CLAMP : Crowdsourcing a LArge, in-the-wild haptic dataset with an open-source device for learning a Multimodal robot Perception model

Author Names Omitted for Anonymous Review. Paper-ID [706]

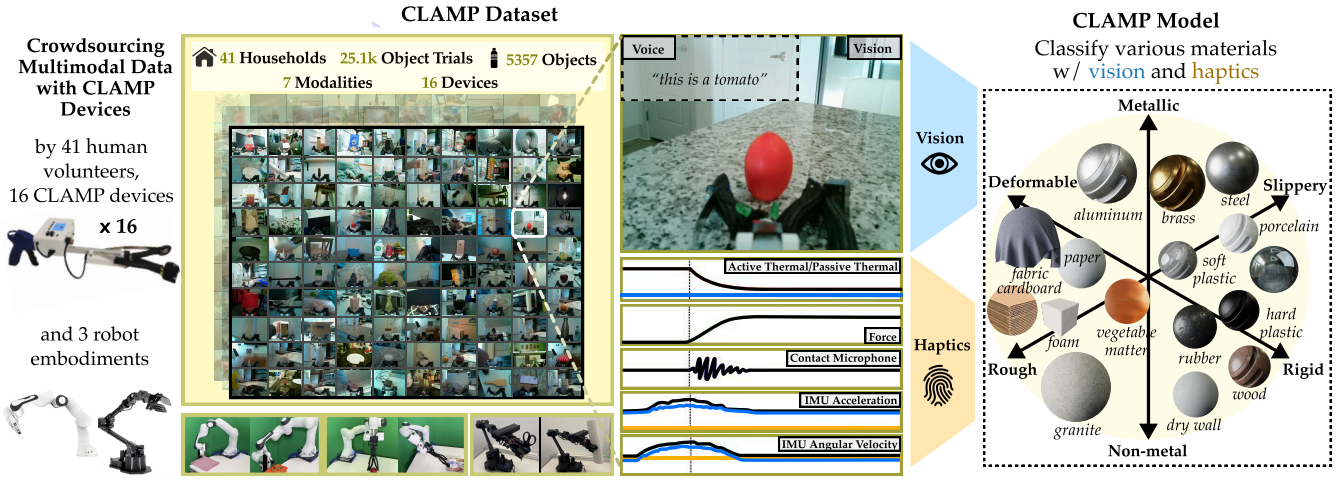

Fig. 1: We present the CLAMP device, dataset and model.

***Abstract*—Haptic sensing enables robots to perceive object properties such as temperature, hardness, and texture, and manipulate objects effectively. However, the performance of state-of-the-art haptic sensing methods is limited by contemporary haptic datasets, which are relatively small and lack diverse, multimodal data. In this work, we demonstrate that *in-the-wild multimodal haptic data collected at scale* can power haptic perception models capable of material recognition in real-world scenarios. We introduce the CLAMP device, a low-cost (<$200) reacher-grabber equipped with sensors for five haptic modalities—active thermal, passive thermal, force, vibration, and proprioception—that features onboard power, storage and an interactive display for non-expert data collection. Using 16 CLAMP devices deployed to 41 participants, we compiled the CLAMP dataset, the largest open-source haptic dataset, which contains 12.3 million multimodal data points from 5357 real-world objects. We leverage our large-scale dataset to propose the CLAMP model, a visuohaptic model capable of performing material recognition in challenging conditions, such as under visual ambiguity and on data from asynchronous and intermittent contact. We show that the CLAMP model can be fine-tuned to learn other object properties such as hardness. Finally, we demonstrate the performance of the CLAMP model on robot data gathered from multiple robots and with varying end-effectors.**

I. Introduction

Haptic sensing, a combination of touch and kinesthesia [1], is essential for perceiving and manipulating objects. Humans combine haptic sensing with vision to infer object properties such as texture, hardness, and temperature [2, 3], and to grasp objects effectively [4]. In robotics, haptic sensing has been used to infer material properties such as object texture, temperature, and surface friction [5–7] and to associate these properties with adjectives [8, 9]. Haptic sensing has also been used to improve robot manipulation under occlusion [10] and visual uncertainty [11, 12].

Despite its potential, haptic sensing remains far less utilized in real-world robotics than vision or language. One reason for this is that haptic sensing methods in the literature [6, 20] are constrained by the limitations of existing haptic datasets. State-of-the-art haptic datasets [16–18, 20] lack one or more of the following attributes:

1) **Size**: Haptic data cannot be scraped from the web; it must be collected by an agent (human or robot) with a haptic device. This makes data collection labor-intensive. Moreover, haptic datasets collected from different sources cannot be aggregated easily since haptic data depends on the kinematics and sensing capabilities of the device, along with the control policies of the agent.
2) **Diversity**: Most haptic datasets are collected with a curated list of objects that are homogeneous in material [13, 18], and do not capture interaction with heterogeneous household objects (e.g. *mobile phones, packaged snacks*). These datasets are also collected by a single expert agent [9, 16, 20]. As a result, they offer

Dataset	Object inst.	Touches	Data Samples	Source	Modalities
Penn Haptic Texture Toolkit [13]	100	200	*	Human (Author)	3
The Feeling of Success [14]	106	9.3k	*	Robot	2
TVL [15]	*	*	44k	Human (Author)	3
Open Access Haptic Database [16]	47	1340	*	Human (Author)	5
Touch100k [17]	*	*	100k	Human (Derived)	3
Penn Haptic Adjective Toolkit [9]	60	600	*	Robot	5
SSVTP [18]	*	*	4500	Robot	2
Proton [19]	357	1.1k	*	Human (Author)	3
Touch and Go [20]	3971	13.9k	*	Human (Author)	2
CLAMP Dataset (Ours)	**5357**	**25.1k**	**12.3M**	Human (Crowdsourced)	**7**

TABLE I: Comparison of the CLAMP dataset with various existing datasets and their characteristics. The CLAMP dataset is the largest open-source haptic dataset to date in terms of data samples and modalities (5 haptic + vision and language).

* indicates unreported figure in source

limited variation in the motion of the device and do not capture real-world scenarios such as varying initial conditions during contact and unavailability of some sensor modalities during contact.

3) **Grasping Interaction:** Most haptic datasets focus on non-prehensile interactions with objects with one virtual finger, such as tapping [9], sliding [21], and pushing [16]. Consequently, these datasets fail to capture real-world scenarios that occur during grasping, such as intermittent and asynchronous contact of haptic sensors on different fingers.
4) **Multimodality**: Haptic sensing is multimodal—it involves temperature [22, 23], force [24, 25], vibration [13, 26], texture [20], and proprioception [27]. Most state-of-the-art haptic datasets do not capture the multimodality of these interactions [14, 15, 20].

In this work, we present CLAMP, a framework for crowdsourcing a large, in-the-wild haptic dataset with an open-source device, for learning a multimodal robot perception model. Our key insight is that *crowdsourced haptic data is diverse, yet structured enough to support learning a mapping between haptic data and object properties.*

Our first contribution is the **CLAMP device**, a low-cost reacher-grabber customized with a camera, a microphone, and a haptic sensing suite that captures five haptic modalities: active thermal, passive thermal, force, vibration, and proprioceptive sensing. The device is portable and lightweight and contains an interactive display, allowing non-experts to grasp objects in their homes and collect multimodal in-the-wild haptic data at scale. An onboard GUI guides users through data collection. The device supports long-term deployment: users gather data at their convenience, power it off when not in use, and return it for maintenance and data backup.

We gather haptic data from 16 CLAMP devices deployed to 41 people and present the **CLAMP dataset**—the largest open-source haptic dataset for material recognition—comprising data samples from 5357 objects. The CLAMP dataset contains diversity in objects, including objects that are heterogenous in material, and in control policies used for interaction, such as grasp orientation, speed, and force. Our dataset contains instances of contact under varying initial temperatures as well as asynchronous sensor contact.

We train a haptic encoder on the CLAMP dataset to perform classification on 14 materials (e.g., `cardboard, steel, aluminum, wood`) that encompass 98% of the objects in our dataset, and show that our encoder can be fine-tuned to learn object properties such as hardness. We combine this encoder with a pretrained visual encoder to present the **CLAMP model**, a multimodal haptic model for material recognition. Through extensive experiments, we demonstrate that the CLAMP model can perform recognition on unseen objects in challenging conditions, such as:

- Imperfect vision, e.g., occlusion and visual ambiguity,
- Imperfect haptics, e.g., varying initial contact conditions and asynchronous sensor contact
- New embodiments, e.g., perception on real robots.

In summary, our contributions are as follows:

1) The CLAMP device, a low-cost, open-source device for crowdsourcing multimodal in-the-wild haptic data.
2) The CLAMP dataset, the largest open-source haptic dataset to date, with 12.3 million samples spanning 5357 object instances collected by 41 people.
3) The CLAMP model, a visuo-haptic model capable of material recognition over diverse objects, grasping interactions, and agents.

II. Related Work

Data Acquisition in Haptics: Scalable approaches for haptic data collection have been underexplored in robotics. Prior research in haptic recognition collected haptic data from robot manipulators [9, 11, 14, 18, 28] or from human experts [18, 20, 21]. These approaches produce in-domain data for method-specific tasks, but cannot scale without the availability of robots and human experts. Haptic sensing rigs [15, 29, 30] provide a pathway to scalable data collection. However, existing haptic sensor rigs are not designed for non-expert user deployment as they are not optimized for size [29], user-friendliness [30], and cost [15]. Moreover, these devices lack streamlined workflows for data collection and annotation. In contrast, the CLAMP device is designed for deployment

in non-expert user homes; the device is lightweight and has a user-friendly interface. Our design approach is inspired by data collection tools in the robot learning community that crowdsource action demonstrations [31–33]. The cost of building a CLAMP device is $200, which is less than that of commercially-available haptic sensors [34–36], allowing anyone to assemble CLAMP devices at scale.

Haptic datasets: Haptic datasets in the literature [6, 9, 11, 14, 16–18, 20, 37–39] are much smaller in size than state-of-the-art datasets in vision [40], language [41] or robot actions [42], despite dataset size being linked to performance improvements in learned models in other domains [43, 44]. Prior to this work, Yang et al. [20] created the largest haptic dataset by number of objects, collecting data from 3971 indoor and outdoor surfaces with short contact durations. Cheng et al. [17] aggregated the largest dataset by data points, combining touch-vision pairs with language descriptions. In comparison, the CLAMP dataset contains 5357 household objects and 12.3 million multimodal data samples, representing a sizable increase in data scale (Table I).

Existing haptic datasets also lack data diversity in terms of the control policies used for interaction and in terms of the objects themselves. Data diversity is known to improve the generalization capabilities of robots [45] and hence is an important quality for robotics datasets. Most datasets are collected in controlled lab settings with a fixed set of materials [22, 23, 46] or objects [14, 16, 38]. These sets often span only one axis of diversity, such as texture [13] or thermal conductivity [47], and thus cannot capture the complex, multi-faceted nature of real-world objects. Datasets are also often collected by robots or expert humans with uniform haptic interactions such as consistent contact duration [9, 14, 16, 20, 38, 48]. Calandra et al. [14] collect haptic data from a robot that randomizes grasping actions over gripping force and end-effector position. These datasets do not cover haptic interaction in non-ideal interactions, such as asynchronous sensor contact, grasping failures, and contact under varying initial conditions. The CLAMP dataset contains in-situ haptic interaction with a large, diverse set of household objects. The data is collected by non-expert human users who grasp objects in an unstructured, natural manner, reflecting realistic, everyday interactions.

Many haptic datasets also lack haptic multimodality [18, 20, 48]. Multimodal data is known to help create better representations and improve model performance in recognition tasks [49]. Specifically in haptic recognition, modalities such as temperature, force and kinesthesia provide complementary information and boost material recognition performance when combined [6, 16]. Haptic datasets from prior works have collected data for force [24, 50], vibration [26], thermal [22, 47], and tactile sensing [15, 20, 21, 51, 52], while some include a combination of the above [9, 16, 37–39]. The CLAMP dataset captures 5 haptic modalities along with vision and structured language, covering more modalities than any other haptic dataset in robotics literature.

Haptic models for recognition: The broader field of haptic recognition is well-studied, with multiple surveys discussing contemporary work [53, 54]. Several works in robotics have used haptic data to perform various recognition tasks such as object property prediction [9], slip detection [55, 56], force estimation [57, 58], material texture recognition [5, 9], and contact state estimation [59]. Works in object material recognition for robots have proposed several methods, however, most of these methods do not explicitly account for imperfect haptic data such as intermittent haptic contact and recognition on a subset of sensor modalities. Erickson et al. [6] and Bhattacharjee et al. [16] crop time-series data to retain only contact moments, but defining contact is subjective when sensors on different fingers touch at different times. Our approach of using a time-series representation of sensor contact helps us perform haptic recognition on data indicating asynchoronous contact, and in test-time unavailability of sensor data.

Haptic recognition methods in the literature seldom report performance on unseen objects. Prior works utilize machine learning tools from SVMs and HMMs to deep neural networks and transformers, which generalize over objects to varying degrees. Erickson et al. [6] report high accuracy scores for their semi-supervised learning approach, however they report a significant performance drop when testing on unseen objects. Lin et al. [60] perform cross-modal recognition of unseen object instances by classifying image-touch pairs as positive or negative pairs. Yang et al. [20] perform material recognition on the in-the-wild TAG dataset, but do not report performance on unseen objects. No prior work in material recognition considers heterogenous objects. In comparison, we show a minimal performance drop on unseen objects and show that our model can identify up to two materials on heterogenous objects despite being trained only on homogenous objects.

Material recognition has also been explored from a computer vision lens. Prior work has highlighted the limitations of existing vision foundation models for vision such as CLIP [61] towards material recognition, as these models primarily use semantic information [62]. Bell et al. [63] created the MINC dataset for material recognition and show that a large dataset helps model performance. Drehwald et al. [62] use a dataset of natural and synthetic images and use self-similarity to identify a material in different objects and environments. However these works point out that one-shot recognition for real-world images can be aided by other sensing modalities. We demonstrate the complementarity of visual and haptic data, with our visuohaptic model outperforming vision-only and haptic-only encoders.

III. The CLAMP Device

Our first contribution is the **CLAMP device** (Figure 2), which enables the crowdsourcing of multimodal haptic data in-the-wild. In this section, we present hardware details and design principles.

A. Hardware

The CLAMP device features a modified reacher-grabber equipped with sensors on both end-effector cups that capture

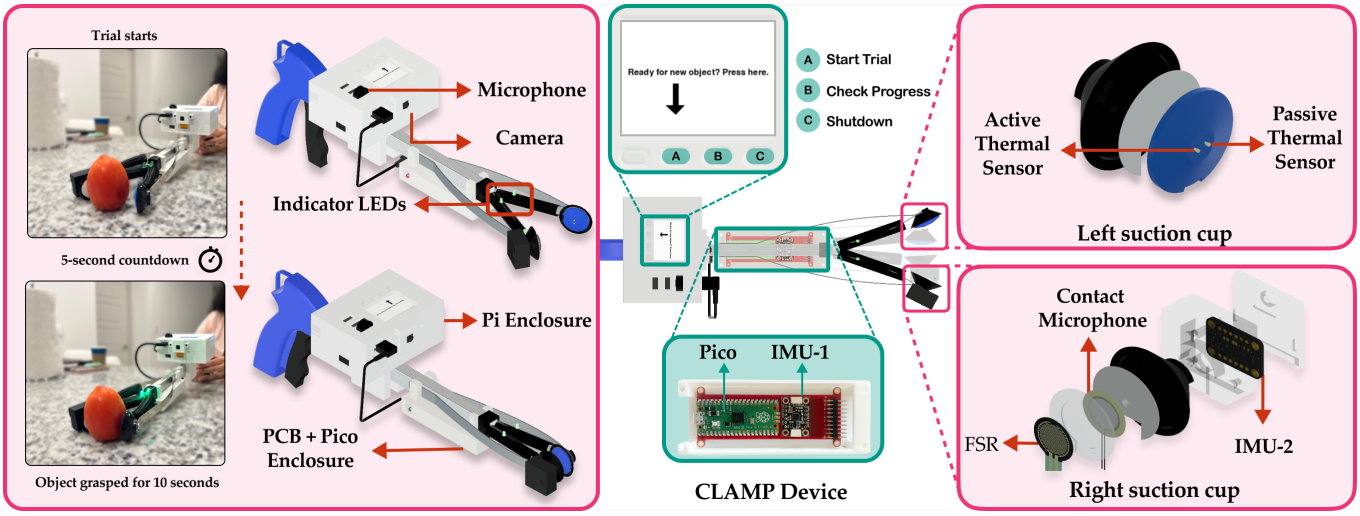

Fig. 2: **Device overview:** The CLAMP device features a modified reacher-grabber equipped with sensors that capture five haptic modalities. Designed to be easy to **carry**, easy to **use**, and easy to **scale**, our device enables non-expert users to collect haptic data in-the-wild.

five forms of haptic data upon contact. These sensors are embedded in 3-D printed sensor beds as shown in Figure 2. We now describe each of the sensors.

Active and passive thermal sensing. The CLAMP device uses two 10 kΩ B57541G1103F NTC thermistors for active and passive thermal sensing (two different modalities). The active thermal sensor is maintained at a temperature of 55°C. The change in active thermal sensor readings over time, when the sensor comes in contact with an object, is indicative of the heat capacity of the object. The passive thermal sensor measures the surface temperature of the object in contact.

Force sensing. The CLAMP device uses an Interlink UX 402 force-sensing resistor that can measure forces up to 50N. We calibrated the force-sensing resistor with an FX29 load cell. The resulting calibration curve is described by an exponential function of the voltage, achieving an R2 = 0.980. While not as accurate as load cell sensors or MEMS force sensors over long-term cyclic use, force-sensing resistors offer significant advantages in terms of resilience and cost.

Vibration sensing. The CLAMP device uses a 20 mm diameter piezo disc, also known as a contact microphone, and a MAX4466 amplifier with a gain of 25 to measure audio signals resulting from contact.

Proprioceptive sensing. An important aspect of haptics is that sensing depends on action. The contact forces generated while grasping an object depend on the velocity and the angle at which the object is grasped. To address this, the CLAMP device has two 6-axis MPU6050 IMUs. The axes of the two IMUs are oriented such that the z-axis of IMU-1 aligns with the y-axis of IMU-2.

Compute and storage. The CLAMP device integrates a Raspberry Pi Pico microcontroller and a Raspberry Pi Zero 2W single-board computer (SBC). The Pico is mounted on a custom printed circuit board (PCB) and samples data from the sensors at a rate of 50 Hz for all sensors except the contact microphone, which is sampled at 100Hz. Data is transferred to the SBC via UART. All data is stored on-board and retrieved when the user is finished with the device.

B. Design Principles

We designed the CLAMP device to facilitate large-scale data collection by non-expert users in-the-wild. To make this possible, we insisted that the device be:

- Easy to Use: The CLAMP device can be used with minimal instruction. The reacher-grabber is also intuitive to use for grasping. For example, the device provides feedback about sensor contact using two LEDs, corresponding to each sensor bed, which light up when the sensor bed makes contact with a surface.
- Easy to Carry: The CLAMP device is easy for users to carry around and take home. The device is lightweight (1.3 lbs) to minimize user fatigue during extended use. This was achieved by choosing the Pi Zero 2W as a lightweight SBC, and by using small end-effectors that keep the center of gravity of the device close to the handle. A consequence of size-weight constraints is that onboard compute has read/write constraints that prevent collection of high bandwidth data; however, haptic data streams lie within these constraints.
- Easy to Deploy: The CLAMP device is designed to allow users to collect and annotate data independently. The device can be operated through a user-friendly interface that consists of button inputs on a 2.2” display screen. The interface guides users to annotate objects using vision and speech. The interface also includes features for tracking progress (number of objects collected) and offers a power-down option when idle.

We have open-sourced the hardware for the CLAMP device, including assembly instructions, CAD files, as well as the user interface software. The device can be assembled from scratch

within 5 hours (excluding 3-D printing time). The total cost of building a CLAMP device is $< \$200$.

IV. The CLAMP Dataset

Our second contribution is the **CLAMP dataset**, the largest open-source haptic dataset to date. The CLAMP dataset consists of 5357 object instances, 25.1k object touches, and 12.3M individual data samples (Table I). These data were collected by 41 people sharing 16 CLAMP devices. In addition to releasing the raw data, we provide post-processed features and object property annotations to facilitate visuo-haptic model training. In this section, we describe the data collection procedure, post-processing and featurization, and data quality.

A. Data Collection Process

We describe the procedure that users follow for collecting data with the CLAMP device.

User on-boarding. We recruited 41 users to collect data with 16 CLAMP devices. Each user is given a short tutorial and a manual on how to use the device.

User interface. Users are guided through the data acquisition process on the graphical interface that is mounted on the device. In between acquisition trials, the number of objects collected so far is displayed on the screen. The user presses a button to start acquisition and presses the button again to advance through each of the following stages.

Image capture. The first step of data acquisition is collecting an image of the object using a 5 MP ArduCam camera mounted on the device. The user interface displays a preview and a countdown before the image is captured.

Audio annotation. Users are then prompted to provide an audio annotation of the object in the form: "This is an *object type*. It is made of *material*." Audio is captured with a Mini USB Microphone mounted on the device.

Grasping trials. Haptic data is collected in five consecutive grasping trials. In each trial, the user grasps and maintains contact with the object for up to 10 seconds. The CLAMP device LEDs light up if there is successful contact.

B. Data Post-Processing and Featurization

Given raw data collected with the CLAMP device, we start by time-synchronizing all sensor data to within 2 ms across modalities, while preserving the exact timestamp for each sensor reading. We first apply a moving average filter to mitigate sensor noise. Then, in addition to the five raw sensor inputs—active thermal, passive thermal, force, vibration (contact microphone), and proprioception (IMU)—we generate four **derived inputs**: active thermal difference, passive thermal difference, force difference, and impedance. The three difference features are computed by subtracting consecutive timesteps in the respective raw inputs.

Impedance is derived from force as follows:

$$Z(t) = \begin{cases} \dfrac{F'(t)}{\omega(t)} & \text{if } \omega(t) \geq \delta \\ 0 & \text{if } \omega(t) < \delta \end{cases}$$

where $F'(t)$ denotes the force difference at time t, $\omega(t)$ denotes the angular velocity of the CLAMP device right finger (proprioception), and δ denotes a threshold for angular velocity, which we fix as $3^\circ/s$ for our experiments. While impedance is typically computed via linear velocity, we use angular velocity because our IMU-based proprioception is more accurate in this dimension. Furthermore, we fix a lower bound of $3^\circ/s$ to exclude spurious values of high impedance that we observe at low angular velocities. This often happens when users attempt to change the grasp contact point, resulting in a sudden increase in force. Finally, we do not use Fast Fourier Transforms due to the low-frequency nature of our data and the diversity in control policies that users exhibit while grasping.

Given the nine sensory modalities (five raw and four derived), we next process the data with modality-specific filters. We then crop the time series such that the first instant indicates the onset of contact for the finger that the sensor is on, and the last instant indicates loss of contact. This step is crucial in synchronizing contact for the two sensor beds, and helps in tackling data from asynchronous contact with both fingers. Details of the modality specific filters and cropping are described in Appendix A. Finally, we align the cropped time series and pad them to ensure all features are of length 490 time steps. The final output of post-processing an individual object touch attempt is thus a 490×9 time series of feature vectors. With 25.1k trials, the overall CLAMP dataset includes $490 \times 25.1\text{k} \approx 12.3\text{M}$ feature vectors.

C. Annotating Object Properties

To enable training our own visuo-haptic model and facilitate future research using the CLAMP dataset, we provide additional annotations for the objects in the data. In particular, each object is labeled with:

1) One of 16 **material** labels: {`aluminium`, `brass`, `cardboard`, `dry wall`, `fabric`, `foam`, `glass`, `granite`, `paper`, `hard plastic`, `soft plastic`, `porcelain`, `rubber`, `steel`, `vegetable matter`, `wood`},
2) Whether the object has two different materials on opposing surfaces, i.e., is **heterogeneous**: {`yes`, `no`},
3) A **hardness adjective**: {`soft`, `hard`},
4) A **temperature adjective**: {`cold`, `neutral`, `warm`}.

We generate these labels automatically using the visual and audio data collected on the CLAMP device. First, we use Whisper [64] to transform the user's spoken audio annotation (e.g., *"This is a water bottle. It is made of plastic."*) into text. We pass the audio transcription and raw image to GPT-4o [65], along with in-context examples, to generate the four annotations listed above. Finally, expert humans verify the audio transcription and material labels and re-run the annotation pipeline for any corrections made.

Material annotation for heterogeneous objects involves some degree of subjectivity. We rely on user-provided information in the audio annotation about grasp parameters such as grasp location (*"I am grasping the handle of the knife, it is made*

of plastic") and object orientation (*"I am grasping a mobile phone made of glass and metal, the glass is on the left side"*). We link this grasp information to a material in the heterogeneous object through in-context examples.

D. Dataset Analysis

The CLAMP dataset contains multi-faceted diversity in objects and in grasp interactions. In particular, the dataset contains 5357 objects, 757 (14%) feature heterogeneous surfaces. For the remaining homogeneous objects, the distribution of material labels is shown in Figure 3. We see the material labels contain significant class imbalance, with the largest class size (`hard plastic`) containing approximately 1000x as many samples as the smallest class (`dry wall`). The dataset also contains substantial diversity in grasps. Figure 4 visualizes this diversity in terms of two grasp parameters: grasp speed and grasping force.

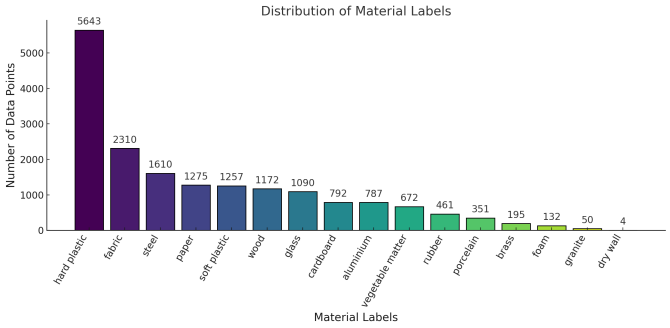

Fig. 3: Material Labels in CLAMP Dataset.

V. The CLAMP Model

In this section, we use the CLAMP dataset to train the **CLAMP model**, a visuo-haptic perception model that uses a haptic encoder to generate haptic representations and a visually-conditioned language model (VLM) to generate visual representations before fusing them together (Figure 5).

A. Generating Haptic Representations

We start by training a haptic encoder to predict a 2048-dimensional haptic embedding from the 490×9 time-series haptic data input. Our haptic encoder is a time-series convolutional neural network based on the InceptionTime architecture [66]. We use 6 InceptionTime blocks. To train the network, we feed the haptic embedding to a 3-layer MLP that generates logits for material prediction. We use the material property annotations in the CLAMP dataset, excluding the two least common classes (`dry wall` and `granite`). We also exclude objects with heterogeneous materials for simplicity. We split the data into 80:10:10 train/val/test and train the model to minimize multi-category cross entropy loss weighted inversely by class size.

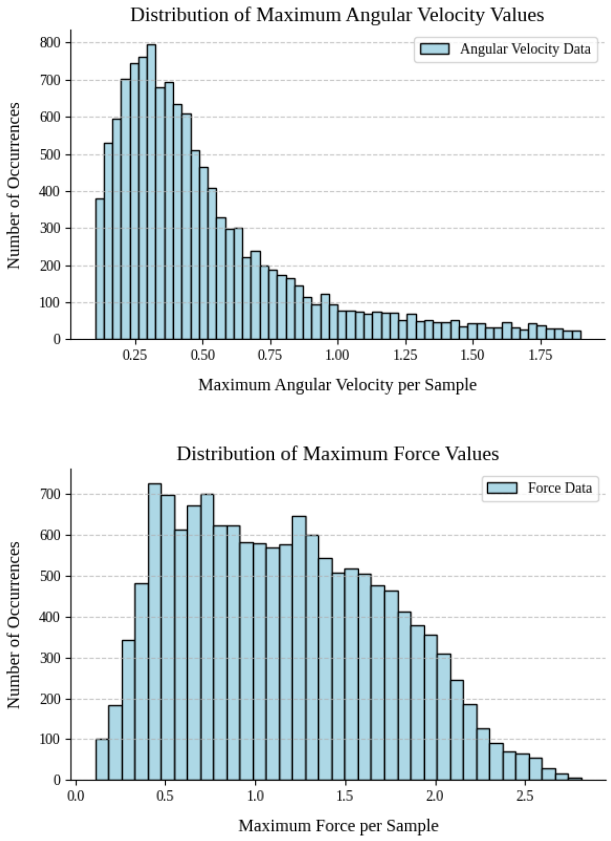

Fig. 4: Diversity in Grasp Parameters.

B. Generating Visual Representations

After preliminary experiments with CLIP [61] and other visual embedding techniques, we found the following approach to work best for incorporating vision into the CLAMP model. Given an object image, we prompt a pretrained VLM (GPT-4o [65]) to directly predict object material. The logits output by the VLM are normalized to create a distribution over the 14 material classes. This 14-dimensional vector serves as our visual representation. For prompting the VLM, we use two stages: first, the VLM is asked to predict the object instance; then, it is asked to predict the material class. We found that this two-stage pipeline outperforms predicting material class directly (see also Section VI-B).

C. The CLAMP Model: Combining Haptics and Vision

We fuse the representations generated by the haptic encoder and VLM to develop the CLAMP model. In particular, given a time-series haptic input and object image, we input them into the pretrained haptic encoder and the VLM respectively to generate two 14-dimensional vectors. These two vectors are concatenated and passed through a 2-layer MLP to make a final material class prediction. To train this combined model, we freeze the haptic encoder and optimize the MLP to minimize weighted cross-entropy loss. We additionally found that regularizing the model to stay close to the predictions of

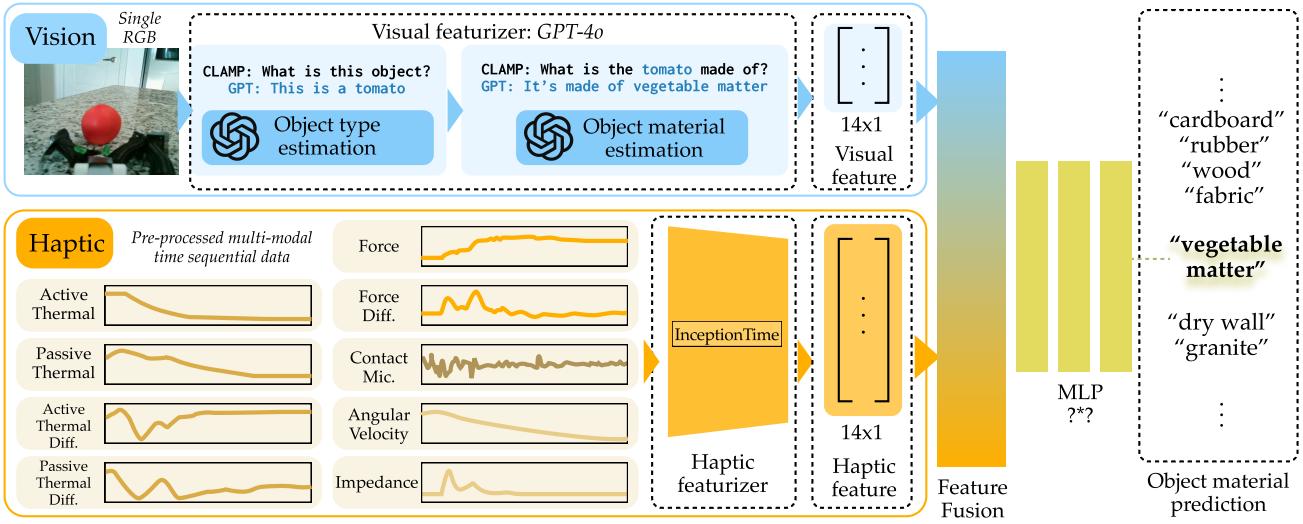

Fig. 5: **Model overview:** We propose the CLAMP model, a visuohaptic model that fuses outputs from a GPT visual encoder and a pretrained InceptionTime-based haptic encoder.

the VLM was helpful for mitigating the effects of spurious correlations in the haptic data. The final loss function is:

$$\mathcal{L} = \mathcal{L}_{WCE} + \lambda_{KL} \cdot \mathcal{L}_{KL}(\mathcal{P}||\mathcal{V})$$

where λ_{KL} is a weighting factor for KL divergence, $\mathcal{P}$ is the predicted probabilities, and $\mathcal{V}$ is the VLM prediction.

VI. Experiments and Results

In this section, we evaluate the effectiveness of the CLAMP model compared to haptic-only and vision-only baselines. We also present a suite of ablation studies to justify our design choices. We then show that the haptic encoder in the CLAMP model can be transferred to other tasks. Finally, we analyze the extent to which the logit predictions of the CLAMP model can be used to measure model uncertainty.

A. Analysis of Haptic + Vision Model Performance

We first compare the CLAMP model to two baselines:

1) Haptic-Only: The haptic encoder (Section V-A) alone.
2) Vision-Only: The VLM classifier (Section V-B) alone.

We evaluate our model on the lowest-performing seed for the haptic-only model. Our main metrics of interest are class-weighted accuracy and normalized Matthews Correlation Coefficient (nMCC), which both take into account the significant class imbalance in the CLAMP dataset. Results are shown in Table II. Overall, we find that the CLAMP model exhibits better performance than both baselines. This confirms that the model is able to fuse low-dimensional information from vision and haptic encoders to generate predictions that are better than either individually. This result also shows that the haptic data in the CLAMP dataset contains useful signal for material recognition.

Perception Model	Test Accuracy	nMCC
Haptic-Only	0.58	0.76
Vision-Only	0.66	0.80
CLAMP Model	0.73	0.84

TABLE II: Decomposing Haptics and Vision.

B. Ablation Studies

We next conduct ablation studies to examine the role of several design choices in the CLAMP model.

Contribution of Each Haptic Modality: We now examine the extent to which all five haptic modalities contribute to the strong performance of the CLAMP model. We compare the original CLAMP haptic encoder to five ablations where each of the haptic modalities are removed. In Table III, we see that the original classifier with all modalities performs best, confirming the importance of each modality.

Multimodal Haptic Fusion Strategy: The timing of feature fusion is a key design decision in multimodal learning [67]. Our haptic encoder performs *early fusion*—all modalities are combined at the start. We also consider *late fusion*, where inputs from sensors on the left finger (active thermal, passive thermal) and right finger (force, motion, vibration) are passed through two InceptionTime networks separately and combined at the classification head; and *hybrid fusion*, where both sets of sensors are passed through two smaller (3-block) InceptionTime networks and then joined for the remaining 3 blocks. Table IV shows that all three fusion strategies perform similarly on material recognition. We thus opt for early fusion for a more unified haptic representation.

Haptic Encoder Architecture: In addition to the InceptionTime architecture, we consider a simple Random Forest classifier, which has been used in past work on haptic classification [68]. In Table V, we see that the InceptionTime architecture consistently outperforms the Random Forest.

Absent Modality	Number of Features	Test Accuracy	nMCC
None	9	0.58	0.76
Active thermal sensing	7	0.49	0.70
Passive thermal sensing	7	0.46	0.68
Force sensing	6	0.22	0.57
Contact microphone	8	0.57	0.75
Proprioceptive sensing	7	0.54	0.74

TABLE III: Haptic Model Performance with Ablated Modalities.

Fusion Strategy	Test Accuracy	nMCC
Early fusion	0.58	0.76
Hybrid fusion	0.59	0.76
Late fusion	0.58	0.75

TABLE IV: Haptic Fusion Strategies.

Model Architecture	Test Accuracy	nMCC
InceptionTime	0.59 ± 0.01	0.76 ± 0.006
Random Forest	0.52 ± 0.001	0.71 ± 0.001

TABLE V: Haptic Encoder Architectures.

Generating Visual Representations from VLMs: We now consider multiple variations of the VLM prompting scheme used to generate visual representations in the CLAMP model. First, in addition to GPT-4o, we consider CLIP [61]. Then, for each of the two pretrained vision models, we consider two alternatives to prompting with raw images: images cropped to the object instance using Grounding Dino [69], and the image masked to the object instance by Grounded SAM [70]. In Table VI, we see that GPT-4o prompted with the raw images yields the best performance. We also note that for both models, the raw image produces the best accuracy, indicating that both models use semantic information to perform material recognition.

Model	Image Type	Metrics	
		Test Accuracy (%)	NMCC
GPT-4o	Raw Image	0.66	0.81
	Segmented Image	0.57	0.75
	Cropped Image	0.65	0.80
CLIP	Raw Image	0.58	0.75
	Segmented Image	0.56	0.74
	Cropped Image	0.56	0.74

TABLE VI: GPT-4o and CLIP on Different Image Types.

C. Transferring the Haptic Encoder

We next evaluate the extent to which our pretrained haptic encoder can be used as a general model of object haptics, beyond the specific material recognition task that we have considered thus far. We use the hardness adjective annotations in the CLAMP dataset to devise a new classification problem. We freeze the haptic encoder and use the 2048-dimensional embeddings as input to a 2-layer MLP of size [512, 128]. The haptic encoder is able to learn hardness adjectives accurately, achieving an accuracy of 0.88 on the test set and an nMCC score of 0.869. This result suggests the general applicability of the haptic encoder.

D. Analyzing CLAMP Model Uncertainty

Uncertainty quantification is important for a variety of downstream tasks [50, 71]. As a final experiment, we analyze the extent to which CLAMP model predictions contain useful signal for uncertainty quantification. We use information from the softmax layer of our model determine prediction uncertainty. Specifically, we mark a prediction as *uncertain* if the corresponding softmax probability is less than a threshold $p_1 = 0.5$, or if the difference between the largest and next largest softmax value is less than a threshold $p_2 = 0.3$. Table VII shows that by removing unknown and uncertain predictions, performance metrics improve, suggesting that there is signal in the model's predicted uncertainty.

Scenario	Predictions (%)	Test Accuracy	nMCC
Original	100%	0.58	0.76
Uncertain Excluded	73%	0.83	0.90

TABLE VII: Uncertainty Quantification Analysis.

VII. Demonstrating the CLAMP Model on Real Robot Embodiments

Our ultimate goal is to imbue robots with multimodal perception capabilities. In this section, we show that the CLAMP model can be used for real-robot perception. We consider three different robot embodiments. For each, we collect a small amount of embodiment-specific data for fine-tuning. We then evaluate the extent to which the fine-tuned CLAMP model can predict material properties for held-out data.

A. Robot Embodiments

We consider the following embodiments:

1) **Franka+Device:** We attach a truncated version of the CLAMP device to the end effector of a Franka Emika Panda. The device's tension-loading steel strips are connected to a Kevlar string, which is actuated by a pair of interlocked and actuated by a pair of Dynamixel XC330-M288-T motors. Finally, a 3-D printed mount, held by the Franka Hand, houses the motors and the Raspberry Pico that powers the sensor suite and secures the steel strips. The mount is designed with grooves that allow the Franka Hand to hold it.
2) **Franka+Sensors:** We directly attach haptic sensor suite to the Franka hand using 3D printed mounts. An enclosure on the Franka Hand houses the Pi.
3) **WidowX+Sensors:** We similarly directly attach haptic sensor suite to the end effector of a WidowX robot using 3D printed mounts. An enclosure again houses the Pi.

In all three embodiments, the arrangement of haptic sensors is the same. Nonetheless, these embodiments represent a significant departure from the human crowdsourcing used to collect the CLAMP dataset.

B. Robot Data Collection and Fine-Tuning

We use 60 objects for fine-tuning and evaluation. However, given the narrow grasping width of the Franka Hand and WidowX gripper, we use only 30 of the 60 for the second and third embodiments. We place each object at a known position in front of the robot. An image of each object is also captured using a camera from a CLAMP device. We then execute a pre-defined motion for grasping and hold for 10 seconds, mirroring the protocol for CLAMP data collection. We post-process and featurize the collected data following the steps described in Section IV-B. For the Franka+Sensors and WidowX+Sensors embodiments, to handle the difference in kinematics with respect to the CLAMP device, we modify the impedance feature calculation to use linear acceleration instead of angular velocity. For each embodiment, we fine-tune the pretrained CLAMP model on varying amounts (7%, 10% and 20% data) of embodiment-specific data.[1]

C. Real-Robot Perception Results

We evaluate each model for each embodiment on the respective held-out data. As shown in Table VIII, we find that fine-tuning on increasing amounts data helps improves performance. We also see a large gap between the Franka+Device model trained on 7% data versus the zero-shot model. This indicates that a combination of diverse pretraining haptic data and small amounts of in-distribution fine-tuning data can help robots perform multimodal haptic recognition with high accuracy.

Embodiment	Scenario	Test Accuracy	nMCC
Franka+Device	Zero-shot	0.64	0.80
	Fine-tuning (7%)	0.75	0.86
	Fine-tuning (10%)	0.78	0.87
	Fine-tuning (20%)	0.80	0.89
Franka+Sensors	Fine-tuning (7%)	0.40	0.64
	Fine-tuning (10%)	0.50	0.70
	Fine-tuning (20%)	0.85	0.92
WidowX+Sensors	Fine-tuning (7%)	0.41	0.65
	Fine-tuning (10%)	0.59	0.77
	Fine-tuning (20%)	0.63	0.78

TABLE VIII: Real-Robot Perception Results.

VIII. Limitations and Future Work

In this work, we presented the CLAMP device, dataset, and model, which collectively represent a significant step towards scaling visuo-haptic perception for robotics. We conclude with a discussion of limitations in the current framework and opportunities for future work.

CLAMP Device Design Constraints. The CLAMP device was designed in an attempt to optimize weight, cost and portability. This lightweight design came with trade-offs. For example, we opted against using a more powerful microprocessor like the Raspberry Pi 4, which would have enabled higher bandwidth data collection, such as video streams, high-frequency contact microphone data, and vision-based tactile sensing. In addition, the current suction cup design allows for agile handling, but its low surface area and convex sensor bed sometimes hinder consistent contact with objects that have curved surfaces. The design of the device is also centered around grasping and does not permit other exploratory actions [9] such as touching or tapping.

CLAMP Dataset Limitations. While the CLAMP dataset is the largest haptic dataset to date, it is still far smaller than counterparts in vision and language. It remains to be seen how much data is required to train large visuo-haptic foundation models that have emergent generalization capabilities analogous to large language and vision-language models. Furthermore, while the CLAMP dataset features diversity along multiple axes, it inherently does not have diversity in terms of the collection device itself. Combining the CLAMP dataset with other datasets could be one path towards even more data diversity.

CLAMP Model Limitations. The CLAMP model performs better than vision-only and haptic-only perception models, but there is much room for future work on learning better visuo-haptic models with the CLAMP dataset. For example, we excluded objects in the CLAMP dataset that have heterogeneous surfaces; incorporating them could lead model improvements if care is taken to handle the heterogeneity. Another direction to explore in future work is training larger models that operate directly on the raw visuo-haptic inputs. The structure we imposed by featurizing the haptic data and prompting a VLM for visual encoding was helpful for the current CLAMP model, but this may change as data and compute continue to scale.

[1] We normalize the dataset using the mean and standard deviation from the CLAMP pretraining dataset, as our fine-tuning data is too limited to provide reliable statistics.

References

[1] Blake Hannaford and Allison M Okamura. Haptics. *Springer handbook of robotics*, pages 1063–1084, 2016.

[2] T Aisling Whitaker, Cristina Simões-Franklin, and Fiona N Newell. Vision and touch: Independent or integrated systems for the perception of texture? *Brain research*, 1242:59–72, 2008.

[3] Elisabeth Baumgartner, Christiane B. Wiebel, and Karl R. Gegenfurtner. Visual and haptic representations of material properties. *Multisensory research*, 26 5:429–55, 2013.

[4] Ivan Camponogara and Robert Volcic. Integration of haptics and vision in human multisensory grasping. *Cortex*, 135:173–185, 2021. ISSN 0010-9452. doi: https://doi.org/10.1016/j.cortex.2020.11.012. URL https://www.sciencedirect.com/science/article/pii/S0010945220304299.

[5] Yang Gao, Lisa Anne Hendricks, Katherine J Kuchenbecker, and Trevor Darrell. Deep learning for tactile understanding from visual and haptic data. In *2016 IEEE*

international conference on robotics and automation (ICRA), pages 536–543. IEEE, 2016.

[6] Zackory Erickson, Sonia Chernova, and Charles C Kemp. Semi-supervised haptic material recognition for robots using generative adversarial networks. In *Conference on Robot Learning*, pages 157–166. PMLR, 2017.

[7] Tran Nguyen Le, Francesco Verdoja, Fares J. Abu-Dakka, and Ville Kyrki. Probabilistic surface friction estimation based on visual and haptic measurements. *IEEE Robotics and Automation Letters*, 6(2):2838–2845, 2021. doi: 10.1109/LRA.2021.3062585.

[8] Vivian Chu, Ian McMahon, Lorenzo Riano, Craig G McDonald, Qin He, Jorge Martinez Perez-Tejada, Michael Arrigo, Naomi Fitter, John C Nappo, Trevor Darrell, et al. Using robotic exploratory procedures to learn the meaning of haptic adjectives. In *2013 IEEE International Conference on Robotics and Automation*, pages 3048–3055. IEEE, 2013.

[9] Vivian Chu, Ian McMahon, Lorenzo Riano, Craig G McDonald, Qin He, Jorge Martinez Perez-Tejada, Michael Arrigo, Trevor Darrell, and Katherine J Kuchenbecker. Robotic learning of haptic adjectives through physical interaction. *Robotics and Autonomous Systems*, 63:279–292, 2015.

[10] Adithyavairavan Murali, Yin Li, Dhiraj Gandhi, and Abhinav Gupta. Learning to grasp without seeing. In *International Symposium on Experimental Robotics*, pages 375–386. Springer, 2018.

[11] 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(4):3300–3307, 2018.

[12] Hao Dang and Peter K. Allen. Learning grasp stability. *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, 2012.

[13] Heather Culbertson, JJ Lopez Delgado, and Katherine J Kuchenbecker. The penn haptic texture toolkit for modeling, rendering, and evaluating haptic virtual textures. *Departmental Papers (MEAM)*, 299, 2014.

[14] 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? *arXiv preprint arXiv:1710.05512*, 2017.

[15] 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.

[16] Tapomayukh Bhattacharjee, Henry M Clever, Joshua Wade, and Charles C Kemp. Multimodal tactile perception of objects in a real home. *IEEE Robotics and Automation Letters*, 3(3):2523–2530, 2018.

[17] Ning Cheng, Changhao Guan, Jing Gao, Weihao Wang, You Li, Fandong Meng, Jie Zhou, Bin Fang, Jinan Xu, and Wenjuan Han. Touch100k: A large-scale touch-language-vision dataset for touch-centric multimodal representation. *arXiv preprint arXiv:2406.03813*, 2024.

[18] 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. *arXiv preprint arXiv:2209.13042*, 2022.

[19] Alexander L Burka. *Instrumentation, data, and algorithms for visually understanding haptic surface properties*. PhD thesis, University of Pennsylvania, 2018.

[20] Fengyu Yang, Chenyang Ma, Jiacheng Zhang, Jing Zhu, Wenzhen Yuan, and Andrew Owens. Touch and go: Learning from human-collected vision and touch. *arXiv preprint arXiv:2211.12498*, 2022.

[21] Samson Yu, Kelvin Lin, Anxing Xiao, Jiafei Duan, and Harold Soh. Octopi: Object property reasoning with large tactile-language models. *arXiv preprint arXiv:2405.02794*, 2024.

[22] Tapomayukh Bhattacharjee, Joshua Wade, and Charles C Kemp. Material recognition from heat transfer given varying initial conditions and short-duration contact. In *Robotics: Science and Systems*, volume 2015, 2015.

[23] Tapomayukh Bhattacharjee, Joshua Wade, Yash Chitalia, and Charles C Kemp. Data-driven thermal recognition of contact with people and objects. In *2016 IEEE Haptics Symposium (HAPTICS)*, pages 297–304. IEEE, 2016.

[24] Tapomayukh Bhattacharjee, James M Rehg, and Charles C Kemp. Haptic classification and recognition of objects using a tactile sensing forearm. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4090–4097. IEEE, 2012.

[25] Alin Drimus, Gert Kootstra, Arne Bilberg, and Danica Kragic. Classification of rigid and deformable objects using a novel tactile sensor. In *2011 15th International Conference on Advanced Robotics (ICAR)*, pages 427–434, 2011. doi: 10.1109/ICAR.2011.6088622.

[26] Jivko Sinapov, Vladimir Sukhoy, Ritika Sahai, and Alexander Stoytchev. Vibrotactile recognition and categorization of surfaces by a humanoid robot. *IEEE Transactions on Robotics*, 27(3):488–497, 2011.

[27] Jivko Sinapov, Connor Schenck, and Alexander Stoytchev. Learning relational object categories using behavioral exploration and multimodal perception. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5691–5698, 2014. doi: 10.1109/ICRA.2014.6907696.

[28] Joseph M. Romano and Katherine J. Kuchenbecker. Methods for robotic tool-mediated haptic surface recognition. In *2014 IEEE Haptics Symposium (HAPTICS)*, pages 49–56, 2014. doi: 10.1109/HAPTICS.2014.6775432.

[29] Alex Burka, Siyao Hu, Stuart Helgeson, Shweta Krishnan, Yang Gao, Lisa Anne Hendricks, Trevor Darrell, and

Katherine J. Kuchenbecker. Proton: A visuo-haptic data acquisition system for robotic learning of surface properties. In *2016 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, pages 58–65, 2016. doi: 10.1109/MFI.2016.7849467.

[30] Joshua Wade, Tapomayukh Bhattacharjee, and Charles C. Kemp. A handheld device for the in situ acquisition of multimodal tactile sensing data, 2015. URL https://arxiv.org/abs/1511.03152.

[31] Sarah Young, Dhiraj Gandhi, Shubham Tulsiani, Abhinav Gupta, Pieter Abbeel, and Lerrel Pinto. Visual imitation made easy. In *Conference on Robot Learning*, pages 1992–2005. PMLR, 2021.

[32] Nur Muhammad Mahi Shafiullah, Anant Rai, Haritheja Etukuru, Yiqian Liu, Ishan Misra, Soumith Chintala, and Lerrel Pinto. On bringing robots home. *arXiv preprint arXiv:2311.16098*, 2023.

[33] Shuran Song, Andy Zeng, Johnny Lee, and Thomas Funkhouser. Grasping in the wild: Learning 6dof closed-loop grasping from low-cost demonstrations. *IEEE Robotics and Automation Letters*, 5(3):4978–4985, 2020.

[34] 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(3):3838–3845, July 2020. ISSN 2377-3774. doi: 10.1109/lra.2020.2977257. URL http://dx.doi.org/10.1109/LRA.2020.2977257.

[35] Mike Lambeta, Tingfan Wu, Ali Sengul, Victoria Rose Most, Nolan Black, Kevin Sawyer, Romeo Mercado, Haozhi Qi, Alexander Sohn, Byron Taylor, et al. Digitizing touch with an artificial multimodal fingertip. *arXiv preprint arXiv:2411.02479*, 2024.

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

[37] Matthias Kerzel, Erik Strahl, Connor Gaede, Emil Gasanov, and Stefan Wermter. Neuro-robotic haptic object classification by active exploration on a novel dataset. In *2019 International joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2019.

[38] Lasse Emil R Bonner, Daniel Daugaard Buhl, Kristian Kristensen, and Nicolás Navarro-Guerrero. Au dataset for visuo-haptic object recognition for robots. *arXiv preprint arXiv:2112.13761*, 2021.

[39] Sibel Toprak, Nicolás Navarro-Guerrero, and Stefan Wermter. Evaluating integration strategies for visuo-haptic object recognition. *Cognitive computation*, 10: 408–425, 2018.

[40] Baoyuan Wu, Weidong Chen, Yanbo Fan, Yong Zhang, Jinlong Hou, Jie Liu, and Tong Zhang. Tencent ml-images: A large-scale multi-label image database for visual representation learning. *IEEE access*, 7:172683–172693, 2019.

[41] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad, 2018. URL https://arxiv.org/abs/1806.03822.

[42] Open X-Embodiment Collaboration, Abby O'Neill, Abdul Rehman, Abhinav Gupta, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, Albert Tung, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anchit Gupta, Andrew Wang, Andrey Kolobov, Anikait Singh, Animesh Garg, Aniruddha Kembhavi, Annie Xie, Anthony Brohan, Antonin Raffin, Archit Sharma, Arefeh Yavary, Arhan Jain, Ashwin Balakrishna, Ayzaan Wahid, Ben Burgess-Limerick, Beomjoon Kim, Bernhard Schölkopf, Blake Wulfe, Brian Ichter, Cewu Lu, Charles Xu, Charlotte Le, Chelsea Finn, Chen Wang, Chenfeng Xu, Cheng Chi, Chenguang Huang, Christine Chan, Christopher Agia, Chuer Pan, Chuyuan Fu, Coline Devin, Danfei Xu, Daniel Morton, Danny Driess, Daphne Chen, Deepak Pathak, Dhruv Shah, Dieter Büchler, Dinesh Jayaraman, Dmitry Kalashnikov, Dorsa Sadigh, Edward Johns, Ethan Foster, Fangchen Liu, Federico Ceola, Fei Xia, Feiyu Zhao, Felipe Vieira Frujeri, Freek Stulp, Gaoyue Zhou, Gaurav S. Sukhatme, Gautam Salhotra, Ge Yan, Gilbert Feng, Giulio Schiavi, Glen Berseth, Gregory Kahn, Guangwen Yang, Guanzhi Wang, Hao Su, Hao-Shu Fang, Haochen Shi, Henghui Bao, Heni Ben Amor, Henrik I Christensen, Hiroki Furuta, Homanga Bharadhwaj, Homer Walke, Hongjie Fang, Huy Ha, Igor Mordatch, Ilija Radosavovic, Isabel Leal, Jacky Liang, Jad Abou-Chakra, Jaehyung Kim, Jaimyn Drake, Jan Peters, Jan Schneider, Jasmine Hsu, Jay Vakil, Jeannette Bohg, Jeffrey Bingham, Jeffrey Wu, Jensen Gao, Jiaheng Hu, Jiajun Wu, Jialin Wu, Jiankai Sun, Jianlan Luo, Jiayuan Gu, Jie Tan, Jihoon Oh, Jimmy Wu, Jingpei Lu, Jingyun Yang, Jitendra Malik, João Silvério, Joey Hejna, Jonathan Booher, Jonathan Tompson, Jonathan Yang, Jordi Salvador, Joseph J. Lim, Junhyek Han, Kaiyuan Wang, Kanishka Rao, Karl Pertsch, Karol Hausman, Keegan Go, Keerthana Gopalakrishnan, Ken Goldberg, Kendra Byrne, Kenneth Oslund, Kento Kawaharazuka, Kevin Black, Kevin Lin, Kevin Zhang, Kiana Ehsani, Kiran Lekkala, Kirsty Ellis, Krishan Rana, Krishnan Srinivasan, Kuan Fang, Kunal Pratap Singh, Kuo-Hao Zeng, Kyle Hatch, Kyle Hsu, Laurent Itti, Lawrence Yunliang Chen, Lerrel Pinto, Li Fei-Fei, Liam Tan, Linxi "Jim" Fan, Lionel Ott, Lisa Lee, Luca Weihs, Magnum Chen, Marion Lepert, Marius Memmel, Masayoshi Tomizuka, Masha Itkina, Mateo Guaman Castro, Max Spero, Maximilian Du, Michael Ahn, Michael C. Yip, Mingtong Zhang, Mingyu Ding, Minho Heo, Mohan Kumar Srirama, Mohit Sharma, Moo Jin Kim, Naoaki Kanazawa, Nicklas Hansen, Nicolas Heess, Nikhil J Joshi, Niko Suenderhauf, Ning Liu, Norman Di Palo, Nur Muhammad Mahi Shafiullah, Oier Mees, Oliver Kroemer, Osbert Bastani, Pannag R Sanketi, Patrick "Tree" Miller,

Patrick Yin, Paul Wohlhart, Peng Xu, Peter David Fagan, Peter Mitrano, Pierre Sermanet, Pieter Abbeel, Priya Sundaresan, Qiuyu Chen, Quan Vuong, Rafael Rafailov, Ran Tian, Ria Doshi, Roberto Mart'in-Mart'in, Rohan Baijal, Rosario Scalise, Rose Hendrix, Roy Lin, Runjia Qian, Ruohan Zhang, Russell Mendonca, Rutav Shah, Ryan Hoque, Ryan Julian, Samuel Bustamante, Sean Kirmani, Sergey Levine, Shan Lin, Sherry Moore, Shikhar Bahl, Shivin Dass, Shubham Sonawani, Shubham Tulsiani, Shuran Song, Sichun Xu, Siddhant Haldar, Siddharth Karamcheti, Simeon Adebola, Simon Guist, Soroush Nasiriany, Stefan Schaal, Stefan Welker, Stephen Tian, Subramanian Ramamoorthy, Sudeep Dasari, Suneel Belkhale, Sungjae Park, Suraj Nair, Suvir Mirchandani, Takayuki Osa, Tanmay Gupta, Tatsuya Harada, Tatsuya Matsushima, Ted Xiao, Thomas Kollar, Tianhe Yu, Tianli Ding, Todor Davchev, Tony Z. Zhao, Travis Armstrong, Trevor Darrell, Trinity Chung, Vidhi Jain, Vikash Kumar, Vincent Vanhoucke, Wei Zhan, Wenxuan Zhou, Wolfram Burgard, Xi Chen, Xiangyu Chen, Xiaolong Wang, Xinghao Zhu, Xinyang Geng, Xiyuan Liu, Xu Liangwei, Xuanlin Li, Yansong Pang, Yao Lu, Yecheng Jason Ma, Yejin Kim, Yevgen Chebotar, Yifan Zhou, Yifeng Zhu, Yilin Wu, Ying Xu, Yixuan Wang, Yonatan Bisk, Yongqiang Dou, Yoonyoung Cho, Youngwoon Lee, Yuchen Cui, Yue Cao, Yueh-Hua Wu, Yujin Tang, Yuke Zhu, Yunchu Zhang, Yunfan Jiang, Yunshuang Li, Yunzhu Li, Yusuke Iwasawa, Yutaka Matsuo, Zehan Ma, Zhuo Xu, Zichen Jeff Cui, Zichen Zhang, Zipeng Fu, and Zipeng Lin. Open X-Embodiment: Robotic learning datasets and RT-X models. https://arxiv.org/abs/2310.08864, 2023.

[43] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *CoRR*, abs/2001.08361, 2020. URL https://arxiv.org/abs/2001.08361.

[44] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12104–12113, 2022.

[45] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024.

[46] Matti Strese, Jun-Yong Lee, Clemens Schuwerk, Qingfu Han, Hyoung-Gook Kim, and Eckehard Steinbach. A haptic texture database for tool-mediated texture recognition and classification. In *2014 IEEE International Symposium on Haptic, Audio and Visual Environments and Games (HAVE) Proceedings*, pages 118–123, 2014. doi: 10.1109/HAVE.2014.6954342.

[47] Haoping Bai, Haofeng Chen, Elizabeth Healy, Charles C Kemp, and Tapomayukh Bhattacharjee. Analyzing material recognition performance of thermal tactile sensing using a large materials database and a real robot. In *2022 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 2255–2262. IEEE, 2022.

[48] Yunzhu Li, Jun-Yan Zhu, Russ Tedrake, and Antonio Torralba. Connecting touch and vision via cross-modal prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10609–10618, 2019.

[49] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. Multimodal deep learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 689–696, Madison, WI, USA, 2011. Omnipress. ISBN 9781450306195.

[50] Tapomayukh Bhattacharjee, Ariel Kapusta, James M Rehg, and Charles C Kemp. Rapid categorization of object properties from incidental contact with a tactile sensing robot arm. In *2013 13th IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, pages 219–226. IEEE, 2013.

[51] Ning Cheng, You Li, Jing Gao, Bin Fang, Jinan Xu, and Wenjuan Han. Towards comprehensive multimodal perception: Introducing the touch-language-vision dataset. *arXiv preprint arXiv:2403.09813*, 2024.

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

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

[54] Lucia Seminara, Paolo Gastaldo, Simon J Watt, Kenneth F Valyear, Fernando Zuher, and Fulvio Mastrogiovanni. Active haptic perception in robots: a review. *Frontiers in neurorobotics*, 13:467142, 2019.

[55] Claudio Melchiorri. Slip detection and control using tactile and force sensors. *IEEE/ASME transactions on mechatronics*, 5(3):235–243, 2000.

[56] Jianhua Li, Siyuan Dong, and Edward Adelson. Slip detection with combined tactile and visual information. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7772–7777. IEEE, 2018.

[57] Carolina Higuera, Akash Sharma, Chaithanya Krishna Bodduluri, Taosha Fan, Patrick Lancaster, Mrinal Kalakrishnan, Michael Kaess, Byron Boots, Mike Lambeta, Tingfan Wu, et al. Sparsh: Self-supervised touch representations for vision-based tactile sensing. *arXiv preprint arXiv:2410.24090*, 2024.

[58] Rishabh Madan, Skyler Valdez, David Kim, Sujie Fang, Luoyan Zhong, Diego T Virtue, and Tapomayukh Bhattacharjee. Rabbit: A robot-assisted bed bathing system with multimodal perception and integrated compliance. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, pages 472–481, 2024.

[59] Kuan-Ting Yu and Alberto Rodriguez. Realtime state

estimation with tactile and visual sensing. application to planar manipulation. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7778–7785. IEEE, 2018.

[60] Justin Lin, Roberto Calandra, and Sergey Levine. Learning to identify object instances by touch: Tactile recognition via multimodal matching. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 3644–3650. IEEE, 2019.

[61] 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.

[62] Manuel S Drehwald, Sagi Eppel, Jolina Li, Han Hao, and Alan Aspuru-Guzik. One-shot recognition of any material anywhere using contrastive learning with physics-based rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23524–23533, 2023.

[63] Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. Material recognition in the wild with the materials in context database. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3479–3487, 2015.

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

[65] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[66] Hassan Ismail Fawaz, Benjamin Lucas, Germain Forestier, Charlotte Pelletier, Daniel F Schmidt, Jonathan Weber, Geoffrey I Webb, Lhassane Idoumghar, Pierre-Alain Muller, and François Petitjean. Inceptiontime: Finding alexnet for time series classification. *Data Mining and Knowledge Discovery*, 34(6):1936–1962, 2020.

[67] Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*, 16:345–379, 2010.

[68] Vinicius Prado da Fonseca, Xianta Jiang, Emil M Petriu, and Thiago Eustaquio Alves de Oliveira. Tactile object recognition in early phases of grasping using underactuated robotic hands. *Intelligent Service Robotics*, 15(4): 513–525, 2022.

[69] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2025.

[70] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.

[71] Allen Z Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, Peng Xu, Leila Takayama, Fei Xia, Jake Varley, et al. Robots that ask for help: Uncertainty alignment for large language model planners. *arXiv preprint arXiv:2307.01928*, 2023.

APPENDIX A
POST-PROCESSING DATA

We describe the post-processing pipeline to convert raw data from a CLAMP device into features.

We first apply a smoothing filter of window size 5 to every feature. This is done to eliminate sensor noise. Next, we compute the derived features, i.e. active thermal difference, passive thermal difference, force difference, and impedance. We then apply signal processing filters to extract information from smoothed sensor data. Table X outlines the filters applied to generate features that the haptic model uses for training.

Once the features are generated, we crop them according to rules that indicate sensor contact. We impose different rules for contact for the left and right sensor suites. The rules are as follows:

- For the left sensing suite, active thermal response determines the state of contact. The left sensor suite is said to make contact at a time instant if the sensor is not in contact at the previous instant and the active thermal difference at that instant is less than a threshold. We set this threshold to $0.01°/s$. The sensor suite is said to break contact if it was in contact at the previous time instant, the active thermal difference at the time instant is positive, and the active thermal reading is less than $53°$. This prevents the active thermal response due the on-off controller from being marked as contact.
- For the right sensing suite, force response determines the state of contact. The right sensor suite is said to make contact at a time instant if the sensor is not in contact at the previous instant and the force response at that instant is more than a threshold. We set this threshold to 0.01 V. The sensor suite is said to break contact if the sensor was in contact at the previous time instant and the force response at the time instant is less than the threshold.

APPENDIX B
VISION MODELS

For vision-based material classification, we preprocess images using Grounded SAM to remove distractor items and CLAMP grippers from the input images. We then evaluate our model on three sets of images: raw images, cropped images, and blackened images. Here, cropped images are cropped to just the bounding box of the object whereas blackened images preserves the original dimensions but mask out all areas except for the segmented object. Thesed images are evaluated using two large Vision-Language models: OpenAI CLIP and GPT-4o.

A. OpenAI CLIP

We fine-tune the final six layers of the CLIP model and append a two-layer multilayer perceptron (MLP) followed by a final fully connected layer for material classification. Fine-tuning focuses on leveraging pre-trained visual and textual embeddings to adapt the model to the CLAMP Image Dataset.

B. OpenAI GPT-4o

We employ a two-step prompting approach for better material classification. We initially feed in an image to identify the object in question. Then, we further prompt GPT with the determined object and input image to use its common-sense reasoning and textual understanding of the image for material classification.

In order to combine logits from GPT with our haptic model, we obtain the top 16 log probs from GPT-4o and combine it with the 16 logits before the softmax of the haptic model. To enforce object, material, and adjective outputs to adhere to our format, we use the following methods:

- In-Context Examples: We provide 14 in-context examples indicating our desired output given a sample image input.
- Logit Bias: We store invalid texts that gpt commonly outputed, such as "please", "sorry", and gave them a -100 logit bias. In contrast, we provide a +100 logit bias if we desired an output from a closed set.
- Structured Outputs: We use OpenAI's structured outputs to enforce the outputs to adhere to a defined JSON schema.

APPENDIX C
HYPERPARAMETER DETAILS

See Table IX.

CLAMP-Net	
Learning Rate	1×10^{-5}
Label Smoothing (ϵ)	0.1
Epochs	100
No. of filters (nf)	256
Batch Size	64
Inception Blocks	6
MLP Hidden Sizes	(256, 128)
MLP Dropout	0.1
GPT Temperature	6
GPT Max Tokens	1

TABLE IX: Hyperparameters used across all domains.

Feature	Sensor Suite	Filter	Parameters
Force	Right	Moving average filter	Window size = 10
Force Difference		Moving average filter	Window size = 10
Contact Microphone		Debiasing	-
IMU		Debiasing	-
Active Thermal	Left	Moving Average	Window size = 10
		Butterworth filter	5nd-order, Cutoff frequency = 50 Hz
Active Thermal Difference		Moving average filter	Window size = 5
Passive Thermal		Moving Average	Window size = 10
		Butterworth filter	2nd-order, Cutoff frequency = 50 Hz
Passive Thermal Difference	Left	Moving average filter	Window size = 5
Impedance	Right	Moving average filter	Window size = 10

TABLE X: Overview of filtering techniques for features