable gate function. In our experiments, we use pretrained h_{zero} , and plug our UniTouch embedding in.

X-to-touch generation We conduct our X-to-touch generation model based on stable diffusion. While most existing multimodal tactile datasets only contain vision and touch, we first train an image-to-touch diffusion model and we are able to conduct text-to-touch and audio-to-touch *zero shot* by replacing the image conditioning as they are already aligned. We use the Adam optimizer with a base learning rate of $1e-6$. Models are all trained with 30 iterations using the above learning rate policy. We train our model with a batch size of 48 on 4 RTX A40 GPUs. Since we want to use the aligned condition embeddings, the conditional model is frozen during training. The condition embeddings are integrated into the model using cross-attention. We use the frozen, pre-trained VQGAN to obtain our latent representation, with a spatial dimension of 64×64 . During the inference, we conducted the denoising process for 200 steps and set the guidance scale $s = 7.5$.

A.3. Evaluation Details

Touch-to-image generation Following [130], we use three evaluation metrics of Frechet Inception Distance (FID), Contrastive Visuo-Tactile Pre-Training (CVTP), and Material Classification Consistency. FID is a standard evaluation metric in image synthesis that compares the distribution of real and generated image activations using a trained network. CVTP [130] is a metric similar to CLIP but measures the cosine similarity between the visual and tactile embeddings learned for the generated images and conditioned tactile signals, which used an off-the-shelf network. Material classification consistency [130] uses a material classifier to categorize the predicted and ground truth images and measure the rate at which they agree, where we use CLIP as the zero-shot material classifier by feeding the prompt of "material of [CLS]".

Touch-LLM. We feed each vision language model (including our Touch-LLM) with a touch image and text prompt: "You will be presented with a touch image from an object/surface. Can you describe the touch feeling and the texture?". In the end, we use GPT-4 to perform the automatic evaluation for each model following prior work [5]. Specifically, we provide GPT-4 with: 1) a system prompt describing the desired evaluation behavior; 2) the question; and 3) a human-crafted reference response; 4) each model's generation result (more details see supp.). We instruct GPT-4 to rate each model's generations on a scale of 1 to 5 given the reference response. The template is shown in Fig. 7.

X-to-touch. We test the effectiveness of the x-to-touch model on the Touch and Go dataset, which is the only real-world dataset that contains objects and scenes in the wild. As the objects in this dataset are closely related to the material properties, we measure the material classification consistency between different touches generated from different modalities. We use our UniTouch embedding as the off-the-shelf zero-shot material classifier. For quantitative results for text-to-touch generation, we use the 400 human-labeled text captions as the input. For audio-to-touch generation, as there is no impact sound correlated to this dataset, we manually select audios from ObjectFolder 2.0 as the input that have the same material properties or geometry with the visual image for qualitative evaluations, as shown in Fig. 10.

A.4. Additional Experiments

In-batch sampling mix rate selection. We evaluate different choices of σ for in-batch sampling, where σ denotes the percentage of the data that comes from the same dataset while the rest from others. We set σ to $\{0, 0.5, 0.75, 1.0\}$ and evaluate their zero-shot material classification performance on all six datasets, as shown in Fig. 6. We observe

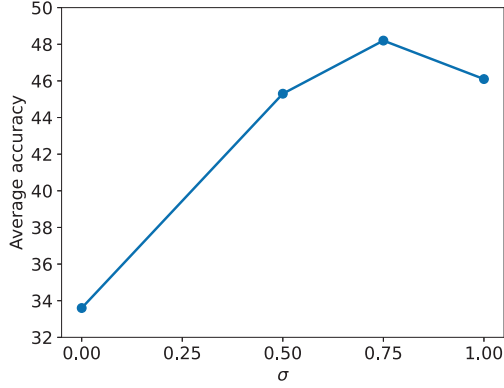


Figure 6. **Effect of σ for in-batch sampling.** We compare the average zero-shot material classification accuracy from six datasets using different σ of 0, 0.5, 0.75, 1.

that if we select $\sigma = 0$, the ability to distinguish between intra-sensor samples is significantly undermined thus leading to inferior performance. As the σ is increasing, the model is able to better distinguish between intra-sensor samples. In the extreme case when $\sigma = 1.0$ where all samples come from the same dataset, the model will have no exposure to the inter-class negatives. We observe that the performance in this case is actually decreasing. This demonstrates the effectiveness of design to balance between inter-sensor and intra-sensor negatives. We empirically found that selecting $\sigma = 0.75$ obtains a good trade-off between these factors.

Image synthesis with touch. We leverage our aligned UniTouch embedding and pretrained text-to-image stable diffusion model [104] to generate more qualitative results of touch-to-image generation and tactile-driven image stylization as presented in Fig. 8. It shows that our UniTouch embedding can guide image synthesis successfully in a zero-shot manner.

X-to-touch generation. We show more examples of X-to-touch generations on the Touch and Go [129] dataset in Fig. 10, where we generate touch images using image, text, and audio.

Touch-LLM. We show more touch image question answering examples in Fig. 9.

system prompt (human authored)

You are EvaluationGPT, an expert language model at judging whether or not a response adequately answers the question. More specifically, you will be given the following:

1. A question: This is a question which requires machine to generate a reasonable answer.
2. A high-quality reference answer: This is a human crafted answer, which should address the question.
3. A candidate response: This is a response generated by model trying to answer the question.

Your job is to judge whether the response adequately answers the question given the reference answer. Please output the number from 1 to 5.

- 1 represents the response is totally irrelevant to the question. 5 means the response fully addresses the question given reference answer.

user (human authored)

I will provide question to you. Then, I will provide a reference answer which is an example of a high quality output for that question in the context of the image (image is not provided). Then I will offer you a candidate response that tries to address the question. Your job is to rate the response on a scale of 1 to 5.

assistant (human authored)

Sure, please provide the question, reference answer, and the candidate response. Then I will rate the response on a scale of 1 to 5 to judge whether the response adequately address the question.

user (human authored)

OK. Here is the question, the high-quality reference answer, and the candidate response.

Question: You will be presented with an touch image from a object/surface. Can you describe the touch feeling and the texture?

Reference: The surface appears to be a piece of fabric. The fabric looks soft and textured with a knit or woven pattern. Touching it would likely feel warm, soft, and slightly bumpy due to the textures. The material would be flexible and malleable, conforming easily to pressure and touch, providing a comfortable and gentle tactile experience.

Response: The touch image shows a soft, fuzzy, and cozy texture, which is characteristic of the material used to make the sweater. The sweater is likely made from a warm and comfortable material, such as cotton or a blend of cotton and other fibers. The texture of the sweater is smooth and inviting, making it an ideal choice for a comfortable and relaxed outfit.

Figure 7. **GPT-4 evaluation template.** We use this template to instruct GPT-4 for automatic evaluation of our Touch-LLM and other selected open-source VLM baselines.

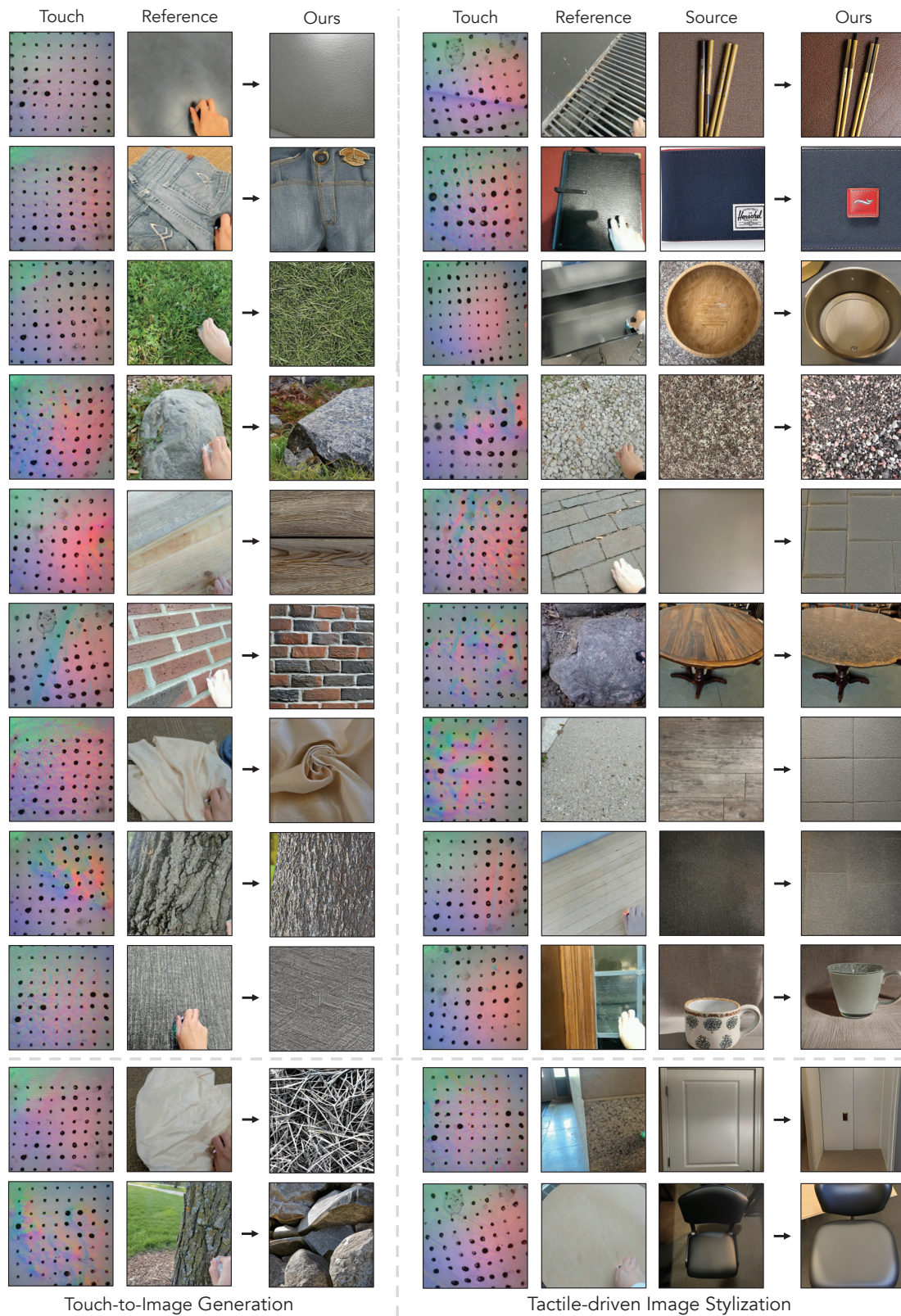


Figure 8. **More examples of zero-shot image synthesis with touch.** (Left) We generate an image of a scene given a tactile signal. (Right) We perform tactile-driven image stylization to manipulate an image to match a given touch signal. We denote “reference” as visual images paired with the input touch in the dataset, which are not seen by the model but *only* shown for demonstration purposes. The last two rows are failure cases.

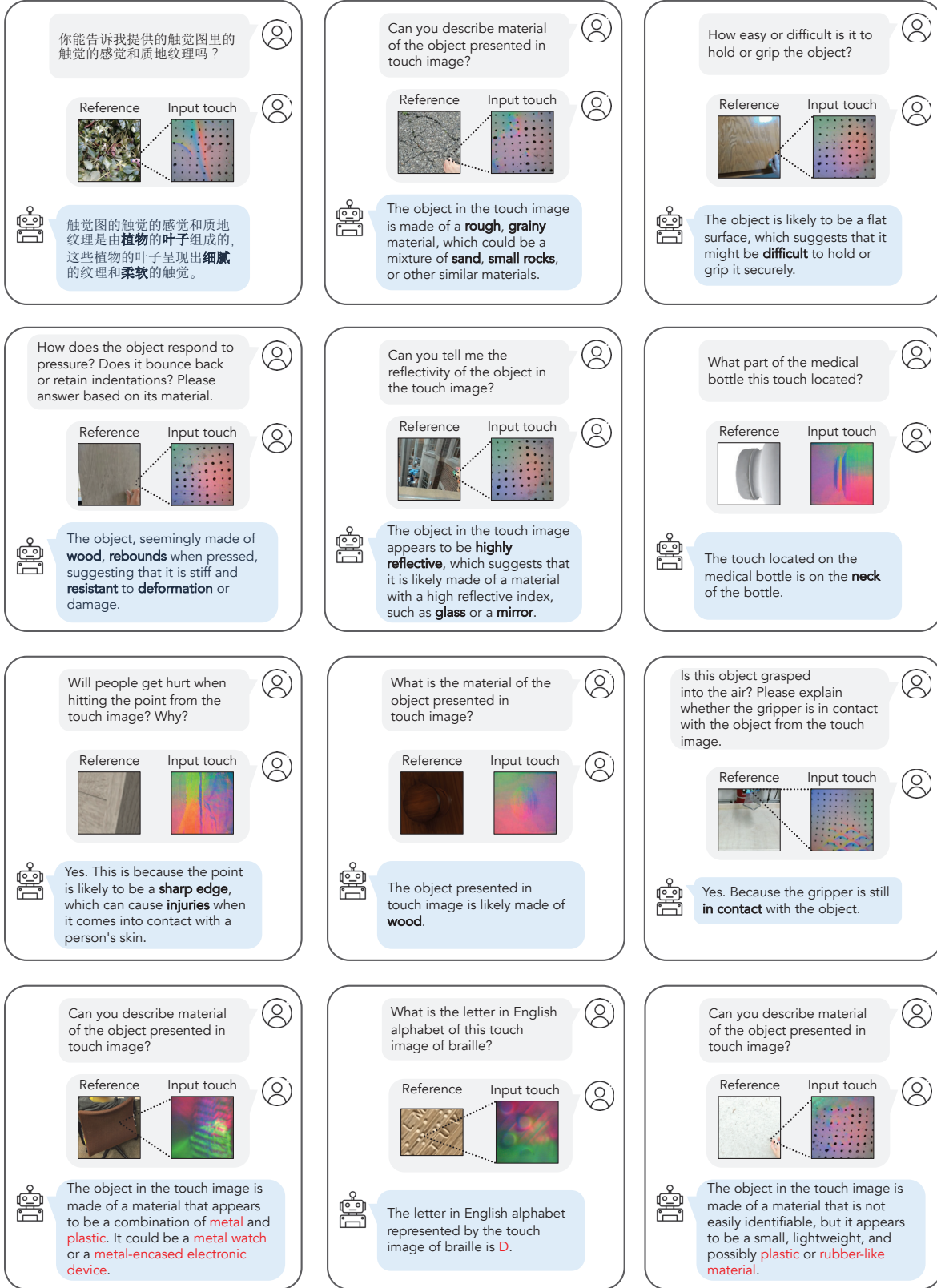


Figure 9. **More examples of Touch-LLM.** We show more question-and-answering examples for touch images using our Touch-LLM. We denote “reference” as visual images paired with the input touch in the dataset, which are not seen by the model but *only* shown for demonstration purposes. The last row is the failure case. Incorrect portion is highlighted in red.

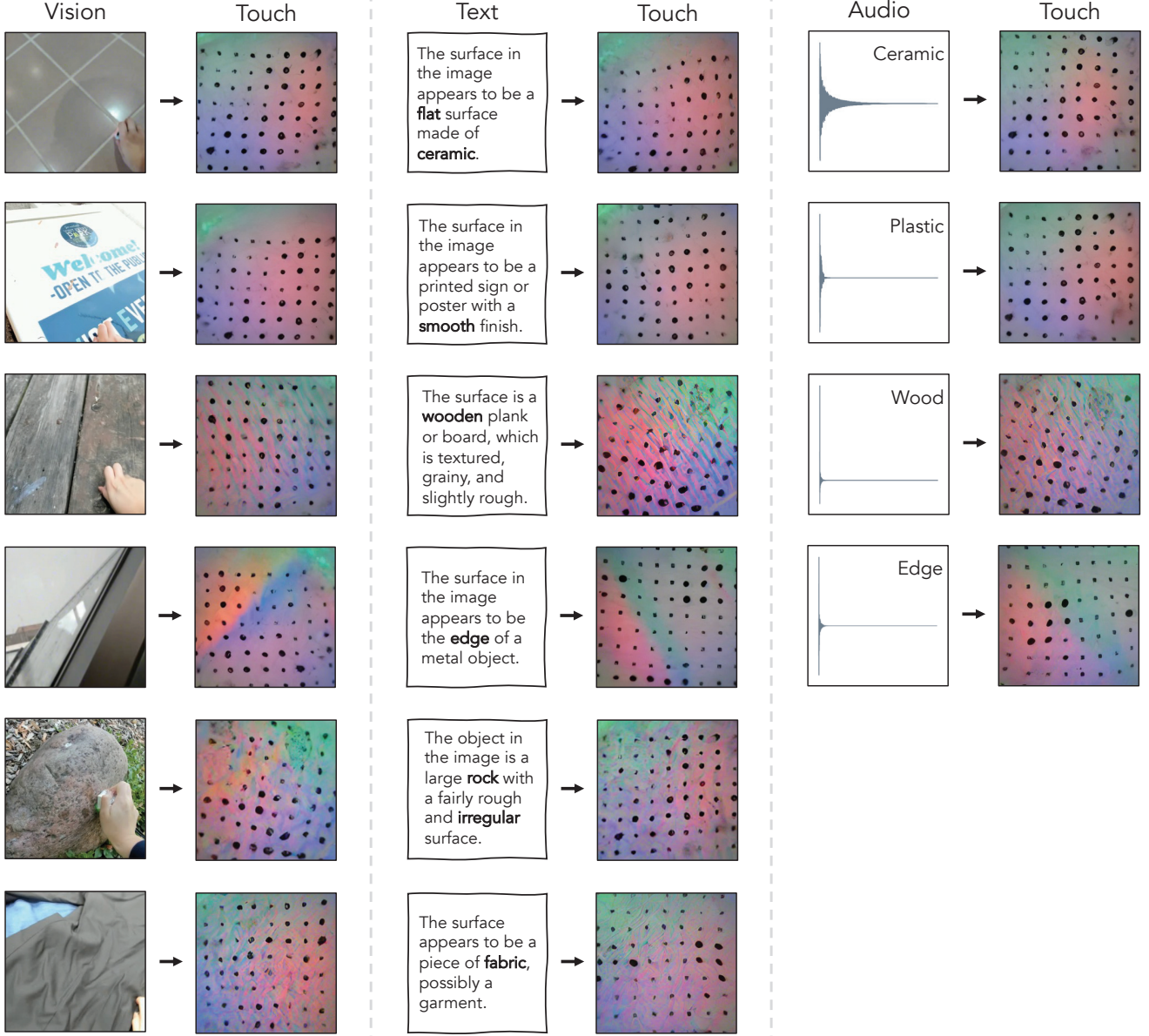


Figure 10. **More examples for X-to-touch generation.** We show more examples of x-to-touch generations on the Touch and Go [129] dataset. We manually select audios from ObjectFolder 2.0 [38] matching the vision input. Since the overlapping material categories between [38] and [129] are limited and [38] only contains rigid objects, impact sound for materials like stone and cloth can not be found.