

# CAVER: Curious Audiovisual Exploring Robot

Luca Maccesanu\*, Boueny Folefack\*, Samik Singh†, Ruchira Ray†,  
Ben Abbatematteo, Roberto Martín-Martín

**Abstract**—Multimodal audiovisual perception can enable new avenues for robotic manipulation, from better material classification to the imitation of demonstrations for which only audio signals are available (e.g., playing a tune by ear). However, to unlock such multimodal potential, robots need to learn the correlations between an object’s visual appearance and the sound it generates when they interact with it. Such an active sensorimotor experience requires new interaction capabilities, representations, and exploration methods to guide the robot in efficiently building increasingly rich audiovisual knowledge. In this work, we present CAVER, a novel robot that builds and utilizes rich audiovisual representations of objects. CAVER includes three novel contributions: 1) a novel 3D-printed end-effector, attachable to parallel grippers, that excites objects’ audio responses, 2) an audiovisual representation that combines local and global appearance information with sound features, and 3) an exploration algorithm that uses and builds the audiovisual representation in a curiosity-driven manner that prioritizes interacting with high uncertainty objects to obtain good coverage of surprising audio with fewer interactions. We demonstrate that CAVER builds rich representations in different scenarios more efficiently than several exploration baselines, and that the learned audiovisual representation leads to significant improvements in material classification and the imitation of audio-only human demonstrations. More information: <https://robin-lab.cs.utexas.edu/CAVER>

## I. INTRODUCTION

Humans learn and exploit multimodal audiovisual cues in everyday life to obtain a more complete understanding of their environment and broader manipulation capabilities. We routinely fuse audio and vision to understand materials and reproduce behaviors: tapping a mug reveals glass vs. ceramic, and hearing a melody lets a musician find the right key. Building similar capabilities in robots would increase their robustness and autonomy, but requires a representation that couples *how things look* with *how they sound when interacted with*, and a way to acquire that representation efficiently through interaction.

Underlying the multimodal audiovisual abilities of humans is a rich developmental process in which children interact with and explore the physical world around them [1]. When children play with objects, they generate multimodal stimuli from which they learn audiovisual correlations in a self-supervised manner [2]. Their exploration is incremental and, critically, *curious*: children prioritize unfamiliar experiences so as to quickly and efficiently learn the audiovisual properties of the world [3], [4]. While prior work in machine learning and robotics has explored how to build such audiovisual representations [5]–[8], they rely on previously acquired

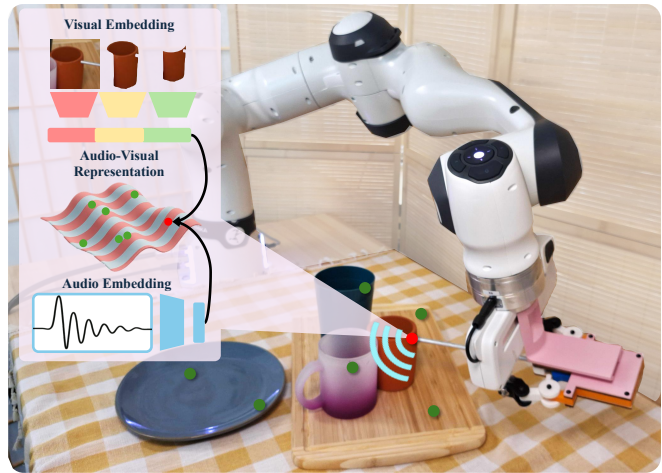


Fig. 1: Curiously Building and Exploiting an Audiovisual Representation with CAVER. CAVER incrementally builds a KNN-based audiovisual representation capable of audio-to-visual and visual-to-audio prediction. To explore an environment, CAVER considers many candidate interaction points (dots), and ranks their uncertainty by comparing their visual features to those of prior samples. CAVER then collects an audio sample of the most uncertain point (red) using our novel impact tool. The audio and visual features are added as a pair to the audiovisual representation in an intrinsically motivated process that results in efficient interactive exploration.

large datasets of visual and audio pairs collected manually, largely failing to account for how a robotic agent might explore an environment autonomously and efficiently.

In this work, we introduce CAVER, Curious Audiovisual Exploring Robot, a robot that autonomously builds and exploits an audiovisual representation of objects (Fig. 1). CAVER integrates a novel, custom-built impact tool that enables a robot to “tap” on object surfaces to elicit an audio response. CAVER draws inspiration from human curiosity to guide the robot through interaction with uncertain objects, first actively exploring visually unfamiliar objects to learn correlations between their appearance and impact acoustics, then using those correlations for downstream tasks. Concretely, a strategy of uncertainty-guided sampling chooses where to strike next to generate a visual [9], [10] and auditory [11] feature representation of the interaction point, while a simple but effective k-nearest-neighbor (KNN) methodology on the audiovisual features enables immediate retrieval in either direction (vision→audio and audio→vision) to exploit the information for manipulation (see Sec. IV-D).

We test CAVER by exploring various household scenes and incrementally building its audiovisual representation, and evaluate it on downstream tasks including audio pre-

\* , † denotes equal contribution. All authors are with The University of Texas at Austin

diction, material classification, and audio-based imitation. Across several household environments, CAVER enables rapid learning of object visual and acoustic properties in an autonomous and self-supervised manner. Through our experiments, we demonstrate that CAVER’s curious, uncertainty-driven exploration strategy yields a marked improvement in audio prediction quality over several sampling and exploration alternatives, and its audiovisual representation achieves 87% accuracy in material classification, 66% accuracy in audio-based imitation of melodies, and 42% accuracy in action recognition (compared to a 27% accurate human baseline).

In summary, in CAVER our contributions include: 1) a reproducible, 3D-printed *impact end-effector*; 2) a *multi-scale audiovisual representation* combining visual and audio features in a KNN model, and learned feature weights aligning visual to audio distance; 3) an *uncertainty-guided exploration* policy that prioritizes visually distinct object parts using farthest-first selection; and 4) a demonstration that this representation is immediately *useful* across tasks—audio prediction, material classification, musical imitation, and audio-only manipulation inference—without large-scale pretraining on paired data.

## II. RELATED WORK

CAVER is a robot that can perceive, curiously explore, and quickly learn an audiovisual representation that can predict sound based on images and be applied to solve audio-related downstream tasks. In this section, we contrast it to representative prior work that explored multimodal robot learning, interactive perception, and action sound generation.

**Multimodal and Audiovisual Robot Learning.** Leveraging multi-modal data in robotic tasks is less explored than the conventional vision-based learning setting [12]–[15]. While several methods have combined vision with tactile information [16], [17] or force feedback [18], [19] for increased performance, leveraging audiovisual signals remains relatively underexplored. Only a few methods have explored leveraging audio in imitation learning [20], to handle occlusions [21], or as supervision for skill learning [22]. This primarily stems from a lack of large annotated audio datasets, as well as the inherent complexity of working with audio, including source ambiguity, environmental variability, and temporal dynamics. The ObjectFolder benchmark [8], [23] provides both synthetic and real impact sounds collected manually from a wide variety of household objects. DiffImpact [7] explores differentiable rendering approaches to generating impact sounds, decomposing sounds into force profiles, and object acoustic properties. Still, none of these methods answer the question of how we can enable a robot to autonomously gather data to learn about the acoustic properties of the world around it. CAVER is able to autonomously generate its own dataset of audiovisual samples with which to learn a representation and solve downstream tasks.

**Interactive Perception.** CAVER is a robot that learns to perceive audiovisual information by leveraging interactions with the environment in the form of hitting actions that reveal

audio signals. This is a form of *interactive perception* [24], a family of techniques that enrich robots’ perceptual capabilities by leveraging physical interactions. Examples include interactively segmenting a scene [25]–[27], perceiving kinematics [28], [29], and recognizing objects [30], [31]. Some interactive perception works [32]–[34] demonstrate material property inference by interactively generating audio data. However, none of these methods provides an account of curiosity-driven exploration of the audiovisual properties of a scene or a sufficiently general representation that can be applied to diverse downstream tasks.

**Curious Exploration.** Curiosity as an intrinsic metric to guide exploration has proven its utility in reinforcement learning domains [35]–[37]. Prior works have been able to build accurate dynamics representations of their environment by exploring state action pairs whose effects are unknown [36]. CAVER extends this approach to the audiovisual domain by exploring uncertain regions of the environment to learn their audio properties.

**Action Sound Generation.** Previous approaches have attempted to predict the audio that an interaction would create based on the visual appearance of the interacted area. Several methods use physics simulation to create virtual audio of interactions [38]–[40]. These methods require detailed models of the objects (e.g., shape, material), which are costly to obtain in unstructured environments. Owens et al. [5] introduced the idea of visually indicated sounds, audio signals that can be predicted based on the visual appearance of the interacted objects. Since then, several works have explored generating audio data conditioned on video, typically using large, manually curated datasets of humans interacting with objects [6], [41]–[46]. Despite the ability of these methods to generate plausible sounds accompanying videos, they fail to address the active exploration problem required to guide an agent to autonomously acquire its own audiovisual representation of the environment in the real world. Moreover, many of these methods typically predict audio based on observed video input (sequence of images), whereas CAVER learns to predict impact sounds only from a single image.

## III. CAVER: CURIOUS AUDIOVISUAL EXPLORING ROBOT

CAVER explores the visual and acoustic properties of objects in its surroundings in an efficient, self-supervised manner so as to build a rich audiovisual representation that enables downstream tasks (see Fig. 2). Equipped with a novel, custom-built end-effector tool that enables the robot to hit different interaction points gently, CAVER takes a curiosity-driven approach that encourages exploration in the areas that promise to generate the most novel audiovisual signals. Our assumption is that visually distinct object regions are likely to generate different sounds, and CAVER leverages this insight to identify object parts with high visual uncertainty relative to the current dataset, providing fast coverage of diverse audiovisual signals.

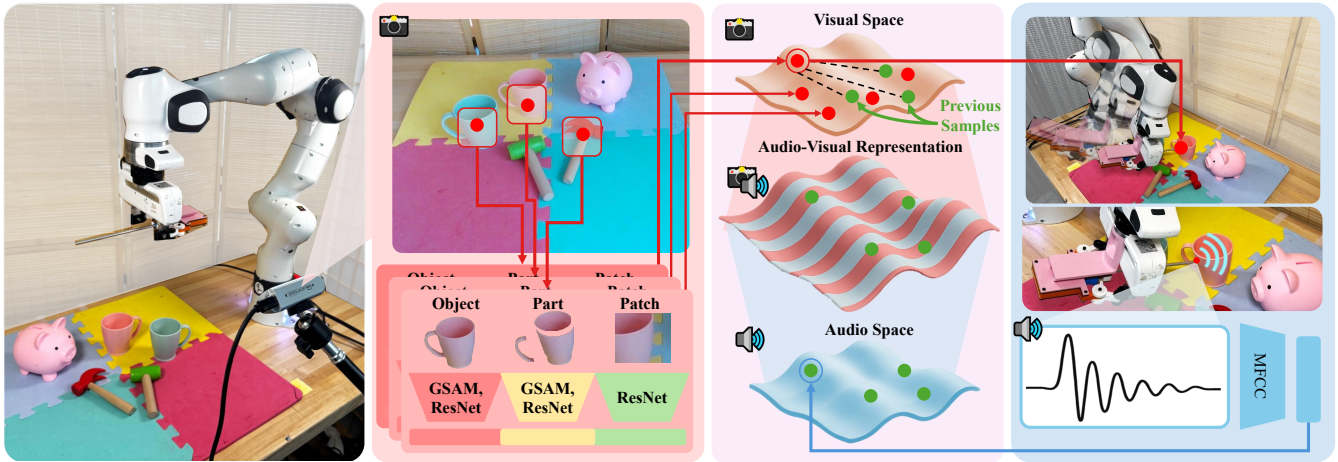


Fig. 2: **Overview of CAVER’s audiovisual representation and curious exploration.** CAVER curiously and efficiently learns correlations between object visual appearance and acoustic properties. Given candidate hitting points, CAVER uses a KNN model with fine tuned features from foundation vision models to predict corresponding impact sounds. To select informative interaction points, CAVER selects the most uncertain candidate hitting location using distance in visual feature space between a candidate (red) and all prior samples (green) as a proxy for uncertainty. After sampling the most uncertain candidate, the corresponding visual and audio embeddings are paired as two sides of the Audio-Visual representation. This can best be thought of as a bi-directional mapping where audio features can be used to predict visual features, visual features can be used to predict audio features, and concatenating the embeddings gives an informative multimodal representation of that sample point.

At its core, CAVER relies on a KNN mapping visual embeddings to audio samples. The dependence of audio on environmental factors indicates that in domain data is likely important for accurate audio prediction. As such, rather than trying to train a generative model on a large dataset, CAVER instead aims to learn ground truth per object sounds in the relevant environment. In this way CAVER’s retrieval-based approach leverages the simplicity of KNNs as a tool for maximizing coverage of the audio domain in the context of its environment.

To maximize the information gained with a new audiovisual sample, avoid repeatedly interacting with previously explored objects, and infer the most promising interaction points for downstream tasks, CAVER analyzes candidate samples on both the global and local level. CAVER evaluates the differences between the visual appearance of the candidate interaction points and previously interacted ones and selects the most uncertain one (the most distinct one), in a curiosity-inspired mechanism. Conversely, given an audio sample, CAVER can also predict the most promising interaction point to recreate that sound by inverting the similarity-based search process. Both capabilities (uncertainty-based curious exploration and similarity-based exploitation) are necessary to 1) efficiently build a proper audiovisual representation and 2) achieve downstream tasks such as audiovisual material classification or reproducing a tune or an audio-demonstrated manipulation. In the following, we describe each of the components of CAVER in detail.

### A. Impact Tool

To generate consistent, informative impact sounds across diverse objects, we equip the robot with a mechanically actuated, single degree of freedom, 3D-printed, spring-loaded impact tool that can be grasped/mounted on a standard

parallel gripper (Fig. 3). Closing the gripper actuates a cam-follower that preloads the spring and releases a short, repeatable strike with a metal rod; a directional microphone mounted near the scene camera records the impact. The force of the impact is dictated by the thickness of the 3D-printed spring, which is set at 2mm after qualitative adjustments for consistent sample collection. This arrangement both standardizes excitation energy across trials and allows safe, gentle contacts at surface normals computed from the depth point cloud, so interaction is precise yet does not cause motion. Our data collection indicated that even after 1800 samples, the performance of this hitting implement did not measurably deteriorate. The result is audio that is comparable across objects and scenes, enabling a simple KNN model to perform surprisingly well in the low-data regime targeted by CAVER. Thanks to our custom tool, CAVER turns each interaction into a clean audiovisual training sample, powering both exploration and the downstream results in Sec. IV-D.

### B. CAVER Audiovisual Representation

At its core, CAVER autonomously builds and maintains an increasing audiovisual representation of the interacted points that guides exploration and allows exploitation of the information for downstream tasks. CAVER’s audiovisual representation is a growing set  $\mathcal{S}$  of tuples  $\{(x, \mathbf{f}(x), \mathbf{a}(x))\}$ , where  $x$  is an interaction point on an object,  $\mathbf{f}(x)$  is a multi-scale *visual* embedding of that point and its context, and  $\mathbf{a}(x)$  are *audio* descriptors from the recorded impact. The representation is intentionally retrieval-friendly so it can be queried during exploration and downstream tasks.

Given a segmented scene, CAVER extracts visual features that preserve semantics and geometry at complementary scales. We use GroundedSAM (GSAM) [9], [47], [48] to obtain mask embeddings for the *object* and *object part*, and

a pre-trained ResNet to compute appearance features for the *object crop*, *part crop*, and a small *local patch* centered at  $x$ . We concatenate these into a single visual feature of the form:

$$\mathbf{f}(x) = [\mathbf{f}_{\text{obj}}^{\text{SAM}}, \mathbf{f}_{\text{part}}^{\text{SAM}}, \mathbf{f}_{\text{obj}}^{\text{RN}}, \mathbf{f}_{\text{part}}^{\text{RN}}, \mathbf{f}_{\text{patch}}^{\text{RN}}] \quad (1)$$

After a point is interacted with, CAVER acquires and creates an audio feature from the resulting impact waveform. First, CAVER level-normalizes and clips the audio signal around the strike; then, it computes mel-frequency cepstral coefficients (MFCCs) and stores them as audio features. MFCC features represent the shape of the spectral envelope of a sound, describing timbre instead of the intensity of the impact waveform. This makes them effective for identifying contact points independently of the strength of the interaction. To evaluate and match audio to vision for downstream tasks (see Sec. IV-D), CAVER uses mel-cepstral distortion (MCD) [11] between MFCC features as an audio-space distance metric between samples.

*Aligning Audio and Visual Distances:* Distances in the concatenated visual space may not reflect acoustic similarity. We therefore learn per-component weights so that the weighted *visual* distance correlates strongly with the *audio* distance. Given two candidate points  $x_i, x_j$ , we define

$$d_v(x_i, x_j) = \sum_k w_k \|\mathbf{f}_k(x_i) - \mathbf{f}_k(x_j)\|_2, \quad (2)$$

where  $k \in \{\text{SAM-obj}, \text{SAM-part}, \text{RN-obj}, \text{RN-part}, \text{RN-patch}\}$  indexes feature types. We fit the scalar weights  $\{w_k\}$  by linear regression so that  $d_v$  best predicts the corresponding MCD between paired impact audios. To compute these weights before online exploration, we used synthetic data from the ObjectFolder [8] dataset. This alignment lets the system i) better retrieve plausible sounds from vision and most promising interaction points to recreate given audio by nearest neighbors, and ii) more accurately measure novelty for exploration directly in the visual space.

### C. Uncertainty-guided Exploration

CAVER uses uncertainty to decide where to interact next so that each strike adds the most information. At each iteration, first generate per-object and per-part masks by prompting GroundedSAM [9] with a ground-truth list of objects in the scene. The object masks are then projected onto the scene point cloud, and their novelty is computed with respect to the current audiovisual representation  $\mathcal{S}$ :

$$U(x) = \min_{s \in \mathcal{S}} d_v(x, s). \quad (3)$$

The policy selects the farthest candidate  $x^* = \arg \max_x U(x)$ , subject to a simple cycling rule over detected objects so that each instance is sampled before revisiting. Intuitively, this expands the frontier of audiovisual knowledge in the most visually unfamiliar regions first, while still allocating budget fairly across objects. In practice, combining cycling with curiosity delivers faster MCD drop than baselines (see Sec. IV-D-A, and Fig. 4), and it seeds a more discriminative embedding that boosts downstream classification (see Sec. IV-D-B, and Fig. 5).

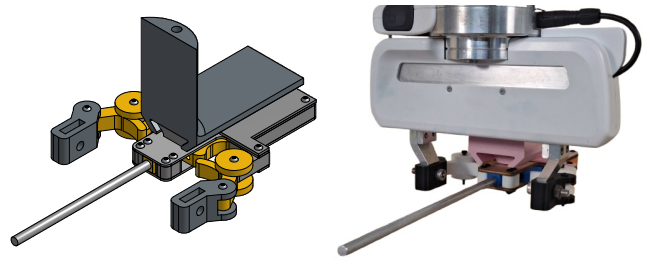


Fig. 3: **Robotic Impact Tool.** To generate consistent impact sounds, we designed a 3D-printed, spring-loaded impact tool that attaches to the robot’s gripper. When the gripper closes, a cam-follower retracts the rod, building tension in the spring until the cam slips and the stored energy in the spring drives the metal rod forward, impacting the object. A directional microphone is mounted near the scene camera to record the resulting sound at 44.1 kHz.

### D. Exploiting the Representation for Downstream Tasks

CAVER efficiently learns audiovisual correlations and gathers audiovisual data of its environment in a self-supervised, autonomous manner. Given this data, multiple downstream tasks are possible, leveraging the same nearest-neighbor-based retrieval machinery. Here, we study three in particular: audio generation, material classification, and audio-based imitation learning.

**(i) Efficient exploration towards best audio prediction from vision.** Given a prospective hit, CAVER uses a nearest-neighbor retrieval under  $d_v$  to obtain the audio exemplar that most closely matches the expected acoustics for that location. While this model yields modest audio generation capabilities compared to generative AI models, it proves effective and accurate in our low-data regime. This retrieval methodology reveals the benefits of CAVER’s exploration strategy, which decreases MCD the fastest for a test set of held interaction points (Fig. 4).

**(ii) Material classification.** To evaluate this task, we train lightweight material prediction heads on visual, audio, or concatenated audiovisual embeddings, trained on a subset of interaction points data labeled manually with ground truth material labels. We observe that while CAVER’s visual features collected with the curiosity-driven mechanism already provide strong predictive capabilities (as in humans) that are enhanced when audio is integrated in the audiovisual variant, yielding strong and sample-efficient accuracy (up to 87% in our scenes; Fig. 5).

**(iii) Audio-guided imitation.** CAVER enables a robot to reproduce audio demonstrations by inverting the nearest neighbor computation: given a candidate audio sample, we can calculate the most similar audio in the dataset by finding the sample in the dataset with the lowest MCD to the query. Given this audio sample, we can retrieve the corresponding visually closest sample and identify the object part that most likely produced the sound. In our experiments, we perform two such imitation tasks: reproducing musical notes, and identifying objects involved in a pick-and-place manipulation performed by a human, closing the loop from *listen* to *do* (see Sec. IV-C and Sec. IV-D).

## IV. EXPERIMENTAL EVALUATION

We evaluate CAVER exploration-exploitation capabilities on several downstream tasks that require learning and exploiting correlations between audio and visual appearance. We study four tasks in particular: predicting audio given visual input and hitting location, predicting material based on audio and vision, inferring musical notes played by a human, and identifying objects in a pick-and-place manipulation demonstrated by a human. Our evaluation is designed to test CAVER’s capabilities to efficiently explore the environment to build a rich and useful audiovisual representation and the accuracy of exploiting it for the four downstream tasks.

Using a fixed RGB-D camera, CAVER converts point-clouds into surface normal maps to help an exploration policy select and orient target impact poses for a Franka Research 3 robot. The system then executes collision-free paths via RRT [49] and Octomaps [50], utilizing audio magnitude thresholds to automatically filter failed impacts before autonomously proceeding to the next sample. To process audio, each impact recording is aligned by peak magnitude and clipped to a fixed window of 10 ms before and 100 ms after the peak. We apply per-clip amplitude normalization and compute features for MCD by resampling to 22.05 kHz and extracting the spectral envelopes (frame period 5.0 ms, FFT size 512). We convert these envelopes to 13th-order mel-cepstral coefficients (MCEP;  $\alpha=0.65$ ) and flatten the resulting coefficient sequence into a fixed-length vector per interaction. We measure end-to-end sampling time from timestamps between consecutively collected samples. This includes time spent on failed attempts and any resulting replanning/resampling. The average time per sample is 52.9 s across all environments ( $\approx 68$  samples/hour):  $53.7 \pm 3.7$  s (Garage),  $56.2 \pm 8.5$  s (Kitchen), and  $48.6 \pm 3.1$  s (Playroom). Our experiments aim to answer the following four questions:

*A. Can CAVER enable a robot to efficiently gather data for training an audio prediction model?*

To assess CAVER’s exploration and audio generation capabilities, we designed an experiment in which the agent interacts with various household environments, sampling scenes to collect data autonomously. We designed three environments (`kitchen`, `garage`, and `playroom`), each consisting of five scenes with representative objects arranged on a table. Objects from each of the environments vary in size, color, texture, and materials, and are depicted in Fig. 4. The scenes are designed to test various aspects of the robotic system, including its ability to handle cluttered scenes, objects of similar visual appearance but different materials, and recurring objects across scenes. The scenes are experienced sequentially in order to test the model’s ability to prioritize novel objects. To evaluate performance, we manually curate a test set of impact sounds and corresponding hitting locations by sampling each part of each object that occurs in the scenes. To measure audio generation accuracy, we compute the mel-cepstral distortion (MCD) between the retrieved audio sample and the ground truth audio sample.

In order to validate the effectiveness of our curiosity metric, we ablate CAVER’s point selection against the following baselines:

- **Random:** Randomly sample valid hitting locations by choosing points in the point cloud uniformly at random.
- **Cycling:** Sample hitting locations by cycling through objects and choosing a random point on each object. This ablation of CAVER proceeds in an arbitrary order rather than prioritizing unfamiliar objects/points.

Fig. 4 shows the results for each household environment. Dashed vertical lines indicate the boundaries between scenes when the objects before the robot change. The test samples for each scene are added to the test set at the start of that scene. Audio prediction error naturally declines over time as the robot encounters more objects that are present in the withheld test set. We see that CAVER more quickly reduces the audio prediction error than the baselines, indicating that the curiosity-driven exploration indeed leads to more efficient learning of the audiovisual properties of the objects in each scene. This is especially apparent in later scenes when many of the objects have already appeared in earlier scenes and are thus already familiar to CAVER, whereas the baselines spend additional time sampling objects for which the uncertainty is already low.

We also observe that CAVER struggles to retrieve quality audio samples when dealing with visually identical objects that have significantly different sound profiles (namely plastic white mug versus a ceramic white mug). This occurs when the same region in visual space maps to two separate regions of audio space, creating ambiguity as to what the object should sound like. This problem is inherent to audiovisual approaches, and instance-level object identification would be required to definitively solve it. In IV-D we explore a workaround where objects can be sampled during tasks to make instance-level predictions.

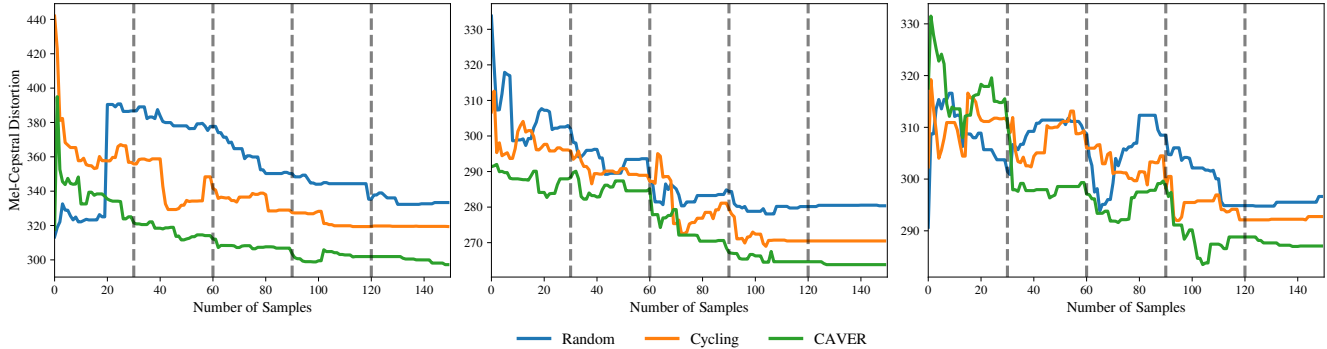
*B. Can CAVER enable efficient learning of object material prediction given audio data?*

In a second experiment, we evaluate CAVER’s embeddings in a material classification task. The setup is identical to Experiment 1 above; however, we evaluate material classification performance rather than audio generation error. Ground truth material labels for each sample are provided by a human annotator. We compare three variants of MLP-based material classification models: `vision-only` uses only CAVER’s visual embeddings; `audio-only` uses audio—featurized using MFCC, Chroma, and Mel Spectrogram [51]—to predict the material; and `audiovisual` uses the concatenated latent features of both modalities during its forward pass. The visual features are encoded using MLPs and audio features with a CNN. The complete list of material classes consists of ceramic, glass, metal, plastic, rubber, and wood.

To evaluate the effectiveness of CAVER’s visual embeddings and the importance of audiovisual data in material classification, we first evaluate material classification accuracy on data from CAVER’s exploration described above

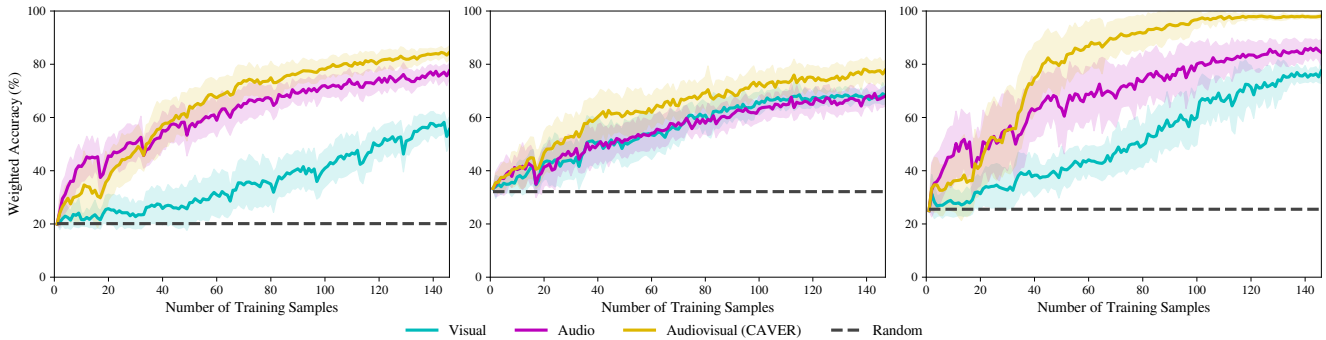


(a) Our experiments take place across 3 environments: kitchen, garage, and playroom.



(b) Audio prediction error as a function of samples collected for different audiovisual exploration approaches for the kitchen, garage, and playroom.

**Fig. 4: Curious Exploration for Audio Prediction.** The top row (a) visualizes the objects that comprise each environment. The bottom row (b) shows the average audio prediction error on a held-out test set, measured as the mel-cepstral distortion, as a function of the number of interaction points sampled by the robot. Vertical lines indicate scenes within an environment in which a subset of the objects appear. Each environment is plotted individually (left to right: kitchen, garage, and playroom). CAVER consistently achieves higher prediction accuracy more quickly than the naive exploration baselines via its curious exploration and audiovisual representation.



**Fig. 5: Material Classification Results.** We measure the efficacy of CAVER’s unified audiovisual embeddings for material classification compared to visual embeddings alone, audio embeddings alone, and a random baseline. The plots show the class-balanced material classification accuracy on a held-out test set as a function of the number of interaction points sampled by the robot, aggregated over 20 runs. Performance is evaluated in the kitchen, garage, and playroom environments, respectively. Incorporating both audio and visual inputs as in CAVER’s unified embedding is necessary to achieve strong material classification performance given ambiguous stimuli.

in Sec. IV-A. The order in which the data is provided to the model is randomized across 20 runs for robustness. This procedure is repeated across each environment. Fig. 5 showcases the classification performance results for each model as a function of samples collected. We observe that the audiovisual model has the strongest performance across all three environments, demonstrating the utility of both combined audio and visual features when classifying materials. In addition, audio-only outperforms vision-only in environments where there are many objects of the same type, demonstrating the utility of the curiously collected data.

### C. Can CAVER’s audiovisual representation enable an agent to reproduce musical sounds?

Human expert musicians are capable of reproducing a tune on their instrument upon hearing it once. In this experiment, we evaluate CAVER’s ability on this same imitation task. We designed two tasks to evaluate this capability: playing the drums, and playing a xylophone. The instruments are pictured in Fig. 6. For each instrument, a test set of sounds is generated by a human demonstrator. We evaluate CAVER’s ability to infer the correct musical notes and compare its performance to a human baseline. The human is allowed to play



(a) CAVER replicating a song on the drums and xylophone.



(b) CAVER imitating previously unheard object interactions.

**Fig. 6: Audio Imitation with CAVER.** CAVER’s audiovisual representation enables imitation of solely auditory demonstrations. In (a), CAVER plays a song by ear after exploring the instruments autonomously. In (b), the robot replicates unseen manipulation actions by predicting what objects sound like together (ie, combining the audio from a wood cup sample and a ceramic plate sample) and comparing them to the observed sound. The ability to reason about multi-object interaction sounds gives CAVER the flexibility to be applied to diverse downstream tasks.

with each instrument for a few minutes before being played test sounds and asked to guess the correct note. CAVER first explores the scene to build its KNN model. To reproduce a given sound, we sample points in the environment and predict their corresponding audio. The predicted audio with the lowest MCD distance to the test sound is selected as CAVER’s prediction. CAVER is able to infer the correct xylophone key 64% of the time, and the correct drum hit with 68% accuracy. These results demonstrate the effectiveness of CAVER’s visual representation and sample retrieval. Please see the supplement for a demonstration of CAVER’s musical ability.

#### D. Can CAVER enable activity recognition by inferring objects that are interacted with based on sound alone?

A final capability enabled by CAVER is imitation of a manipulation behavior based on sound alone: given a test sound, CAVER enables the robot to infer which object(s) were involved in the production of this sound. We evaluate this capability on a pick-and-place task. A human picks an object of varying material and places it on a plate of known material. The robot’s task is to infer which object was placed and reproduce the demonstrated behavior. CAVER interacts with the picked objects and placement objects for a period of time before the evaluation begins. We assume that the placement object is known and only the picked object needs to be determined. To estimate which objects were involved in the manipulation, CAVER synthesizes a combined sound by superimposing the predicted sounds of the picked and placement objects. As above, the interaction point yielding the lowest MCD to the test audio is selected and that object is picked. We find that CAVER is able to correctly infer the manipulated object 42% of the time, whereas the average human performance is only 27%.

#### V. CONCLUSION AND LIMITATIONS

We introduced CAVER, a robot capable of efficiently and autonomously learning the audiovisual properties of objects. By combining a nearest-neighbor retrieval system with

uncertainty-driven exploration, CAVER can generate impact sounds conditioned on visual input and accelerate the acquisition of acoustic properties compared to naive sampling strategies. This enables robots to rapidly acquire auditory knowledge that supports downstream tasks such as sound prediction, material classification, and activity recognition, as demonstrated in our comprehensive evaluation. While CAVER performs well in all our experiments, it presents several limitations that we plan to explore: 1) CAVER does not yet capture continuous sounds (e.g., dragging), 2) it completely separates exploration from inference, 3) it lacks mechanisms to filter background noise that could distort the learned audiovisual representation 4) failure states (unsamplable points, mechanical failure during sampling, etc.) are discarded and 5) CAVER assumes a ground truth list of objects for visual segmentation. Addressing these challenges will enable larger-scale deployments in more open-ended environments, such as homes with diverse objects and ambient conditions, where the benefits of uncertainty-guided exploration are especially critical. We also hope further exploration of the audio-visual field will yield appropriate baselines against which to benchmark CAVER’s performance. Overall, we view CAVER as a step toward equipping robots with richer multimodal understanding, bridging vision and audition to support new capabilities and more adaptive, perceptive, and interactive behavior in the real world.

#### VI. ACKNOWLEDGEMENTS

We would like to thank Dr. David Harwath for his guidance in audio representation and analysis techniques. This project has been partially funded by Good Systems, an initiative by the University of Texas at Austin.

#### REFERENCES

- [1] J. Piaget *et al.*, *The origins of intelligence in children*. International universities press New York, 1952, vol. 8, no. 5.
- [2] L. Smith *et al.*, “The development of embodied cognition: Six lessons from babies,” *Artificial life*, vol. 11, no. 1-2, 2005.
- [3] G. Loewenstein, “The psychology of curiosity: A review and reinterpretation,” *Psychological bulletin*, vol. 116, no. 1, p. 75, 1994.

- [4] J. Jirout *et al.*, “Children’s scientific curiosity: In search of an operational definition of an elusive concept,” *Developmental review*, vol. 32, no. 2, pp. 125–160, 2012.
- [5] A. Owens *et al.*, “Visually indicated sounds,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2405–2413.
- [6] C. Chen *et al.*, “Action2sound: Ambient-aware generation of action sounds from egocentric videos,” in *European Conference on Computer Vision*. Springer, 2024, pp. 277–295.
- [7] S. Clarke *et al.*, “Diffimpact: Differentiable rendering and identification of impact sounds,” in *Conference on Robot Learning*. PMLR, 2022, pp. 662–673.
- [8] R. Gao *et al.*, “The objectfolder benchmark: Multisensory learning with neural and real objects,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17 276–17 286.
- [9] T. Ren *et al.*, “Grounded sam: Assembling open-world models for diverse visual tasks,” 2024. [Online]. Available: <https://arxiv.org/abs/2401.14159>
- [10] K. He *et al.*, “Deep residual learning for image recognition,” 2015. [Online]. Available: <https://arxiv.org/abs/1512.03385>
- [11] R. Kubichek, “Mel-cestral distance measure for objective speech quality assessment,” in *Proceedings of IEEE pacific rim conference on communications computers and signal processing*, vol. 1. IEEE, 1993, pp. 125–128.
- [12] S. Levine *et al.*, “End-to-end training of deep visuomotor policies,” *Journal of Machine Learning Research*, vol. 17, no. 39, pp. 1–40, 2016.
- [13] C. Tang *et al.*, “Deep reinforcement learning for robotics: A survey of real-world successes,” *Annual Review of Control, Robotics, and Autonomous Systems*, 2024. [Online]. Available: <https://www.annualreviews.org/content/journals/10.1146/annurev-control-030323-022510>
- [14] A. O’Neill *et al.*, “Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0,” in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 6892–6903.
- [15] A. Khazatsky *et al.*, “Droid: A large-scale in-the-wild robot manipulation dataset,” *arXiv preprint arXiv:2403.12945*, 2024.
- [16] Y. Chebotar *et al.*, “Learning robot tactile sensing for object manipulation,” in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2014, pp. 3368–3375.
- [17] M. A. Lee *et al.*, “Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks,” in *2019 International conference on robotics and automation (ICRA)*. IEEE, 2019, pp. 8943–8950.
- [18] R. Martín-Martín *et al.*, “Cross-modal interpretation of multi-modal sensor streams in interactive perception based on coupled recursion,” in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 3289–3295.
- [19] W. Liu *et al.*, “Forcemimic: Force-centric imitation learning with force-motion capture system for contact-rich manipulation,” *arXiv preprint arXiv:2410.07554*, 2024.
- [20] Z. Liu *et al.*, “Maniwav: Learning robot manipulation from in-the-wild audio-visual data,” in *8th Annual Conference on Robot Learning*, 2024.
- [21] M. Du *et al.*, “Play it by ear: Learning skills amidst occlusion through audio-visual imitation learning,” *arXiv preprint arXiv:2205.14850*, 2022.
- [22] A. Thankaraj *et al.*, “That sounds right: Auditory self-supervision for dynamic robot manipulation,” in *Conference on Robot Learning*. PMLR, 2023, pp. 1036–1049.
- [23] R. Gao *et al.*, “Objectfolder 2.0: A multisensory object dataset for sim2real transfer,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 598–10 608.
- [24] J. Bohg *et al.*, “Interactive perception: Leveraging action in perception and perception in action,” *IEEE Transactions on Robotics*, vol. 33, no. 6, pp. 1273–1291, 2017.
- [25] A. Mishra *et al.*, “Active segmentation for robotics,” in *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2009, pp. 3133–3139.
- [26] J. Kenney *et al.*, “Interactive segmentation for manipulation in unstructured environments,” in *2009 IEEE International Conference on Robotics and Automation*. IEEE, 2009.
- [27] T. Kontogianni *et al.*, “Interactive object segmentation in 3d point clouds,” in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023.
- [28] R. M. Martin *et al.*, “Online interactive perception of articulated objects with multi-level recursive estimation based on task-specific priors,” in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2014, pp. 2494–2501.
- [29] B. Abbatematteo *et al.*, “Learning to generalize kinematic models to novel objects,” in *Proceedings of the 3rd Conference on Robot Learning*, 2019.
- [30] E. Maiettini *et al.*, “Interactive data collection for deep learning object detectors on humanoid robots,” in *2017 IEEE-RAS 17th international conference on humanoid robotics (Humanoids)*. IEEE, 2017, pp. 862–868.
- [31] E. Martínez-Martin *et al.*, “Object detection and recognition for assistive robots: Experimentation and implementation,” *IEEE Robotics & Automation Magazine*, vol. 24, no. 3, 2017.
- [32] J. Sinapov *et al.*, “Interactive object recognition using proprioceptive and auditory feedback,” *The International Journal of Robotics Research*, vol. 30, no. 10, pp. 1250–1262, 2011.
- [33] E. Torres-Jara *et al.*, “Tapping into touch,” *Fifth International Workshop on Epigenetic Robotics*, 2005.
- [34] Q. Nguyen *et al.*, “Autonomous sensorimotor learning for sound source localization by a humanoid robot,” in *IROS 2018-Workshop on Crossmodal Learning for Intelligent Robotics in conjunction with IEEE/RSJ IROS*, 2018.
- [35] L. Pinto *et al.*, “The curious robot: Learning visual representations via physical interactions,” 2016. [Online]. Available: <https://arxiv.org/abs/1604.01360>
- [36] S. A. Sontakke *et al.*, “Causal curiosity: RL agents discovering self-supervised experiments for causal representation learning,” in *Proceedings of the 38th International Conference on Machine Learning*, 2021.
- [37] D. Pathak *et al.*, “Curiosity-driven exploration by self-supervised prediction,” 2017. [Online]. Available: <https://arxiv.org/abs/1705.05363>
- [38] D. Jin *et al.*, “Differentiable room acoustic rendering with multi-view vision priors,” in *International Conference on Computer Vision (ICCV)*, 2025.
- [39] X. Liu *et al.*, “Hearing anywhere in any environment,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.
- [40] K. Su *et al.*, “Physics-driven diffusion models for impact sound synthesis from videos,” 2023. [Online]. Available: <https://arxiv.org/abs/2303.16897>
- [41] A. Owens *et al.*, “Audio-visual scene analysis with self-supervised multisensory features,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 631–648.
- [42] P. Chen *et al.*, “Generating visually aligned sound from videos,” *IEEE Transactions on Image Processing*, vol. 29, pp. 8292–8302, 2020.
- [43] V. Iashin *et al.*, “Taming visually guided sound generation,” *arXiv preprint arXiv:2110.08791*, 2021.
- [44] S. Luo *et al.*, “Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 48 855–48 876, 2023.
- [45] K. Su *et al.*, “Physics-driven diffusion models for impact sound synthesis from videos,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 9749–9759.
- [46] Y. Du *et al.*, “Conditional generation of audio from video via foley analogies,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2426–2436.
- [47] S. Liu *et al.*, “Grounding dino: Marrying dino with grounded pre-training for open-set object detection,” in *European Conference on Computer Vision*. Springer, 2024, pp. 38–55.
- [48] A. Kirillov *et al.*, “Segment anything,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4015–4026.
- [49] S. M. LaValle *et al.*, “Rapidly-exploring random trees: Progress and prospects: Steven m. lavalley, iowa state university, a james j. kuffner, jr., university of tokyo, tokyo, japan,” *Algorithmic and computational robotics*, pp. 303–307, 2001.
- [50] A. Hornung *et al.*, “Octomap: An efficient probabilistic 3d mapping framework based on octrees,” *Autonomous robots*, vol. 34, no. 3, 2013.
- [51] A. J. Aristorenas, “Machine learning framework for audio-based content evaluation using mfcc, chroma, spectral contrast, and temporal feature engineering,” 2024. [Online]. Available: <https://arxiv.org/abs/2411.00195>