X-Capture: An Open-Source Portable Device for Multi-Sensory Learning

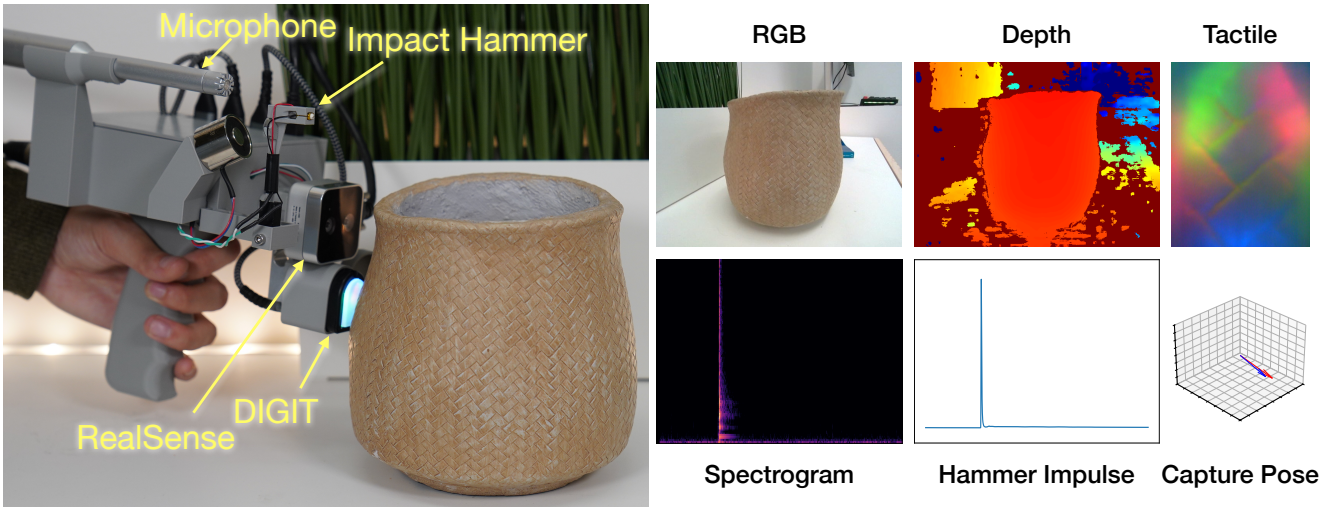


Figure 1. **X-Capture for multi-sensory data capture**. (**Left**) The user captures tactile data from a vase in a living room. (**Right**) The sensor readings for each modality from the same probed point on the vase, as well as a visualization of the hammer impulse and 3D pose vectors for the image and tactile captures, shown in blue and red, respectively.

Abstract

Understanding objects through multiple sensory modalities is fundamental to human perception, enabling cross-sensory integration and richer comprehension. For AI and robotic systems to replicate this ability, access to diverse, high-quality multi-sensory data is critical. Existing datasets are often limited by their focus on controlled environments, simulated objects, or restricted modality pairings. We introduce X-Capture, an open-source, portable, and cost-effective device for real-world multi-sensory data collection, capable of capturing correlated RGBD images, tactile readings, and impact audio. With a build cost under $1,000, X-Capture democratizes the creation of multi-sensory datasets, requiring only consumer-grade tools for assembly. Using X-Capture, we curate a sample dataset of 3,000 total points on 500 everyday objects from diverse, real-world environments, offering both richness and variety[1]*. Our experiments demonstrate the value of both the quantity and the sensory breadth of our data for both pre-training and fine-tuning multi-modal representations for object-centric tasks such as cross-sensory retrieval and reconstruction. X-Capture lays the groundwork for advancing human-like sensory representations in AI, emphasizing scalability, accessibility, and real-world applicability.*

1. Introduction

As humans, we experience objects in our everyday environments through a combination of every sensory modality we possess. Each sensory modality provides us with unique information about the object, which can complement the information evident from other modalities. Touching what visually appears to be a ripe fruit can reveal that it is not in fact ripe yet or has a hidden bruise. Hearing a rigid drinking glass being tapped can disambiguate whether it is made of glass, crystal, or plastic. However, while each of these sensory modalities may complement each other, they are often highly correlated, as they each derive from the underlying physical properties of a given object [25]. Thus we are able to intuitively relate different modalities and generate expectations of other modalities from experiencing only one [26]. In order for robots and agents to understand objects in the same way humans do, they must similarly possess such intuition about the relationships between objects'

[1]Project page, hardware designs, and dataset are available at https://xcapture.github.io.

different sensory modalities.

In this paper, we focus on the sensory modalities of vision (both RGB and depth), sound, and touch—modalities for which popular commercially available sensors exist. Numerous powerful models have been developed to relate these modalities within a shared latent representation for interesting downstream cross-sensory inference and generation tasks [19, 36, 44, 46]. Such models rely on large datasets of examples correlating sensory modalities with each other. Although many such datasets exist, they suffer from key deficiencies that hamper their usefulness in training representations to enhance the understanding of real-world, in-the-wild objects. First, many multi-sensory datasets focus on *scenes* rather than objects, reducing their relevance to applications where object understanding is a priority, such as robotic manipulation. Of those focusing on objects, many only include data collected from *simulated* objects, or from real objects exclusively within controlled environments. Both simulated and controlled real data can present a large domain gap relative to data from real in-the-wild scenarios. Additionally, those collected in controlled environments often require expensive rigs and equipment, a drawback to both accessibility and scalability. Finally, most object-centric multi-sensory datasets lack breadth in the sensory modalities they correlate, often linking only two sensory modalities, such as touch-vision or audio-vision. This hinders representations from forming *direct* rather than emergent alignment among more than two modalities.

To address each of these deficiencies, we introduce X-Capture, a portable, low-cost device for capturing *correlated* RGBD images, tactile images, and impact audio samples from objects in the wild. Our device connects to and is powered by a laptop, with a user interface (UI) that visualizes data during the collection process. We ensure that each sensor takes independent readings (*e.g.*, the touch sensor is not visible in camera images) through careful design of the device and UI, and we explicitly measure an input-output relationship for both touch and audio. These features address common shortcomings of many existing object-centric datasets and collection methods. We open-source the mechanical design and parts list of the device, with a total bill of materials of less than $1000 at time of writing. The device requires only a consumer-grade 3D printer and soldering iron to assemble. Figure 1 shows the X-Capture device in use, along with sensor readings for each modality from a probed sample object. As shown in Table 1, our device is more versatile, portable, and relatively cheaper compared with other data-capturing devices.

We collect a sample dataset of 500 different objects with our device and show how our dataset can be used to fine-tune existing multi-sensory representations to improve their performance in object-centric benchmark tasks such as cross-sensory retrieval, reconstruction, and detection.

Device/Setup	Object Properties				Port.	Cost
	RGB	3D	Audio	Touch		
UBC ACME [35]	✓	✓	✓	✓	✗	N/A*
RealImpact [6]	✓	✓	✓	✗	✗	$8,000
Obj.Folder Real [16]	✓	✓	✓	✓	✗	$11,000
TVL [11]	✓	✗	✗	✓	✓	$560
X-Capture (Ours)	✓	✓	✓	✓	✓	$1000

Table 1. Comparing prior multi-sensory object-centric data capture devices for what object properties they are designed to capture, their portability ("Port."), and their estimated cost. Note that by "Audio" we are referring to impact sounds. The X-Capture device captures images, point clouds, impact audio, and tactile data from objects in a portable and affordable package. (*UBC ACME [35] used equipment which is no longer commercially available at time of writing.)

2. Related Work

Among large foundation models are many popular multi-sensory representation models, which attempt to acquire implicit knowledge of the physical world through their representations. Contrastive Language-Image Pretraining (CLIP) trained a representation between images and their text captions to relate images and text [36]. ImageBind used CLIP as a backbone, along with separate datasets linking images to audio, depth, and thermal data, to train disparate modality-specific encoders to share a single multi-sensory representation [19]. Though each dataset was used to train the model to link images with only one other modality, they showed evidence of *emergent* alignment between other modalities. A Unified Representation of Language, Images, and Point Clouds (ULIP) [44] trained a unified representation of text descriptions, images, and point clouds of objects, showing that the representation was made measurably stronger by aligning all three modalities simultaneously during training, rather than only two at a time as ImageBind had done. Later works used both real and simulated tactile data to bind tactile images to the CLIP representation space as well [4, 11, 46, 49].

Training these representations required a curated dataset linking two or more distinct sensory modalities. Many datasets link images with text through datasets of human-captioned images [12, 39, 40]. Other datasets link sensory signals from two or more modalities, the majority of which consist of multi-sensory data from simulated environments [13, 14], with some works showing how such data could be used to train models for downstream tasks in either simulated [24] or real [15, 17] embodied environments. However, the ObjectFolder Benchmark showed that models fine-tuned with their proposed dataset of 100 *real* objects generally performed significantly better on real-world embodied tasks than those trained only with simulated data [16]. Many multi-sensory datasets were col-

Dataset	Obj.	Correlated Modal			
		RGB	3D	Audio	To
Feeling of Success [3]	106	✓	✓	✗	
VisGel [31]	195	✓	✗	✗	
Touch and Go [45]	3971	✓	✗	✗	
SSVTP [28]	N/A	✓	✗	✗	
HCT [11]	N/A	✓	✗	✗	
Greatest Hits [34]	N/A	✓	✗	✗	
RealImpact [6]	50	✓	✓	✓	
ObjectFolder 2.0 [15]	1000	✓	✓	✓	
ObjectFolder Real [16]	100	✓	✓	✓	
X-Capture (Ours)	500	✓	✓	✓	

Table 2. Comparing publicly available multi-sensc by their number of objects, their *correlated* sensory and the environments in which they are collected (C=Controlled, S=Simulation, T=Tabletop, and W=Wild). The X-Capture dataset includes the widest breadth of correlated sensory modalities of a dataset collected in the wild.

lected using automated setups to scale up data collection [3, 6, 28, 31, 35], but similar to those of ObjectFolder Benchmark, such setups are often costly to replicate, collected in only controlled environments, and naturally impose some constraints on the types of objects which can be scanned. See Table 1 for a comparison of X-Capture to relevant data collection setups and devices, and Table 2 for a comparison of our dataset to prior multi-sensory datasets. We include additional details in Appendix B.

Collecting multi-sensory data *in situ* can produce datasets with distributions matching expected test conditions of embodied agents more closely. The authors of the Touch and Go dataset press a GelSight sensor against objects in the wild while holding a webcam behind the sensor [45]. The authors of the Greatest Hits dataset [34] similarly took video and audio samples of a drumstick striking objects. But without measuring inputs explicitly, such as the contact location and the pressing or striking force, it is impractical to infer the essential input-output relationship between force and touch or impact and sound of an object.

Hand-held sensorized devices can combine the best of both worlds, offering the inter-sensor measurement consistency of sophisticated setups with the portability needed to collect data in situ. Multiple works designed handheld devices for capturing correlated vision and tactile signals [10, 11]. The authors of [2] designed a handheld device for collecting multi-sensory tactile data to characterize thermal properties of objects. Numerous works have proposed sensorized handheld devices that approximated robot end-effectors [5, 42, 48]. Humans used these devices to collect demonstration data by performing tasks with everyday objects, which could be used to train a policy for a robot arm or mobile manipulator [21, 41]. Our device brings the advantages of handheld data collection devices into a unified framework for measuring vision, touch, and audio data from objects, using commercially available sensors popular in many embodied learning applications.

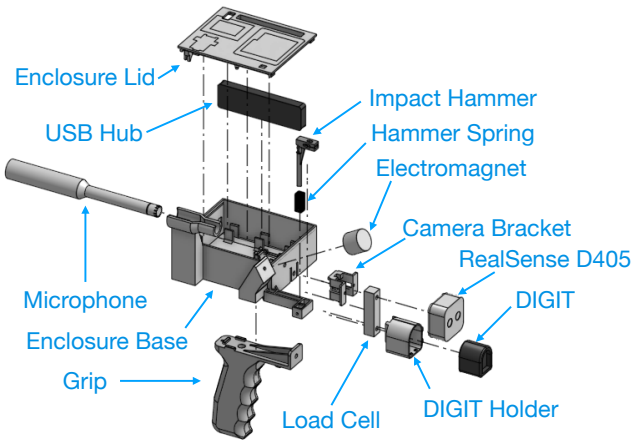


Figure 2. Exploded view of X-Capture. The device rigidly constrains all sensor assemblies into fixed relative poses on a compact chassis with an ergonomic grip. Wires and circuitry are not shown.

3. The X-Capture Device

The X-Capture device supports sensing correlated RGB, depth, tactile, and impact audio samples using a combination of six distinct sensors. It is compact and power-efficient, making it as convenient as possible to carry, only requiring control and power from a laptop. The chassis is constructed out of 3D printed PLA, with an ergonomic grip supporting an enclosure. The enclosure houses a USB hub that routes power and communications to all sensors and connects to a laptop with a single USB-C cable. Figure 2 and the supplementary video show how the device and sensor assemblies physically fit together. We detail the compositions of the sensor assemblies and how they are used below, organized by sensory modality. See Appendix C for additional details on the hardware, including internal photos, a bill of materials, and support for alternative sensors, and the supplementary video for a sequential breakdown of how the sensors and assemblies fit together.

3.1. Touch

For tactile sensing, X-Capture uses an assembly mounted on the front of the device consisting of a DIGIT sensor [30] attached to a load cell. Vision-based tactile sensors are popular for both benchmark datasets [16] and robotic applications [1, 28, 43]. We choose the DIGIT sensor for its commercial availability and relative durability [30]. However, the images the sensor collects during contact with an object are not only a function of the object's intrinsic geometry and stiffness at the point of contact, but also *extrinsic* factors such as the angle and the force at which the DIGIT sensor presses against the object. Thus, we design a novel assembly to explicitly measure this input-output rela-

tionship by annotating DIGIT tactile image readings in real time with relative angle and calibrated normal forces using an accelerometer and load cell, respectively. The device additionally uses the accelerometer to subtract the gravitational force exerted by the DIGIT sensor at the current device angle, dynamically re-zeroing the pressing force. This method uses low-cost, commercially available sensors to obviate the need to collect thousands of training points with a $6,000 force sensor for estimating calibrated forces from the vision-based sensor [9, 23, 51]. While collecting tactile data, our device automatically takes a snapshot of the DIGIT's tactile image at a pressing force within a 0.5N range of different target levels. We use 10, 15, and 20N as target levels during data collection.

3.2. Vision and Depth

X-Capture captures RGBD images of objects with a RealSense D405 stereoscopic depth camera. The camera is compact and lightweight, with a 7cm minimum distance for measuring depth, lending itself well to close-up capture of objects. X-Capture snapshots both the RGBD image and the accelerometer state simultaneously, such that the user can ensure a consistent angle of approach between the RGBD image and the tactile reading of a given point.

3.3. Audio

X-Capture collects audio samples from objects by striking them with an impact hammer and recording the resulting impact sound with a microphone situated behind the point of impact. The impact hammer is 3D printed from PLA, with a Thorlabs PK2JA2P1 piezo stack attached to a small steel rod at its tip. The piezo stack measures the impact force as it excites the object and produces the sound, in order to measure the input-output relation between the acting contact forces and resulting sound at each point. The base of the hammer fits into a 3D-printed elastomer base which functions as a spring. For each audio sample, the user pulls the hammer back until it adheres to an electromagnet mounted behind the hammer, storing potential energy in the elastomer base. After a pre-configured delay, the device automatically deactivates the electromagnet, releasing the hammer to strike the object in a silent and repeatable manner. The device records the impact sounds with a Dayton Audio EMM6 measurement microphone, which has a relatively uniform frequency response within the range of human-audible frequencies.

The impact hammer and the microphone's outputs are recorded by a HiFiBerry DAC2 ADC Pro board attached to a Raspberry Pi Zero 2W. The HiFiBerry board ensures that the hammer and microphone signal recordings are time-synchronized with minimal noise. The board supports high sample rates as well as digitally controlled gains, such that our device can dynamically and repeatably adjust the volume gain of each recording according to the loudness of each object's impact sounds. This ensures that the audio signal is recorded at a maximal level without clipping.

Finally, we design a custom PCB which provides precise power and input conditioning, as well as stable physical connections for all of the above electrical components. This novel assembly collects impact audio data at a comparably high fidelity to prior works [6, 16], while significantly optimizing power usage, size, and cost ($130 vs. $2,000), to improve portability and affordability.

3.4. User Interface and Workflow

Similar to the view shown at the right of Figure 1, the UI visualizes all modalities of data in one screen to provide important feedback during data collection, allowing for retakes and comparisons to previous captures. For each example, the user captures readings from each modality at one specific point on the object. The user first positions X-Capture's RGBD camera such that a target point on the object is centered in the image, at a depth 8 to 13 cm away from the camera, as measured by depth image and displayed on a bar below the image. The UI displays the captured RGB and depth images, with the target point marked by small crosshairs superimposed at the center of both images. The user then presses the DIGIT sensor against the same point, using both the crosshairs on the RGB image and a display of the current angle of the device's accelerometer as a guide to ensure contacting the object at the same point and angle. During contact, a bar below the tactile image displays the current pressing force. The user pushes gradually up to 20N with snapshots automatically collected at 10, 15, and 20N. Finally, the user positions the impact hammer to hover over the target point, initiates a recording, pulls the hammer back to the electromagnet, then holds the device still while the electromagnet releases the hammer to strike the object after a delay. The UI displays the recorded spectrogram of the impact audio, as well as a time-domain graph of the impulse measured by the hammer, so the user can verify that the hammer made a single, clean impulse. See the supplementary video for a demonstration of this workflow.

4. The X-Capture Dataset

To evaluate our device's data collection pipeline, we present a novel dataset of correlated vision, tactile, audio, and post-processed point cloud data from a total of 3000 points on 500 objects in real-world environments, collected in just under three weeks. Our multi-sensory dataset includes data from objects across a diverse class of materials, geometries, and functional uses. Aided by our device's flexible capabilities and portable design, we capture a wide breadth of objects encountered in everyday settings. We include further details of the objects and environments comprising our dataset in Appendix D.1 and a comprehensive comparison of our datasets to existing alternatives in Appendix B.

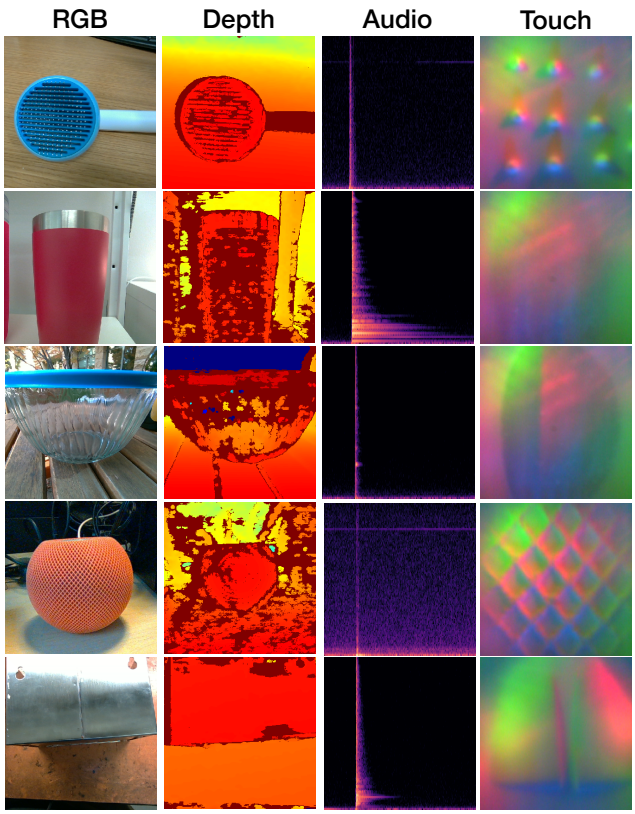


Figure 3. Example multi-sensory data points from the X-Capture dataset. Each row shows aligned multi-sensory data captured from a single point on an object in a distinct natural environment: (from left) RGB image centered on the point, depth image, impact audio spectrogram, and tactile image. (Objects, from top: Cat Brush, Insulated Steel Cup, Glass Storage Bowl, Computer Speaker, and Sheet Metal Container.)

4.1. Collection Procedure

We collect data across nine different environments– one indoor workspace with relatively controlled conditions, and eight in-the-wild locations, ranging from everyday indoor settings to a dynamic outdoor area. We capture 300 diverse objects in the indoor workspace, a quiet office with consistent lighting, providing a distribution baseline. We probe the remaining 200 objects in diverse natural environments, including a Kitchen, Bathroom, Home Office, Workshop, Bedroom, Laundry Room, Living Room, and an outdoor Picnic Table. We capture an equal number of objects in each in-the-wild environment.

For each object, we collect readings of each modality from six distinct points on the object. We choose point locations which cover the breadth of each object's surface and capture unique local features. Each of the six points has a corresponding RGBD image, impact audio recording, and tactile impressions captured at 10N, 15N, and 20N. We use our device and UI to manually register these sensory readings with each other at each point by following the procedure described in Section 3.4.

4.2. Dataset Labeling and Post-Processing

We provide a brief text description of each object, noting salient materials and the relevant state of the object (*e.g.*, if a can is full or empty). We further postprocess our data, using the RGB and depth data to estimate point clouds of objects and using the noted recording gains and hammer signals to normalize the audio. We include details of both these post-processing steps in Appendix D.2.

5. Experiments

We validate the usefulness of the sample dataset we collect with the X-Capture device with three popular multi-sensory benchmark tasks: cross-sensory retrieval, image generation, and point cloud generation. We also demonstrate how we can use our data to train an audio-based object detector. Unless otherwise noted, we randomly split our dataset into 400 training and 100 test objects. See Appendix E.4 for additional details on training and testing procedures.

5.1. Baselines

We evaluate recent cross-sensory representation frameworks on our dataset. The ImageBind framework [19] provides encoders for the RGB, audio, and depth modalities, each pretrained on web-scale datasets pairing images with each modality, but the framework lacks encoders for point clouds or tactile readings. In order to cover all of the modalities we provide in our dataset, we also evaluate the performance of a combination of pretrained encoders from different sources, which each specialize in a specific modality or cross-sensory representation. We encode RGB images with CLIP's [36] pretrained ViT-L encoder. For tactile images, we use the ViT-L encoder publicly released with [11], which has been pretrained on the TVL dataset. We encode our audio recordings with the Audio Spectrogram Transformer (AST) [20], with publicly released weights from pretraining on ImageNet [8] and AudioSet [18]. Finally, for point clouds, we use the PointBERT model [50] with publicly-released weights from pretraining on ULIP [44].

For both the ImageBind encoders and the ensemble of modality-specific encoders, we evaluate three different training configurations. We first evaluate the publicly-released out-of-the-box pretrained weights. Then we test fine-tuning all pretrained models on our dataset's training set, using two distinct formulations of the contrastive InfoNCE loss [33]. In the "Image Loss" formulation, we fine-tune according to a symmetric InfoNCE loss between images and each other available modality, the same technique used by ImageBind. In the "Cross-Sensory Loss" formulation, we fine-tune according to a symmetric InfoNCE loss between each pairing of *all* modalities, similar to the loss used among vision, language, and point cloud encoders in ULIP. For all configurations, we keep the encoders for the RGB image modality frozen during training.

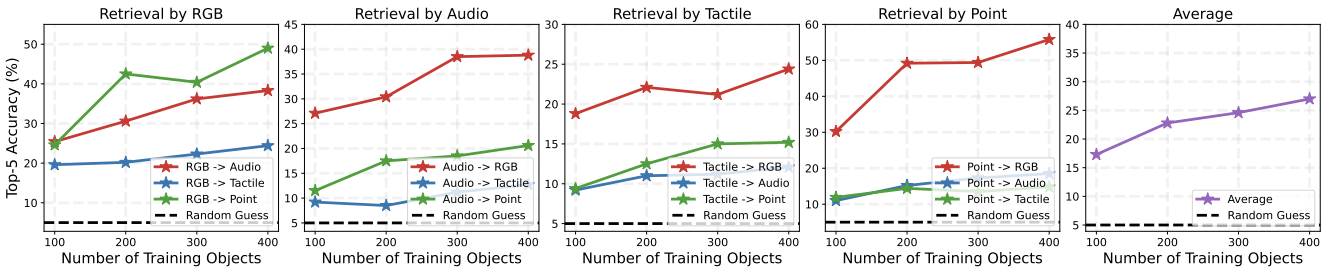


Figure 4. Comparing test retrieval performance of our cross-modal encoders trained with varying quantities of objects, with each plot grouping results by the query modality used for retrieval. The right-most plot shows an average across all modality combinations.

ImageBind	RGB→		Audio→		Depth→		Avg.
Top-5 Accuracy (%)	Audio	Depth	RGB	Depth	RGB	Audio	
Random Guess	5.0	5.0	5.0	5.0	5.0	5.0	5.0
Out-of-the-Box Pretrained	6.2	9.2	6.0	4.6	8.0	6.8	6.8
Fine-Tune w/ Image Loss	**39.0**	62.4	37.4	20.2	63.2	22.6	40.8
Fine-Tune w/ Cross-Sens. Loss	38.0	**64.8**	**41.0**	**21.2**	**64.4**	**24.4**	**42.3**

Table 3. Cross-sensory retrieval top-5 accuracies of ImageBind trained with different strategies using our dataset. The top and bottom column headers denote the query and retrieved modalities, respectively. Out-of-the-Box weights do not generalize well to our data, whereas Fine-Tuning with a Cross-Sensory Loss outperforms other configurations, across all modalities.

ImageBind	RGB→		Audio→		Depth→		Avg.
Top-1 Accuracy (%)	Audio	Depth	RGB	Depth	RGB	Audio	
Random Guess	16.7	16.7	16.7	16.7	16.7	16.7	16.7
Out-of-the-Box Pretrained	16.3	20.0	16.0	18.5	19.3	18.2	18.1
Fine-Tune w/ Image Loss	20.3	32.3	19.8	20.0	**45.8**	20.7	26.5
Fine-Tune w/ Cross-Sens. Loss	**24.0**	**34.5**	**20.3**	**22.0**	44.7	**22.3**	**28.0**

Table 4. Contact localization top-1 accuracies of ImageBind trained with different strategies using our dataset. The top and bottom column headers denote the query and retrieved modalities, respectively. Out-of-the-Box weights generalize poorly to our data, and Fine-Tuning with a Cross-Sensory Loss achieves best average performance.

5.2. Cross-Sensory Retrieval

Cross-sensory retrieval assesses models' abilities to connect multi-sensory information, akin to the human ability to intuitively link sight, sound, and touch, to make valuable inferences about unknown properties of objects. Consequently, it has become a common benchmark in cross-sensory learning [16, 19, 46]. We formulate the task as an inter-object classification task, testing each cross-sensory frameworks' performance as follows: given a randomly selected point from each of N objects, where $N = 100$ for our test dataset, can the framework correctly associate a sensory reading from one modality with another from the same point?

We show the top-5 accuracies for ImageBind in Table 3. Note that for 100 objects, the expected value of random selection would be 5% under our test conditions. While the encoders have been pretrained on very diverse datasets, their out-of-the-box weights struggle on our object-centric data, especially with the audio modality. Our results also suggest that using a full cross-sensory loss comparing all modalities directly provides a generally stronger representation on our data. We show the results for the cross-sensory encoder ensemble in Table 7 of Appendix E.1. Interestingly, in these results we see less of a clear performance advantage to training with the Cross-Sensory Loss versus the Image Loss, as compared to the advantage we observe in our ImageBind experiments. For modality pairings the ensemble and ImageBind have in common, such as RGB→Audio, we see similar results to those of ImageBind.

5.3. Cross-Sensory Contact Point Localization

Contact point localization assesses models' abilities to differentiate between sensory signals coming from different points on the *same* object. We thus formulate contact point localization as a classification task similar to cross-sensory retrieval, except that instead of comparing different sensory signals from a single point from one object to points of other objects, we compare different sensory signals from a single point to those of another modality from all M points on the same object, where $M = 6$ in our dataset.

We show the top-1 accuracies for ImageBind in Table 4. In this task, the expected value of random selection is $\sim 16.7\%$. Once again, ImageBind's pretrained weights struggle to outperform random chance on our object-centric data. However, this is clearly a difficult task even after fine-tuning. ImageBind seems to excel at differentiating object points with RGB and depth much more than it does with audio, perhaps because the impact sounds a single object produces at different points often have very minute differences. We see similar results for our cross-sensory encoder ensemble in Table 8 of Appendix E.2, where the association between RGB and point cloud is much stronger than associations between other modalities.

5.4. Scaling Law

We claim that the X-Capture device makes multi-sensory object-centric data collection more efficient than prior alternatives, allowing us to scale up the collection of a valuable dataset. Though we have already collected a sample dataset of 500 objects, we evaluate how a model's performance improves according to the quantity of our training objects we use during fine-tuning. The results in Figure 4

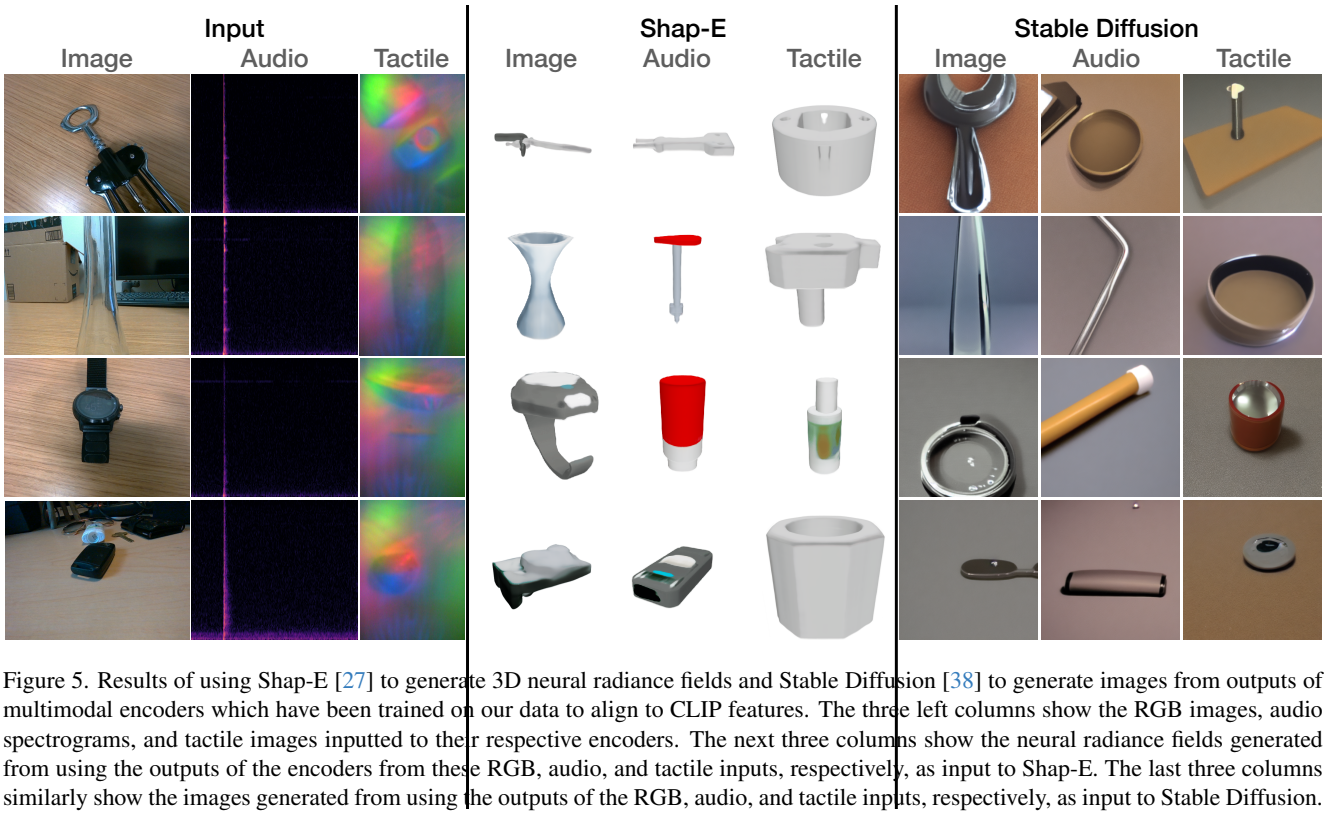


Figure 5. Results of using Shap-E [27] to generate 3D neural radiance fields and Stable Diffusion [38] to generate images from outputs of multimodal encoders which have been trained on our data to align to CLIP features. The three left columns show the RGB images, audio spectrograms, and tactile images inputted to their respective encoders. The next three columns show the neural radiance fields generated from using the outputs of the encoders from these RGB, audio, and tactile inputs, respectively, as input to Shap-E. The last three columns similarly show the images generated from using the outputs of the RGB, audio, and tactile inputs, respectively, as input to Stable Diffusion.

demonstrate that each modality continues to benefit from additional training examples, and the performance improvement seems far from plateauing at 400 objects, confirming the value of scaling up the X-Capture dataset even further.

5.5. Using X-Capture Data as a Pretraining Set

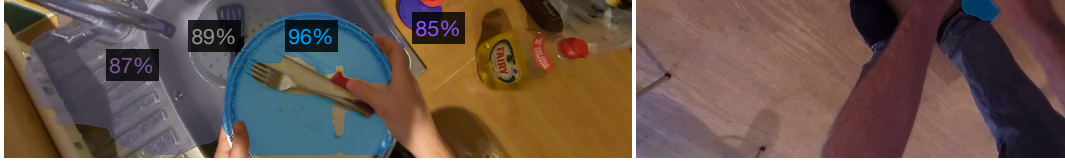
Multi-sensory object understanding is important in many different downstream applications, including robotic manipulation, anomaly detection, and generation. However, each potential application may involve unique data with a domain gap differentiating it from the data in existing large-scale datasets. Fortunately, such large-scale datasets can still be used to pretrain models, which can be fine-tuned on additional data for downstream applications. We evaluate whether data from X-Capture could be similarly helpful by using it as a pretraining set.

We evaluate cross-sensory retrieval on the real object data from the ObjectFolder Benchmark [16]. Though ObjectFolder's subject matter is quite similar to ours, there is a significant domain gap in how each modality is recorded. Their RGB images are background-less renderings from 3D scans of objects rather than real images. Their tactile data is from a GelSight sensor, which has a different texture detail and lighting conditions than the DIGIT. Finally, they collect impact sounds in an acoustically-treated room, with objects suspended by string to reduce the contact-damping usually affecting objects' impact sounds in real-world settings.

We evaluate the cross-sensory encoder ensemble's vision, audio, and tactile encoders trained with three different configurations using the cross-sensory loss. In "Our Pretrained Only" configuration, we train encoders with only X-Capture data. In "Fine-Tuned Only", we train the encoders only with data from the 70 real training objects from ObjectFolder. And in "Our Pretrained + Fine-Tuned", we pretrain the encoders with our data, then fine-tune on the real training objects from ObjectFolder. We evaluate cross-sensory retrieval on the 30 real test objects of ObjectFolder Benchmark and show the top-5 accuracies in Table 5. The expected value of random selection is 16.7%. We see that "Our Pretrained Only" model outperforms random chance, but still performs rather poorly in this zero-shot generalization, likely struggling to surmount the domain gap without fine-tuning on ObjectFolder's data. However, "Our Pretrained + Fine-Tuned" model outperforms the "Fine-Tune Only" model, suggesting that pretraining with the X-Capture dataset helps bridge the generalization gap in the low-data regime of the ObjectFolder Real training set. The improvement is evident in the audio modality, but not in the tactile modality, suggesting that the domain gap between different tactile sensors may be especially challenging.

5.6. X-to-2D/3D Generation

For humans, the sound or feel of an object can often conjure a mental image of the object. We thus evaluate the ability of



ObjectFolder-Real (30 Objects)	I
Top-5 Accuracy (%)	Auc
Random Guess	16
Our Pretrained Only	26
Fine-Tuned Only	25
Our Pretrained + Fine-Tuned	**38**

Table 5. Results of ablating and fine tuning with train ol dataset [16] for downstream e on the real test objects from O

our representations aligned vide a useful representatio plicit function generator ar We use our dataset to train align to the outputs of CLIP's ViT-L encoder of the corresponding image of each point. For this experiment, we use a mean-squared error (MSE) rather than contrastive loss to prioritize alignment to CLIP over cross-sensory association and differentiation. After training alignment on our training objects, we use the signals encoded from our held-out test objects as input to pretrained Shap-E [27] to generate 3D implicit functions and to pretrained Stable Diffusion [38] to generate images. Both models have been pretrained by their authors to use CLIP ViT-L features encoded from text to generate 3D implicit functions and images, respectively.

We show qualitative results for both image generation and 3D neural radiance field generation in Figure 5. The results vary substantially in quality, but they seem to provide some interesting insights into what features our encoders are able to glean from each respective sensory modality. We see that both shape and image generations from the image encoding tend to match the original object in both color and semantic features. Generations from audio encodings often successfully match the salient materials of the object, and in some cases, they show surprising semblance of unique geometric features of the object, such as in the cases of Shap-E's renderings from the audios of the bottle opener and the car key fob. Generations from tactile encodings seem to excel at matching the geometric features local to the contact point, such as the local curvature, and also occasionally match in material properties such as hardness. These results lend further evidence of the complementary information which can be inferred from different sensory modalities when interacting with objects.

5.7. Zero-Shot Audio-Based Object Detection

Similar to ImageBind [19], we use our dataset to train a CLIP-aligned audio embedding which can replace the input for the text-based Detic detection model [52]. We train our audio encoder on our dataset contrastively to ViT/B-32 CLIP features, then use embeddings from this encoder to prompt Detic's CLIP-based object detector. Though our dataset is collected with objects in the wild, all objects are captured in static configurations, whereas humans

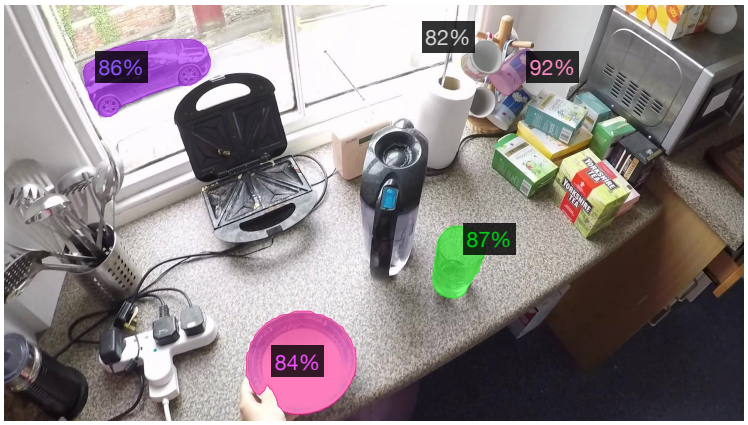


Figure 6. Still frame from prompting Detic [52] with our pretrained audio encoder's embedding of the real sound of the ceramic plate (at bottom, highlighted pink) being placed on the counter. Detic identifies the plate, as well as the ceramic mugs, drinking glass, and car, as likely sources of the sound. It successfully ignores appliances, cardboard tea boxes, and metal utensils.

and robots often perceive objects through dynamic interactions. In order to test generalization to audio from such natural dynamic interactions, we use clips from the EPIC-KITCHENS-100 dataset [7] of humans using their kitchens naturally, and select clips where a human impacts an object against a table or another object and use the audio embedding of the impact sound to prompt the Detic model. We show one such result in Figure 6 and additional results in Figure 11 of Appendix E.3. We include the original video clips with audio in our supplementary video. The detector mostly selects either the correct item or items of similar materials and acoustic properties. ImageBind does not provide weights from this task for comparison.

6. Conclusion and Limitations

We introduced X-Capture, an open-source and low-cost device for collecting multi-sensory data in the wild. Using X-Capture, we collected a sample dataset of correlated RGB, depth, audio, and tactile readings of 3,000 points from 500 objects in natural environments, enabling direct benchmarking of cross-sensory encoding frameworks and loss functions on retrieval and contact localization tasks. Our results suggest that cross-sensory representations can be strengthened by learning from object-centric data correlating as many sensory modalities as possible, and that pretraining on this data yields valuable representations that can be fine-tuned to improve performance on other object-centric tasks. However, a limitation of X-Capture is that it captures objects in static configurations of environments, whereas humans and robots learn about objects interactively and dynamically while manipulating them. We hope our work inspires new, perhaps automated, collection efforts to further scale up multi-sensory learning from real objects.

Acknowledgements. We thank Roger Clarke, Ryan Williams, Anirudh Jain, Mark Rau, and Fernando Lopez-Lezcano for advice with the hardware design, Klemen Kotar, Le Xue, Stephen Tian, and Weiyu Liu for valuable conceptual discussions, and Andrej Krevl and Matt Wright for their facility support. This work is in part supported by NSF CCRI #2120095 and RI #2338203 and ONR MURI N00014-22-1-2740. S. Clarke is supported by the Meta PhD Fellowship.

References

[1] Bo Ai, Stephen Tian, Haochen Shi, Yixuan Wang, Cheston Tan, Yunzhu Li, and Jiajun Wu. Robopack: Learning tactile-informed dynamics models for dense packing. *arXiv preprint arXiv:2407.01418*, 2024. 3

[2] 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. 3

[3] 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? In *Conference on Robot Learning*, pages 314–323. PMLR, 2017. 3, 13

[4] 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. 2

[5] Cheng Chi, Zhenjia Xu, Chuer Pan, Eric Cousineau, Benjamin Burchfiel, Siyuan Feng, Russ Tedrake, and Shuran Song. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. In *Proceedings of Robotics: Science and Systems (RSS)*, 2024. 3

[6] Samuel Clarke, Ruohan Gao, Mason Wang, Mark Rau, Julia Xu, Jui-Hsien Wang, Doug L James, and Jiajun Wu. Realimpact: A dataset of impact sound fields for real objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1516–1525, 2023. 2, 3, 4, 13, 17

[7] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)*, 130:33–55, 2022. 8, 18

[8] 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. Ieee, 2009. 5

[9] Won Kyung Do, Bianca Jurewicz, and Monroe Kennedy. Densetact 2.0: Optical tactile sensor for shape and force reconstruction. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 12549–12555. IEEE, 2023. 4

[10] 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*, pages 26529–26539, 2024. 3

[11] 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. In *Forty-first International Conference on Machine Learning*, 2024. 2, 3, 5, 13

[12] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[13] Chuang Gan, Jeremy Schwartz, Seth Alter, Damian Mrowca, Martin Schrimpf, James Traer, Julian De Freitas, Jonas Kubilius, Abhishek Bhandwaldar, Nick Haber, et al. Threedworld: A platform for interactive multi-modal physical simulation. *arXiv preprint arXiv:2007.04954*, 2020. 2

[14] 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. *arXiv preprint arXiv:2109.07991*, 2021. 2

[15] Ruohan Gao, Zilin Si, Yen-Yu Chang, Samuel Clarke, Jeannette Bohg, Li Fei-Fei, Wenzhen Yuan, and Jiajun Wu. Objectfolder 2.0: A multisensory object dataset for sim2real transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10598–10608, 2022. 2, 3, 13

[16] Ruohan Gao, Yiming Dou, Hao Li, Tanmay Agarwal, Jeannette 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*, pages 17276–17286, 2023. 2, 3, 4, 6, 7, 8, 13

[17] Ruohan Gao, Hao Li, Gokul Dharan, Zhuzhu Wang, Chengshu Li, Fei Xia, Silvio Savarese, Li Fei-Fei, and Jiajun Wu. Sonicverse: A multisensory simulation platform for embodied household agents that see and hear. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 704–711. IEEE, 2023. 2

[18] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017. 5

[19] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023. 2, 5, 6, 8

[20] Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*, 2021. 5

[21] Huy Ha, Yihuai Gao, Zipeng Fu, Jie Tan, and Shuran Song. Umi-on-legs: Making manipulation policies mobile with manipulation-centric whole-body controllers. In *8th Annual Conference on Robot Learning*, 2024. 3

[22] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 17

[23] Erik Helmut, Luca Dziarski, Niklas Funk, Boris Belousov, and Jan Peters. Learning force distribution estimation for the gelsight mini optical tactile sensor based on finite element analysis. *arXiv preprint arXiv:2411.03315*, 2024. 4

[24] Yining Hong, Zishuo Zheng, Peihao Chen, Yian Wang, Junyan Li, and Chuang Gan. Multiply: A multisensory object-centric embodied large language model in 3d world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26406–26416, 2024. 2

[25] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024. 1

[26] Roland S Johansson and J Randall Flanagan. Coding and use of tactile signals from the fingertips in object manipulation tasks. *Nature Reviews Neuroscience*, 10(5):345–359, 2009. 1

[27] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*, 2023. 7, 8

[28] 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 *CVPR*, 2023. 3, 13

[29] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 17

[30] Mike Lambeta, Po-Wei Chou, Stephen Tian, Brian Yang, Benjamin Maloon, Victoria Rose Most, Dave Stroud, Raymond Santos, Ahmad Byagowi, Gregg Kammerer, et al. Digit: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation. *IEEE Robotics and Automation Letters*, 5(3):3838–3845, 2020. 3, 13

[31] 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. 3

[32] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 17

[33] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 5

[34] Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H Adelson, and William T Freeman. Visually indicated sounds. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2405–2413, 2016. 3, 13

[35] Dinesh K Pai, Kees van den Doel, Doug L James, Jochen Lang, John E Lloyd, Joshua L Richmond, and Som H Yau. Scanning physical interaction behavior of 3d objects. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 87–96, 2001. 2, 3, 13

[36] 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. 2, 5

[37] Mark Rau, Orchisama Das, and Elliot K Canfield-Dafilou. Improved carillon synthesis. In *Proceedings of the 22nd international conference on digital audio effects, Birmingham, UK*, pages 1–8, 2019. 17

[38] 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. 7, 8

[39] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 2

[40] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 2

[41] 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. 3

[42] 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. 3

[43] Sudharshan Suresh, Haozhi Qi, Tingfan Wu, Taosha Fan, Luis Pineda, Mike Lambeta, Jitendra Malik, Mrinal Kalakrishnan, Roberto Calandra, Michael Kaess, et al. Neuralfeels with neural fields: Visuotactile perception for in-hand manipulation. *Science Robotics*, 9(96):eadl0628, 2024. 3

[44] Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1179–1189, 2023. 2, 5

[45] 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. 3, 13

[46] Fengyu Yang, Chao Feng, Ziyang Chen, Hyoungseob Park, Daniel Wang, Yiming Dou, Ziyao Zeng, Xien Chen, Rit Gangopadhyay, Andrew Owens, et al. Binding touch to everything: Learning unified multimodal tactile representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26340–26353, 2024. 2, 6

[47] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024. 17

[48] 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. 3

[49] 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. 2

[50] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19313–19322, 2022. 5

[51] Wenzhen Yuan, Siyuan Dong, and Edward H Adelson. Gelsight: High-resolution robot tactile sensors for estimating geometry and force. *Sensors*, 17(12):2762, 2017. 4

[52] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *European conference on computer vision*, pages 350–368. Springer, 2022. 8, 18

X-Capture: An Open-Source Portable Device for Multi-Sensory Learning

The supplementary materials consist of:

A. A video showing the design a X-Capture device
B. A comprehensive comparison of the to prior multi-modal datasets and de
C. Additional details on the device ha the internals, a bill of materials, a support for alternative sensors
D. Additional information on the X more details on the objects and env as the postprocessing steps
E. Additional experimental results and

A. Supplementary Video

The included video (`XCaptureVide` breakdown of the layout and design of vice, then shows the process of using th data from an object, as well as the fee ceives from each sensor on the user in process. We also include qualitative exa from the generation and audio-based det described in Section 5.6 and 5.7, respecti

Specific portions of the video referenc at the following timestamps:

- [**00:49**] A breakdown of the hardw scribed in Section 3
- [**02:40**] A demonstration of using t interface to collect data through scribed in Section 3.4
- [**06:36**] Example video clips, with audio-based detection experiment tion 5.7

B. Comparison to Prior Multi-Sensory Datasets and Devices

While there are prior object-centric multi-sensory datasets, our dataset is the first of its kind to correlate RGB, depth, impact audio, and tactile sensing at a point-level of objects in the wild. This is made possible by the design of the X-Capture device integrating both existing and novel sensor assemblies of difference sensory modalities into a single portable device. We compare to relevant prior datasets and prior devices in more detail below.

Prior Object-Centric Multi-Sensory Datasets We compare the X-Capture dataset to prior works across three dimensions: the quantity of data, the sensory modalities included, and the data collection environment in Table 2 of Section 2. Many datasets correlate only two or three sensory modalities. While some datasets also provide correlated RGB, depth, audio, and tactile data, the X-Capture

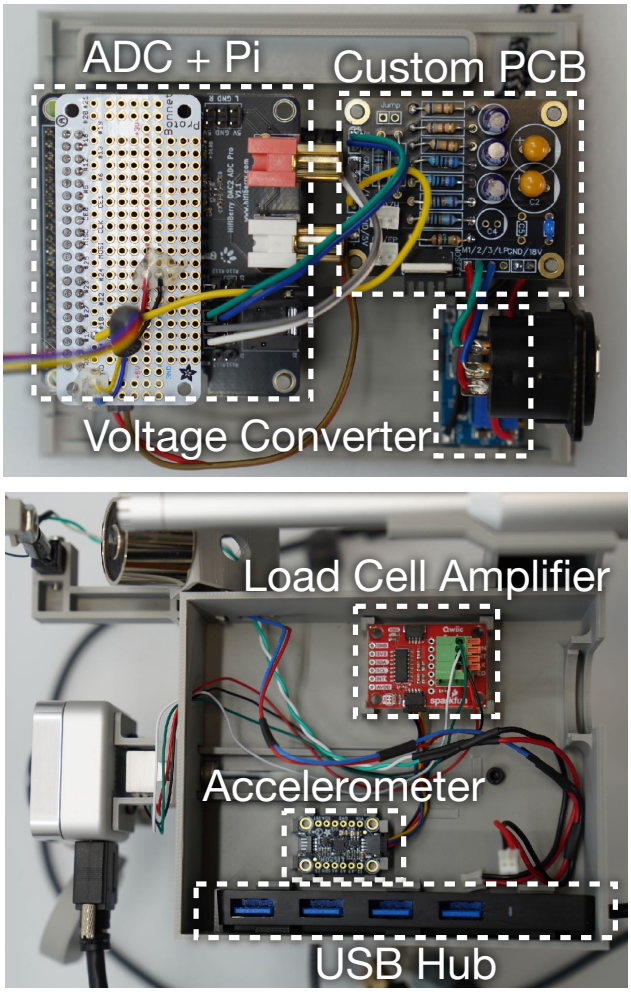


Figure 7. Photos of the X-Capture device internals. (**Top**) The lid of the device has a white prototyping board stacked on top of a HiFiBerry DAC2 ADC Pro, which is also stacked on top of a Raspberry Pi Zero 2W single-board computer. Our custom PCB has circuitry for powering and filtering both the impact hammer and the microphone. The voltage converter, partially occluded in this photo by the microphone jack, provides 18V power for the microphone from USB-powered 5V. (**Bottom**) The base of the device houses an amplifier for the load cell signal, as well as an accelerometer. The USB hub provides USB A ports for the RealSense D405, DIGIT, Raspberry Pi, and voltage converter, connecting them all through a USB C connection to a laptop or desktop computer.

device supports capturing all of these modalities in a correlated fashion *in the wild*.

With the exception of SSVTP [28] and ObjectFolder [15, 16], these datasets do not explicitly correlate at the *point* level of an object. For example the Feeling of Success [3], Touch and Go [45], and HCT [11] correlate tactile images with RGB images where the contact region is specifically occluded by the sensor. Greatest Hits [34] includes videos of a wooden drumstick striking objects and surfaces, where exact contact location is not obvious from the videos due to motion blur at 30 frames per second. We use the X-Capture device to manually register a reading of all four modalities to the same point, such that we can use our data to learn object-centric multi-sensory representations at a point-level resolution.

Of all these datasets, those of ObjectFolder are most similar to ours in covering all four sensory modalities and correlating them at a point level. While ObjectFolder 2.0 [15] includes more objects, all objects are *virtual* and all sensory readings are *simulated* from these virtual objects. ObjectFolder Real [16] has 100 *real* objects, but sensory readings from each modality are collected in *controlled* environments: the authors collect audio from objects suspended in a semi-anechoic chamber, tactile readings from objects rigidly fixed to a robot table top, and RGB images from objects on a turn-table inside a light-box. Note that even for these 100 real objects, the *point-correlated* RGB and depth readings are *simulated* as well, using renderings of the textured 3D model from a 3D scan. Finally, extending ObjectFolder beyond the 100 real objects required purchasing at least $11,000 of equipment which must be powered by a wall socket or generator. X-Capture is powered by a laptop and collects additional data at a similar fidelity to ObjectFolder *in the wild* for $1,000.

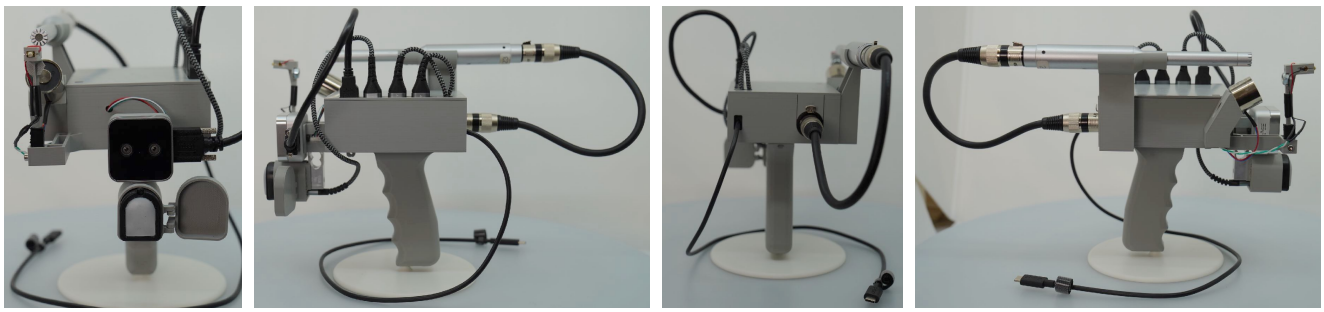
Prior Multi-Sensory Data Collection Devices We compare the X-Capture device to relevant data collection devices, focusing on devices which lend themselves to capturing object-centric data of one or more modalities in addition to vision, in Table 1 of Section 2. The UBC ACME [35] and RealImpact setup [6] both used large, stationary setups where objects were placed in a central position for scanning. Neither were designed to be portable for scanning objects *in situ*. ObjectFolder Real [16] used separate stationary setups for each modality which were also not portable. Most comparable to our device in terms of portability and cost is the novel device introduced with the TVL dataset [11]. The device includes a Logitech webcam and a DIGIT sensor fixed to the same chassis such that the webcam is pointed at an oblique angle toward the DIGIT's contact area. This allows for strict temporal alignment between the webcam video and the DIGIT images, but at the cost of the DIGIT occluding contact area from the webcam during contact. While the webcam is also capable of recording ambient audio during contact with the DIGIT, there is no provision for estimating objects' impulse responses at various points by measuring the input-output relation between a precise impact and the sound thereof with the calibrated microphone of our setup.

C. Additional Details on the X-Capture Device Hardware

The device's enclosure contains many important components including a Raspberry Pi, a HiFiBerry DAC2 ADC Pro, a custom PCB, a USB hub, a voltage converter, an amplifier for the load cell, and an accelerometer. We show photos of the inside of the enclosure in Figure 7. We also include photos of the X-Capture device's exterior from different views in Figure 8.

The X-Capture device can be replicated with consumer-grade tools, a 3D printer, and a soldering iron. We include a full bill of materials in Table 6, showing the total cost of parts is less than $1000 (not including shipping costs or taxes).

Alternative Sensor Support We specifically chose the sensors we used on the X-Capture device for their properties that are well-suited to collecting object-centric data in the wild. However, to support additional applications of multi-sensory data capture, we provide designs of mounts for alternative sensors. For vision, we also provide a mount design which supports the RealSense D415, D435, and D435i RGBD cameras, which each have longer minimum and maximum depth ranges than the D405 camera we chose for our dataset and may be especially well-suited for collecting larger objects or scene-centric datasets. For tactile sensing, we provide a mount design which supports the GelSight Mini vision-based tactile sensor, similar to the tactile sensor used in ObjectFolder Real [16]. The GelSight Mini produces high-quality tactile images, at slightly higher cost and lower durability than the DIGIT [30]. We show the X-Capture device with both the RealSense 435 and the GelSight Mini mounted on it in Figure 9.



t, back, and right side view.

Figure 9. The X-Capture device in a supported alternative configuration, with a RealSense D435 RGBD camera for vision and a GelSight Mini for tactile sensing. The device supports a choice of RealSense D405, D415, D435, or D435i for camera and the DIGIT or GelSight Mini for tactile sensor.

D. Additional Details on the X-Capture Dataset

D.1. Objects and Environments

The X-Capture Dataset features a diverse range of objects and environments, which we detail below. We show photos as well as example object images from each environment in Figure 10.

The *indoor workspace* features consistent artificial lighting and minimal noise. Objects consist of diverse materials such as glass, plastic, metal, and ceramic, with varying textures and geometries.

The *kitchen* environment has mixed natural and artificial lighting, creating moderate shadows and highlights. Ambient noises include faint sounds such as a humming refrigerator or occasional outside noise. Objects are primarily food-related, such as packaging, utensils, and glassware, made from cardboard, metal, glass, and plastic.

The *bathroom* environment features artificial lighting with moderate shadows. The matte ceramic sink countertop reduces reflections, and the enclosed space causes slight reverberation. Objects include personal care products with smooth, cylindrical shapes, made from glass and plastic.

The *home office* environment is lit by natural light from windows, supplemented by artificial light. Objects include technology devices (*e.g.*, headphones, remotes), stationery, and decorative items. Audio occasionally includes aquarium bubbling or outdoor noises.

The *workshop* environment is brightly lit with overhead lighting. Objects include tools, hardware, and electronics, made from durable materials like metal and plastic.

The *bedroom* environment has warm artificial lighting that casts strong shadows. Faint sounds from fans or neighboring rooms may be present. Objects include books, plants, and clothing accessories, made from fabric, glass, plastic, and wood.

The *laundry room* features bright, uniform artificial lighting that minimizes shadows. Objects are predominantly cleaning supplies such as detergents, sprays, and bleaches, in smooth plastic containers with vivid packaging. Ambient noises include faint mechanical sounds, but data was collected with laundry machines turned off.

The *living room* environment has soft artificial lighting from overhead and accent lights. Surfaces include metal shelves and wooden tables, with objects like decor, books, plants, and electronics.

The *picnic table* environment features bright, diffuse outdoor lighting. Objects include food items, food storage containers, cutlery, and outdoor accessories, made from plastic, glass, metal, and organic textures (*e.g.*, fruit skins). Ambient noises include occasional distant activity or building hums.

D.2. Postprocessing

While our RGB, depth, and touch data can be used in their raw state for many useful learning tasks, we postprocess the audio data to normalize differences between recordings

Part	Unit Cost (USD)	Quantity	Total Cost (USD)
Sensors			
RealSense D405	272	1	272
DIGIT	350	1	350
LIS3DH Accelerometer	4.95	1	4.95
10kg Load Cell	12.95	1	12.95
Piezo Stack	79	1	79
Dayton Audio EMM6 Measurement Microphone	54.98	1	54.98
Cables			
1ft MicroUSB	2.66	3	7.99
USB Micro B with screws	7.99	1	7.99
4-port USB C hub, 2ft	9.99	1	9.99
1ft XLR Cable	6.49	1	6.49
100mm JST connector	0.75	2	1.5
50mm Qwiic Cable	0.95	1	0.95
Qwiic Breadboard Jumper	1.6	1	1.6
XLR Female Panel Mount	7.63	1	7.63
Female/Male Jumper Cables, 6 inch	1.95	1	1.95
Breakout Boards			
Raspberry Pi Zero 2W	15	1	15
Raspberry Pi Headers	1.05	1	1.05
HiFiBerry DAC2 ADC Pro	74.9	1	74.9
Qwiic Scale NAU7802	16.5	1	16.5
PermaProto Bonnet	4.5	1	4.5
Custom Electronics			
Aluminum electrolytic capacitor, 33uF 50V 20% Resistor, 1k Ohm	1.01	2	2.02
Resistor, 100k Ohm	0.24	3	0.72
Resistor, 150k Ohm	0.24	3	0.72
Tantalum capacitor, 33uF	2.51	2	5.02
Ceramic capacitor, 3.3uF	0.5	1	0.5
JST connection header	0.14	2	0.28
Resistor, 680 Ohm	0.1	3	0.3
N-channel MOSFET	1.84	1	1.84
Heat Sink	0.5	1	0.5
Custom PCB	3.64	1	3.64
Resistor, 1M Ohm	0.1	1	0.1
Chassis			
Bambu Basic PLA, Gray 1kg	19.99	0.8	15.99
Ninjatek Edge Filament, Black 0.5kg	56.29	0.1	5.63
Various ISO Metric Fasteners	0.21	7	1.47
TOTAL			**971.27**

Table 6. Full bill of materials for building the X-Capture device. Prices do not include shipping costs or taxes.

with respect to their volume gain settings and the forces of the hammer impacts. We also use the RGB and depth data to generate object point clouds for some experiments.

Normalizing Audio During data collection, the X-Capture device dynamically adjusts the recording gains of both the microphone and the impact hammer for each recording to ensure high signal while also preventing

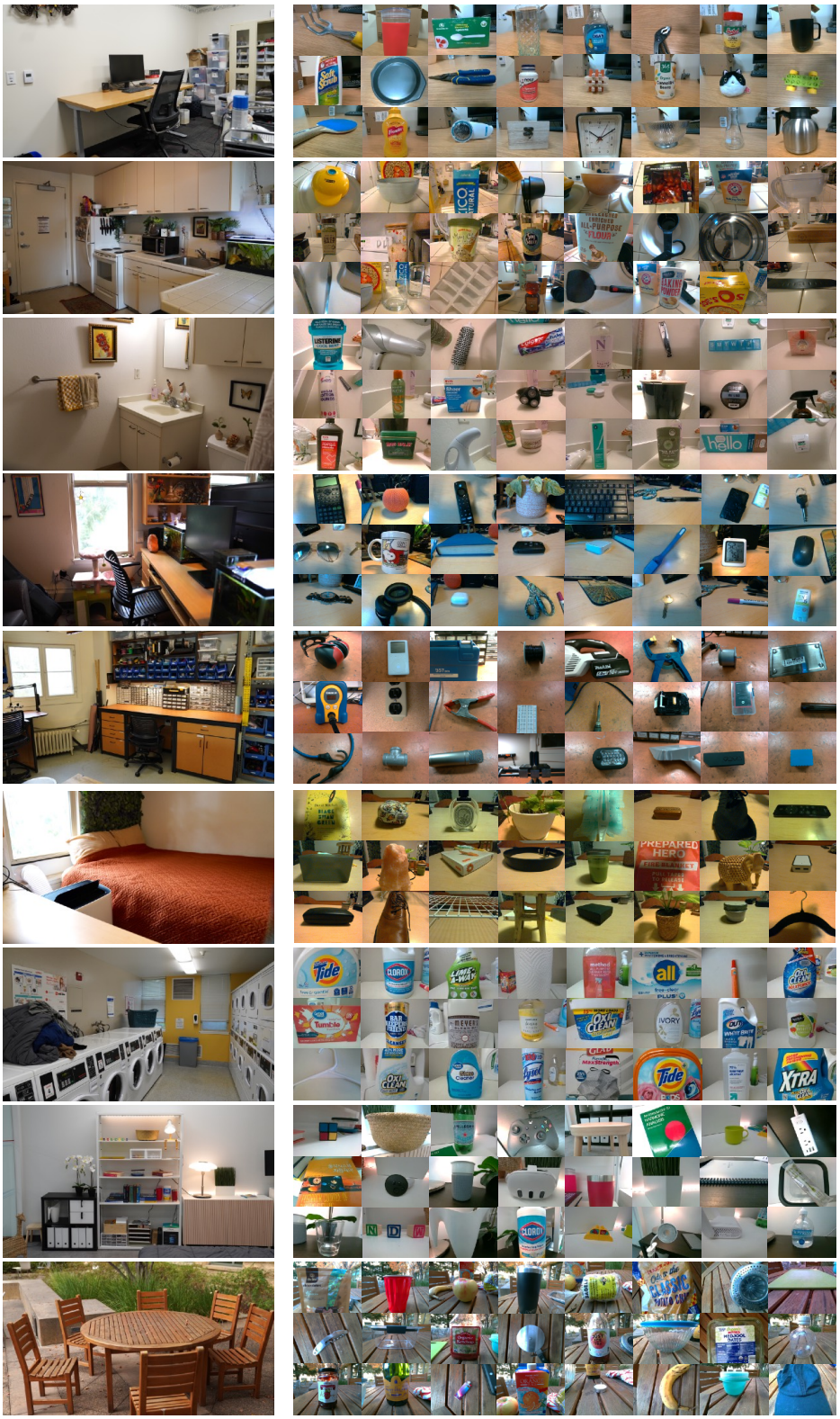


Figure 10. The X-Capture dataset is captured across nine diverse in-the-wild environments. (**Left**) Photos of the different environments in which data was collected. From top: indoor workspace, kitchen, bathroom, home office, workshop, bedroom, laundry room, living room, and picnic table. (**Right**) Montages each show 24 distinct objects from each corresponding environment in the left column of the same row. All images in the montages are object images directly lifted from the X-Capture dataset, as captured by the device.

clipping. After collection, we use the annotated gains for each recording in order to scale all recordings to a common gain. For characterizing the input-output relationship of striking the objects, many works deconvolve impact hammer signals from microphone recordings to estimate objects' impulse responses [6, 37]. However, these works often record more rigid and homogeneous objects with a rigidly positioned impact hammer rig. We found that our hammer signals were too complex to lend themselves to deconvolution without producing filtered noise artifacts. This may be because we record soft, heterogeneous, and articulated objects with a *hand-held* impact hammer rig. Thus, in order to correct for differences in strike force among our audio samples, we simply divide our normalized audio recordings by the peak of their corresponding gain-normalized hammer signals.

Point Cloud Extraction Though the depth readings from the RealSense D405 can be somewhat sparse and noisy depending on the object and capture conditions, we estimate a coherent object point cloud from each captured depth image using the following steps. First, we use DepthAnythingV2 (DAV2) [47] on the captured RGB image to estimate a smooth, dense depth map. Since this predicted depth map lacks a sense of scale, we use least squares regression to estimate a scale and offset which best aligns the prediction map with our sparse real depth recording. To segment the target object from the background and other objects in the scene, we use the Segment Anything Model (SAM) [29] on the overlaid RGB image. Since we collect each RGBD image with the center point on some region of the target object, we query SAM with a small disk of points around the center pixel. SAM provides three mask proposals, and we select the mask with the largest area where the center pixel is activated, to favor selecting an entire composite target object rather than an individual section. We apply an erosion to the SAM mask to eliminate ambiguous outlier points. We then apply this final segmentation mask to the adjusted depth map prediction to construct the final object point cloud.

E. Additional Experimental Results and Examples

E.1. Cross-Sensory Retrieval

We show additional results from the experiment described in Section 5.2 for the multi-modal encoder ensemble in Table 7.

E.2. Cross-Sensory Contact Point Localization

We show additional results from the experiment described in Section 5.3 for the multi-modal encoder ensemble in Table 8.

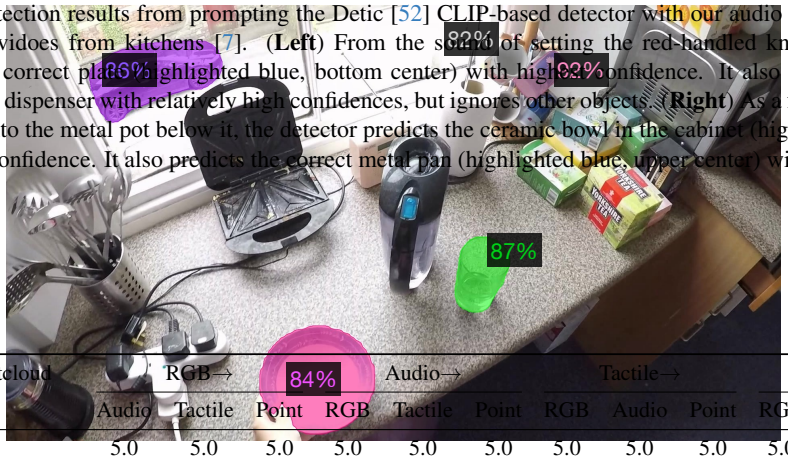
E.3. Zero-Shot Audio-Based Object Detection

We show additional results from the experiment described in Section 5.7 in Figure 11.

E.4. Training and Testing Details

During training for each experiment, we augment the image modality using strong image augmentations as in MoCo [22]. We use all other modalities as is, without augmentation. We train with a batch size of 64, using the AdamW optimizer [32] with a learning rate of 10^{-5}. For all experiments except those in Section 5.5, we train for 500 epochs. For the experiments in Section 5.5, to avoid overfitting on the small fine-tuning set from ObjectFolder Real, we train for only 50 epochs on each dataset. During evaluations for all experiments, we evaluate each model on the entire test set with five different random samplings of object points and report the average to reduce variance.

Figure 11. Additional detection results from prompting the Detic [52] CLIP-based detector with our audio embeddings of natural impact sounds from egocentric videos from kitchens [7]. (**Left**) From the sound of setting the red-handled knife on the plate, the detector successfully predicts the correct plate (highlighted blue, bottom center) with highest confidence. It also predicts the other plate, sink, and body of the handsoap dispenser with relatively high confidences, but ignores other objects. (**Right**) As a failure case, from the sound of the metal pan stacking onto the metal pot below it, the detector predicts the ceramic bowl in the cabinet (highlighted gray at upper right of the image) with highest confidence. It also predicts the correct metal pan (highlighted blue, upper center) with relatively high confidence.



RGB-Audio-Tactile-Pointcloud	RGB→			Audio→			Tactile→			Point→			Average
Top-5 Accuracy (%)	Audio	Tactile	Point	RGB	Tactile	Point	RGB	Audio	Point	RGB	Audio	Tactile	
Random Guess	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0
Out-of-the-Box Pretrained	6.0	5.4	5.4	4.8	5.2	5.4	5.8	6.0	4.8	6.2	5.6	5.4	5.5
Fine-Tune w/ Image Loss	34.0	23.3	**52.7**	**40.2**	12.5	17.5	23.5	**14.4**	11.9	**57.9**	15.6	**14.8**	26.5
Fine-Tune w/ Cross-Sensory Loss	**38.3**	**24.4**	49.0	38.8	**12.7**	**20.6**	**24.4**	12.1	**15.2**	55.8	**18.5**	**14.8**	**27.0**

Table 7. Cross-sensory retrieval top-5 accuracies of the cross-sensory encoder ensemble trained with different strategies using our dataset. The top and bottom column headers denote the query and retrieved modalities, respectively. The encoders' Out-of-the-Box weights do not generalize well to our data across modalities, though Fine-Tuning performance of each loss type varies by modality.

RGB-Audio-Tactile-Pointcloud	RGB→			Audio→			Tactile→			Point→			Average
Top-1 Accuracy (%)	Audio	Tactile	Point	RGB	Tactile	Point	RGB	Audio	Point	RGB	Audio	Tactile	
Random Guess	16.7	16.7	16.7	16.7	16.7	16.7	16.7	16.7	16.7	16.7	16.7	16.7	16.7
Out-of-the-Box Pretrained	16.0	17.2	16.7	16.7	16.7	16.8	16.7	13.9	16.0	15.8	17.7	15.8	16.3
Fine-Tune w/ Image Loss	19.8	19.3	**35.1**	19.4	**20.7**	19.6	**24.3**	**19.4**	20.5	**44.4**	19.7	20.3	23.5
Fine-Tune w/ Cross-Sensory Loss	**20.7**	**23.6**	31.5	**22.2**	18.2	**21.5**	22.9	17.1	**25.6**	41.2	**22.0**	**24.7**	**24.3**

Table 8. Contact localization top-1 accuracies of the cross-sensory encoder ensemble trained with different strategies using our dataset. The top and bottom column headers denote the query and retrieved modalities, respectively. Their Out-of-the-Box weights generalize poorly to our data across all modalities, but the highest performance loss for Fine-Tuning varies by modality.