

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

## Supplementary Material

The supplementary materials consist of:

- A video showing the design and usage of the X-Capture device
- A comprehensive comparison of the X-Capture dataset to prior multi-modal datasets and devices
- Additional details on the device hardware: photos of the internals, a bill of materials, and information on support for alternative sensors
- Additional information on the X-Capture dataset: more details on the objects and environments, as well as the postprocessing steps
- Additional experimental results and details

### A. Supplementary Video

The included video (XCaptureVideo.mp4) shows a breakdown of the design of the X-Capture device, then shows the process of using the device to collect data from an object, as well as the feedback the user receives from each sensor on the user interface during this process. We also include qualitative examples, with audio, from the generation and audio-based detection experiments described in Section 5.6 and 5.7, respectively.

Specific portions of the video referenced in the text begin at the following timestamps:

- [00:49] A breakdown of the hardware assemblies described in Section 3
- [02:40] A demonstration of using the device and user interface to collect data through the workflow described in Section 3.4
- [06:36] Example video clips, with sound, from the audio-based detection experiment described in Section 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 different 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

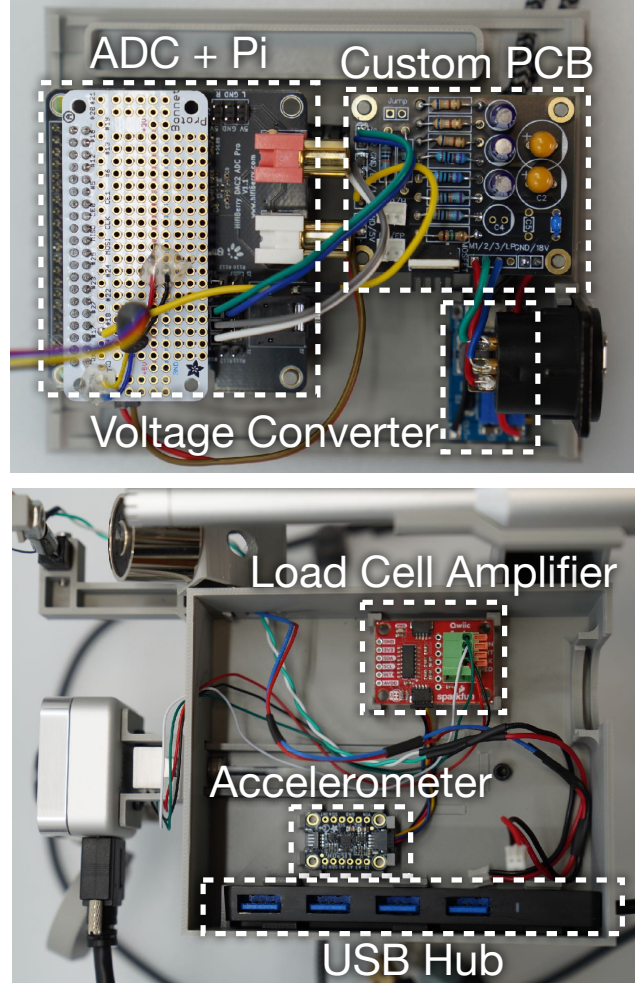


Figure 7. Photos of the X-Capture device internals. **(Top)** The lid of the device has a white prototyping board stacked on top of a HiFiBerry DAC2 ADC Pro, which is also stacked on top of a Raspberry Pi Zero 2W single-board computer. Our custom PCB has circuitry for powering and filtering both the 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.

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

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 generated 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.

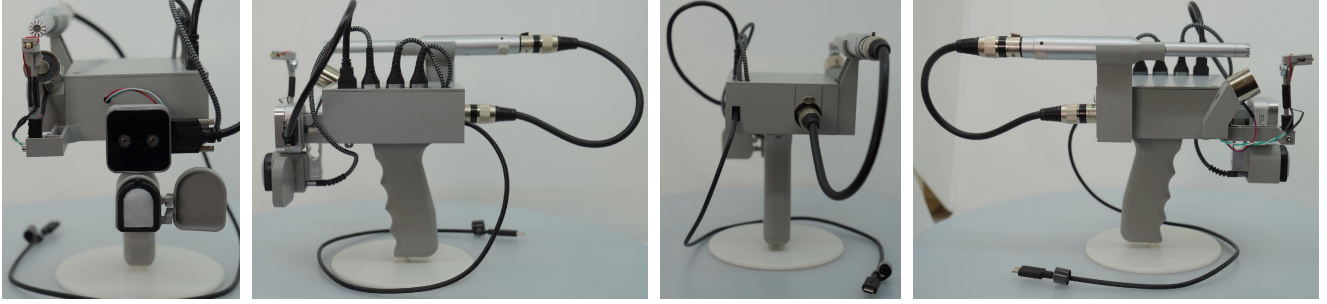


Figure 8. Different views of the X-Capture device. (From left) Front, left, back, and right side view.



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.

Each of these environments has unique acoustic properties, with each having unique static noises (e.g. ventilation), and some having occasional dynamic noises (e.g. voices). While our impact audio includes these static noises from each environment, we avoided collecting audio recordings with dynamic noises.





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, picnic table, kitchenette, patio, and lecture hall. **(Right)** Montages each show 18 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.



Part	Unit Cost (USD)	Quantity	Total Cost (USD)
<b>Sensors</b>			
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
<b>Cables</b>			
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
<b>Breakout Boards</b>			
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
<b>Custom Electronics</b>			
Aluminum electrolytic capacitor, 33uF 50V 20Resistor, 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
<b>Chassis</b>			
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
<b>TOTAL</b>			<b>971.27</b>

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

## D.2. Sensor Variations

Many sensors, such as the RealSense D405 are manufactured with high repeatability, such that different instances of the same sensor produce indistinguishable readings. How-

ever, the DIGIT sensor’s soft skin can be prone to manufacturing differences, as well as differences created by wear and tear. We collect 475 of our objects with one DIGIT sensor, and collect 125 with another. In our experiments, these objects are split randomly across the train-test split.

RGB-Audio-Tactile-Pointcloud Top-5 Accuracy (%)	RGB→			Audio→			Tactile→			Point→			Average
	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	4.8	4.6	6.0	6.0	4.6	5.2	5.2	4.4	6.0	4.4	4.0	4.6	5.0
Fine-Tune w/ Image Loss	<b>44.2</b>	<b>26.7</b>	<b>54.8</b>	<b>44.2</b>	18.5	21.3	<b>29.0</b>	17.9	16.2	<b>61.2</b>	23.5	17.7	<b>31.3</b>
Fine-Tune w/ Cross-Sensory Loss	40.2	24.2	50.8	44.0	<b>20.8</b>	<b>22.5</b>	26.2	<b>19.4</b>	<b>19.0</b>	55.4	<b>26.9</b>	<b>19.8</b>	30.8

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.

### D.3. 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 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. Other than our random data augmentations during training, we do not explicitly crop the RGB images to the object or to the same field of view as the tactile sensor, since the environmental context of the object may be important for its representation. For example, the contact conditions can influence the impact audio of an object, and this may be captured from an image which captures the context around the object.

**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 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 pretrained

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. Multi-Sensory Embedding Space Arithmetic

While humans can gather important information about an object through any single sensory modality, each additional modality can provide complementary insights. Similar to ImageBind [19], we use our aligned data to explore how to combine embeddings from different modalities to better capture the full richness of an object through emergent compositionality. We experiment with some simple methods to combine embeddings from multiple sensory modalities to leverage the complementary insights from each modality.



RGB-Audio-Tactile-Pointcloud Top-1 Accuracy (%)	RGB→			Audio→			Tactile→			Point→			Average
	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	18.0	15.6	15.5	16.6	16.7	16.7	17.2	16.2	16.4	13.9	16.8	15.1	16.2
Fine-Tune w/ Image Loss	<b>20.1</b>	<b>24.1</b>	<b>36.1</b>	19.8	17.4	20.0	<b>24.1</b>	17.8	<b>26.7</b>	<b>46.4</b>	19.5	24.1	24.7
Fine-Tune w/ Cross-Sensory Loss	19.7	22.0	<b>36.1</b>	<b>22.1</b>	<b>18.8</b>	<b>21.5</b>	23.6	<b>19.8</b>	26.4	44.4	<b>24.3</b>	<b>25.5</b>	<b>25.4</b>

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.

ImageBind Training Method	Op.	Target Modality			Avg.
		RGB	Audio	Depth	
Random Guess		5.0	5.0	5.0	5.0
Out-of-the-Box Pretrained	Mean	8.6	6.6	8.0	7.7
Out-of-the-Box Pretrained	Sum	7.8	5.2	10.6	7.9
Fine-Tune w/ Image Loss	Mean	78.0	40.6	58.2	58.9
Fine-Tune w/ Image Loss	Sum	79.0	<b>42.6</b>	<b>60.6</b>	<b>60.7</b>
Fine-Tune w/ Cross-Sens. Loss	Mean	76.8	41.8	56.6	58.4
Fine-Tune w/ Cross-Sens. Loss	Sum	<b>80.4</b>	41.8	58.4	60.2

Table 9. Cross-sensory retrieval top-5 accuracies (%) of combining two sensory modalities to retrieve the third with fine-tuned ImageBind encoders, comparing different training methods and embedding combination operations (“Op.”). The Target Modality columns are labeled by the *target* sensory modality, with the other two modalities combined and used as the query.

ImageBind Training Method	Op.	Target Modality			Avg.
		RGB	Audio	Depth	
Random Guess		16.7	16.7	16.7	16.7
Out-of-the-Box Pretrained	Mean	18.2	17.8	21.2	19.1
Out-of-the-Box Pretrained	Sum	18.2	17.8	21.2	19.1
Fine-Tune w/ Image Loss	Mean	42.5	21.0	32.7	32.1
Fine-Tune w/ Image Loss	Sum	42.5	21.0	32.7	32.1
Fine-Tune w/ Cross-Sens. Loss	Mean	<b>42.8</b>	<b>23.8</b>	<b>31.8</b>	<b>32.8</b>
Fine-Tune w/ Cross-Sens. Loss	Sum	<b>42.8</b>	<b>23.8</b>	<b>31.8</b>	<b>32.8</b>

Table 10. Contact localization top-1 accuracies (%) of combining two sensory modalities to retrieve the third with fine-tuned ImageBind encoders, comparing different training methods and embedding combination operations (“Op.”). The Target Modality columns are labeled by the *target* sensory modality, with the other two modalities combined and used as the query.

Results from ImageBind suggest that a simple arithmetic combination of multiple embeddings can be quite effective, specifically using an additive sum. In addition to an additive sum, we also try using a simple mean of two embeddings for retrieving the third. We show results from using these methods on the embeddings from the fine-tuned ImageBind encoders from Sections 5.2 and 5.3 to perform the same cross-sensory retrieval and contact point localization tasks, respectively. We show results for the cross-sensory retrieval task in Table 9. Interestingly, the encoders fine-tuned with the Image Loss marginally outperform those trained with

the Cross-Sensory Loss on most modalities. Using simple additive sum for the combination operation also tends to outperform using the mean. For contact point localization (Table 10), the Cross-Sensory Loss marginally outperforms the Image Loss. Note that mean and sum produce the same results, regardless of training method. In both the retrieval and localization tasks, we are using a cosine distance for measuring similarity, but in the localization task, there are much fewer embeddings from which to select (6 vs. 100), and all embeddings are from the same object and therefore may be more similar in scale.

However, one premise of combining modalities is that we may be able to perform such cross-sensory retrieval and localization tasks better than using only one modality. Note that our results mostly fail to demonstrate this. When comparing results between using a single modality and two modalities for retrieval in Tables 3 and 9, respectively, using Audio+Depth to retrieve RGB is the only combination that outperforms using the maximally-performing modality of a pairing by itself. Audio may provide especially complementary information to depth, revealing properties important for visual discrimination, such as materials. We see a similar trend for contact point localization by comparing Tables 4 and 10, though using Depth+RGB to retrieve Audio also marginally outperforms using either modality by itself.

For generation, we combine embeddings using a simple mean, to prioritize keeping the magnitude of the combination within the same scale as the original CLIP embeddings that each generation method was designed to consume. We show results in Figure 11. For comparison, we use the same inputs we used in Figure 5. Similar to retrieval, the only combination for which the generations occasionally seem better than those of the best single modality is the Audio+Tactile combination. It seems that combining another modality with the Image modality may push the combined embedding outside the distribution of CLIP embeddings, confusing the generation network.

Our results on each of these tasks highlight the potential for devising more principled approaches to combining cross-sensory representations in order to fully exploit the complementary information each modality provides.

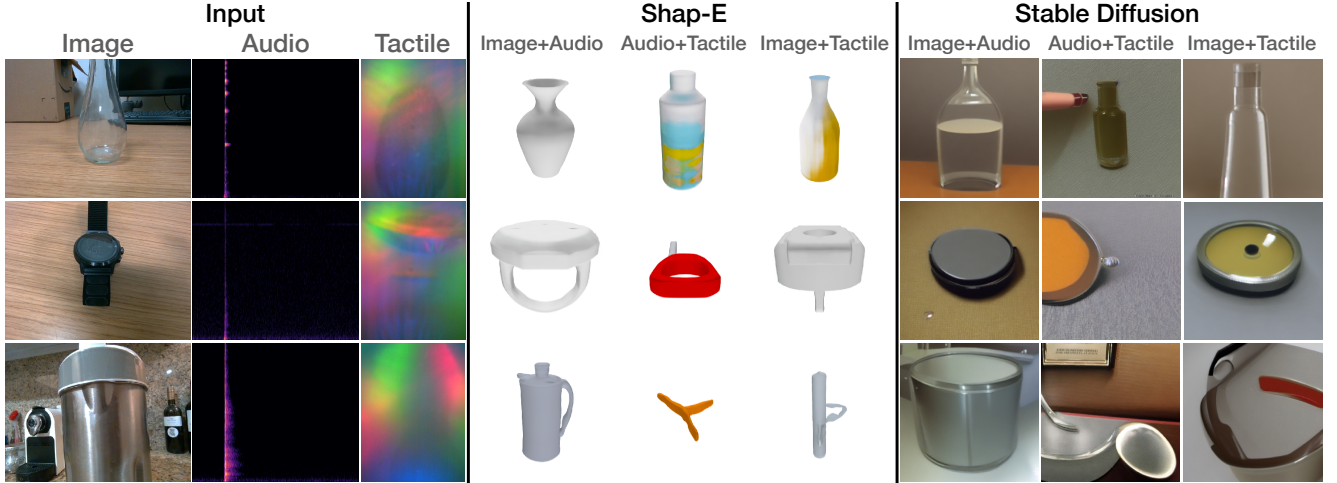


Figure 11. Results of prompting Shap-E [27] to generate 3D neural radiance fields and Stable Diffusion [38] to generate images, using different pairings of modalities’ encodings from our multimodal encoders. We take the average across the embeddings of each modality in the pairing to use as input for each generation method. 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 combining images and audio, audio and tactile, and image and tactile embeddings, respectively, as input to Shap-E. The last three columns similarly show the images generated from the same combinations used as input to Stable Diffusion.



Figure 12. 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.

#### E.4. Zero-Shot Audio-Based Object Detection

We show additional results from the experiment described in Section 5.7 in Figure 12.

#### E.5. 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 retrieval and contact localization 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 100 epochs on each dataset. For all 2D and 3D generation experiments, we train for 1,000 epochs. 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.