2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Contrastive Auditory Knowledge Transfer for Tool-Mediated Robot Interaction with Granular Objects

Si Liu[1], Jindan Huang[1], Zhengyan Huan[2], Michael C. Hughes[1], Jivko Sinapov[1]

***Abstract*— Tool-mediated interactions enable robotics to manipulate and explore granular objects, producing informative auditory signals. A central challenge is transferring this perceptual knowledge across different tools and behaviors without costly data collection for each new context. We address this problem in the domain of audio-based recognition of granular and liquid-like objects. In this work, we leverage audio signals from tool-mediated interactions and learn context-agnostic representations for object recognition. We propose two contrastive learning approaches: a *shared-object transfer* method that performs supervised contrastive learning using audio data, and a *zero-shot transfer* method that integrates both audio and natural language descriptions of interaction contexts. Experiments on real-world data show that both methods achieve strong object recognition performance in unseen contexts, sometimes matching or exceeding a supervised baseline despite limited target-context data. Furthermore, the learned latent spaces exhibit clearly separable clusters by object identity, and the zero-shot method successfully recognizes novel objects, offering a practical solution for robot perception in data-scarce scenarios. The code for this paper is available at: `https://github.com/siliu6487/AuditoryKnowledgeTransfer`.**

I. Introduction

People have long leveraged tool-mediated perception to expand their ability to sense, manipulate, and understand objects beyond direct contact. Similarly, equipping robots with tools not only broadens the range of tasks they can perform but also enriches their perceptual experience [1]. Tool-mediated interaction generates rich streams of non-visual sensory data, such as audio and haptic signals, that reveal intrinsic object properties often invisible to visual-centric systems. These modalities provide complementary information, particularly when visual input is unavailable or unreliable [2]. A key challenge, however, lies in transferring perceptual knowledge across different contexts, such as when a robot switches tools or interaction behaviors. Traditional approaches often require extensive data collection for each new context, which is costly and time-consuming. Transfer learning presents a promising solution by enabling robots to leverage knowledge from prior tool-mediated interactions to recognize objects in new contexts.

In this work, we tackle the problem of cross-tool and cross-behavior knowledge transfer for object classification in robotic perception. We introduce a domain adaptation pipeline that learns context-agnostic audio representations in a shared latent space, preserving object-specific information while mitigating context-specific variations. Specifically, we utilize audio signals generated during tool-mediated interactions and propose two strategies for representation learning. The first relies on limited target-context data from a small set of shared objects between the source and target contexts. The second integrates natural language descriptions of interaction contexts with audio and achieves zero-shot object recognition. The contributions of this paper are threefold:

- We develop shared-latent-space transfer learning to the novel setting of **tool-mediated granular object perception**, enabling robots to generalize object knowledge across tools and behaviors.
- We empirically show that our methods learn **interpretable shared latent spaces**, where embeddings cluster by object identity regardless of tool- and behavior-specific variations. Notably, in some cases our transfer models even outperform baselines that are directly trained on target-context data.
- We introduce a **zero-shot transfer strategy** to demonstrate that robots can recognize entirely novel objects across various interaction contexts. This finding provides a practical path for scalable robot perception in data-scarce scenarios.

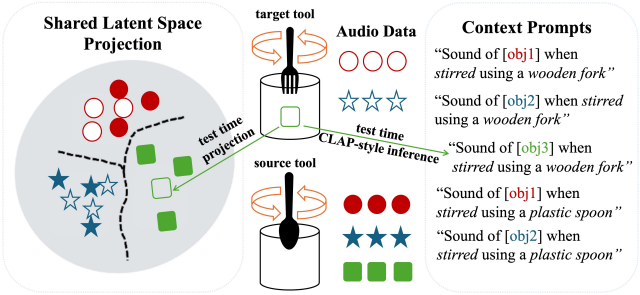

Fig. 1: Overview of knowledge transfer. A robot uses a target tool to interact with an object it has not previously encountered with that tool. However, the robot has prior interactions with the same object (green squares) and a shared set of other objects (red circles and blue stars) using a source tool. On the left, embeddings are learned in a shared latent space, where interactions with the same object are clustered together independent of tool identity. This technique enables novel interactions (green hollow square) to be projected near interactions with the same object. On the right side, a zero-shot CLAP-style learning method can infer the novel object (all green squares) during test time even though this object was never seen during training.

[1]Department of Computer Science; [2]Department of Electrical and Computer Engineering, School of Engineering, Tufts University, Medford, MA, USA. {firstname.lastname}@tufts.edu

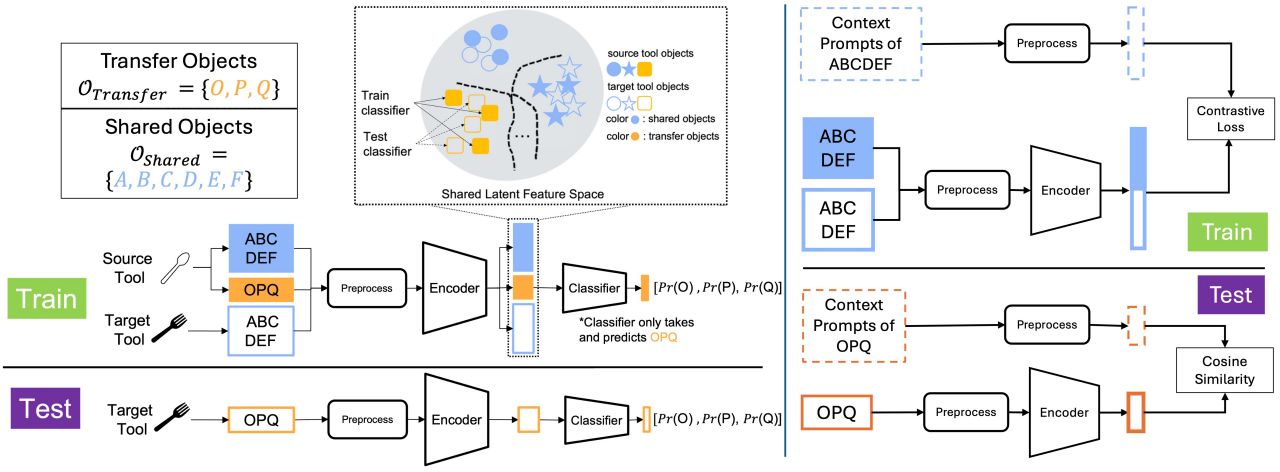

Fig. 2: Pipeline of cross-tool transfer in SINCERE-based and CLAP-based methods. In preprocessing, raw audio is embedded by a frozen pretrained audio encoder, and context prompts are embedded by a frozen pretrained language encoder. **Left (SINCERE-based):** source and target embeddings are projected into a shared latent space using overlapping objects, and a classifier is trained to predict transfer objects. **Right (CLAP-based):** audio embeddings are aligned with text embeddings via contrastive loss, enabling prediction by nearest text similarity at test time. Since transfer objects are unseen during training, CLAP constitutes a true zero-shot object recognition task.

II. Related Work

Researchers have explored a variety of non-visual sensory modalities for object recognition, such as tactile and haptic feedback [3], [4], which captures compliance and stiffness through contact. Among these, audio has emerged as a particularly informative channel, providing rich cues about material types and interaction dynamics [5], [6], motivating our focus on this modality. While previous studies highlight the potential of audio for object recognition, little work has examined its role in tool-mediated scenarios.

Like humans, robots benefit from exploratory interactions with objects to learn their properties [7], [8], [9], [10]. Beyond direct manipulation, tool use further expands a robot's exploratory capabilities, allowing it to learn properties from a wider range of objects. For example, [11] proposed a method to manipulate deformable food items with kitchen tools, enabling inference of physical properties such as elasticity and adhesiveness. [12] and [13] integrated multimodal sensory data gathered through multiple tools and behaviors to perceive object properties. While this type of work highlights the role of tools in mediating perception, it does not address how knowledge acquired in one context (e.g., a specific interaction behavior) can generalize to other new contexts or zero-shot object recognition under various contexts. Motivated by this limitation, a growing body of work explores perceptual knowledge transfer across contexts; however, these studies have primarily focused on rigid objects and have not considered tool use [14]. To fill this gap, our work investigates how robots can leverage audio from tool-mediated interactions to transfer granular object classification knowledge across tools and behaviors.

Recent advances in contrastive and language-grounded representation learning, such as CLIP [15] and CLAP [16], have enabled the learning of shared embedding spaces across modalities. These methods have shown strong generalization capabilities, including zero-shot recognition, by aligning image or audio data with semantic descriptions. CLIP-style methods have been proven effective in various domains. For example, [17] utilized multi-sensory data from robot interactions to learn a unified object property representation with CLIP models. [9] fused traffic scene graphs with visual and textual representations. On the other hand, supervised contrastive objectives such as SINCERE [18] have demonstrated effectiveness in transferable representations. However, these approaches have not been explored in the context of partial-overlap domain adaptation for tool-mediated robotic interaction. In this work, we bridge these directions by studying cross-tool and cross-behavior knowledge transfer using audio from tool-mediated interactions. We instantiate contrastive and language-aligned objectives in context-induced distribution shifts in non-visual robot perception.

III. Methodology

A. Notation and Problem Formulation

Consider a robot performing exploratory behaviors $\mathcal{B}$ (e.g., stirring, poking) with a set of tools $\mathcal{T}$ (e.g., spoon, chopsticks) on granular objects $\mathcal{O}$ (e.g., wheat, salt), while recording audio data. Each interaction consists of the robot using a tool $t \in \mathcal{T}$ and a behavior $b \in \mathcal{B}$ to interact with an object $o \in \mathcal{O}$, producing auditory data $x_i^{t,b,o}$. We denote the set of data gathered under the context (t, b, o) as $\mathcal{X}^{t,b,o} = \{x_m^{t,b,o}\}_{m=1}^{M}$, where M is the number of repeated trials and m is the index of the trial.

We are interested in recognizing a specific set of objects in new tool-behavior contexts via transfer learning. We denote the set of all tool-behavior pairs by $\mathcal{C} = \mathcal{T} \times \mathcal{B}$. The

set of source contexts is denoted as $\mathcal{C}_{\text{source}} \subset \mathcal{C}$ and the distinct target context is denoted as $\mathcal{C}_{\text{target}} \in \mathcal{C} \setminus \mathcal{C}_{\text{source}}$, where $|\mathcal{C}_{\text{target}}| = 1$. We focus on target contexts that differ from the source contexts in only one element. In cross-tool transfer, the set of represented behaviors is the same across $\mathcal{C}_{\text{source}}$ and $\mathcal{C}_{\text{target}}$, while the tools differ. Fig. 1 illustrates an example of cross-tool transfer setting with a single source tool. Similarly, in cross-behavior transfer, the same tools appear in $\mathcal{C}_{\text{source}}$ and $\mathcal{C}_{\text{target}}$, while the behaviors differ.

We further distinguish between two subsets of objects. First, shared objects $\mathcal{O}_{\text{shared}}$ appear in both the source contexts $\mathcal{C}_{\text{source}}$ and the target context $\mathcal{C}_{\text{target}}$, serving as a bridge for learning transferable representations. In contrast, transfer objects $\mathcal{O}_{\text{transfer}}$ are observed only in source contexts or neither context during model training. These objects represent novel cases that require generalization.

We thus partition all collected data $\mathcal{X}$ into four parts:

$$\begin{aligned}
\mathcal{X}^{\text{shared}}_{\text{source}} &= \{\mathcal{X}^{t,b,o} \mid (t,b) \in \mathcal{C}_{\text{source}},\ o \in \mathcal{O}_{\text{shared}}\},\\
\mathcal{X}^{\text{transfer}}_{\text{source}} &= \{\mathcal{X}^{t,b,o} \mid (t,b) \in \mathcal{C}_{\text{source}},\ o \in \mathcal{O}_{\text{transfer}}\},\\
\mathcal{X}^{\text{shared}}_{\text{target}} &= \{\mathcal{X}^{t,b,o} \mid (t,b) \in \mathcal{C}_{\text{target}},\ o \in \mathcal{O}_{\text{shared}}\},\\
\mathcal{X}^{\text{transfer}}_{\text{target}} &= \{\mathcal{X}^{t,b,o} \mid (t,b) \in \mathcal{C}_{\text{target}},\ o \in \mathcal{O}_{\text{transfer}}\}.
\end{aligned} \tag{1}$$

Our primary goal is to achieve cross-tool or cross-behavior object knowledge transfer. At test time, we evaluate the recognition accuracy of transfer objects in the target context given $\mathcal{X}^{\text{transfer}}_{\text{target}}$. We assume access to interaction data from shared objects, i.e., $\mathcal{X}^{\text{shared}}_{\text{source}}$ and $\mathcal{X}^{\text{shared}}_{\text{target}}$. If the setting is not zero-shot transfer, the training data also includes source-context observations of the transfer set of objects $\mathcal{X}^{\text{transfer}}_{\text{source}}$.

To accomplish this knowledge transfer, we train an encoder e that maps audio data $x^{t,b,o}$ into embeddings z in the space shared by the source and target contexts. This projection preserves object-specific features, while either minimizing context-specific variations or aligning audio data with corresponding contextual text data. The resulting representation then serves as input to the object classifier f, which estimates the probability of the transfer object:

$$f\big(e(x^{t,b,o})\big) \ \rightarrow\ [\Pr(o \mid x^{t,b,o}) \mid o \in \mathcal{O}_{\text{transfer}}].$$

B. Representation Learning in Shared Latent Space

We consider two ways to train encoders that project data into representations in a shared latent space. First, we propose a method based solely on audio data, leveraging supervised contrastive learning. Second, we employ an adaptation of the popular CLAP [16] method to incorporate both audio and text data.

SINCERE loss for audio: Given a dataset of feature vectors x and corresponding object labels y, we can employ supervised contrastive learning to learn representations such that embeddings of the same objects are pushed together while representations of different objects are pulled apart. In particular, we use a loss named Supervised InfoNCE REvisited (SINCERE) [18]. This loss has been shown effective for transfer learning in image classification while avoiding issues with previous supervised contrastive losses like SupCon [19].

The loss function is defined in terms of pairs of examples. Let x_A be an anchor instance and x_p be a positive partner of the same class. The SINCERE loss for this pair is

$$\mathcal{L}_{\text{SINCERE}}(z_A, z_p) = -\log \frac{e^{z_A \cdot z_p/\tau}}{e^{z_A \cdot z_p/\tau} + \sum_{n \in \mathcal{N}} e^{z_n \cdot z_p/\tau}}, \tag{2}$$

where z_A is the embedding of x_A, z_p is the embedding of x_p, $\mathcal{N}$ is the set of all negative embeddings from classes other than z_A's class in the same batch, z_n is any particular negative embedding in $\mathcal{N}$, and the temperature $\tau > 0$ is a hyperparameter that scales the pairwise similarities. The total loss for the batch sums over any pairs of instances that share the same label.

We emphasize that SINCERE facilitates knowledge transfer across contexts because the object class labels instances have in common drive learning. Embeddings of the same object will be pushed closer together despite different contexts.

CLAP-style loss for audio and text: To explore methods that may enable zero-shot object recognition in our transfer setting, we examine an unsupervised language-grounded representation learning called Contrastive Language-Audio Pretraining (CLAP) [16]. CLAP projects the audio and text embeddings in a shared latent space. To obtain text for a specific audio instance $x_i^{t,b,o}$, we represent its t, b, o triple in words as in Fig. 1.

Given a batch of N paired examples $\{(x_i^{\text{aud}}, x_i^{\text{txt}})\}_{i=1}^{N}$, CLAP minimizes a symmetric contrastive loss:

$$\mathcal{L}_{\text{CLAP}} \ = \ -\frac{1}{2N} \sum_{i=1}^{N} \left[\log \frac{e^{C_{ii}}}{\sum_{j=1}^{N} e^{C_{ij}}} \ + \ \log \frac{e^{C_{ii}}}{\sum_{j=1}^{N} e^{C_{ji}}} \right], \tag{3}$$

where C_{ij} denotes cosine similarity between the audio embedding of x_i^{aud} and the text embedding of x_j^{txt}. This loss aligns the diagonal entries of C (paired audio-text from the same instance) while pushing apart off-diagonal entries that represent different instances. Importantly, this is instance discrimination rather than supervised classification; some unpaired instances may share the same object class but not the same behavior or tool.

For text representation, we convert each tool-behavior-object context into a single plain English phrase or prompt. Each prompt can be fed into CLAP's pre-trained text model h to obtain an embedding in the shared latent space. To increase robustness, we generate K paraphrased prompts for each tool-behavior-object triple. For example, with $K = 3$, we use the following templates: `"[obj] in a container being [behav] with a [tool]"`, `"A [tool] is used to [behav] [obj] in a container"`, `"Sound of [obj] when [behav] using a [tool]"`. Each template is encoded with the pre-trained text models h, and their embeddings are averaged to form a unified representation for the interaction:

$$\frac{1}{K} \sum_{k=1}^{K} h\big(\text{template}_k(t,b,o)\big). \tag{4}$$

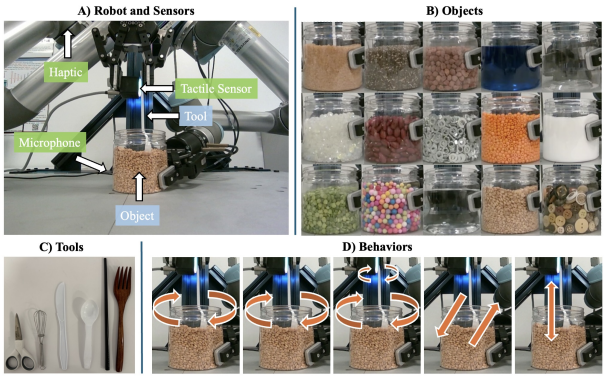

Fig. 3: **A)** Data collected during a robot's interaction with 15 objects using 6 tools. **B)** 15 objects (row-wise, left to right): *cane-sugar*, *chia-seed*, *chickpea*, *detergent*, *empty*, *glass-bead*, *kidney-bean*, *metal-nut-bolt*, *plastic-bead*, *salt*, *split-green-pea*, *styrofoam-bead*, *water*, *wheat*, and *wooden-button*. **C)** 6 tools (left to right): *metal-scissor*, *metal-whisk*, *plastic-knife*, *plastic-spoon*, *wooden-chopstick*, and *wooden-fork*. **D)** 5 behaviors (left to right) and their approximate duration: *stirring-slow* (50s), *stirring-fast* (20s), *stirring-twist* (76s), *whisk* (90s), and *poke* (50s).

IV. Experiment Design and Implementation

A. Data and Preprocessing

We validate the effectiveness of the proposed methods using the dataset from [12], which includes sensory data captured during interactions between a UR5 robot arm and various objects, using six distinct tools to perform five different behaviors (see Fig. 3). Under each tool-behavior combination, the robot interacts with 15 unique objects. Each interaction is repeated 10 times, with each repetition defined as a trial. The dataset consists of 4500 samples in total, derived from 10 trials across 450 unique combinations of tool-behavior-object (6 tools $\times$ 5 behaviors $\times$ 15 objects). We take the **audio** data as non-visual information for object recognition, as audio is the dominant modality for object recognition task with this dataset [13]. We trim all audio samples to 20 seconds to unify the length of all behaviors, and this decision is supported by the finding that the robot only needs the first few seconds of behaviors to maximize the recognition rate [13]. We resample the data at $48,000$ Hz before feeding them into the frozen pre-trained CLAP-style model from [20]. We extract the text embeddings of prompts for all tool-behavior-object combinations using the same model's text encoder. The audio and text are vectors of size 512 and the inputs to our representation learning encoders.

B. Implementation Details

Overall Implementations: To train the encoder and classifier, we use PyTorch's AdamW [21] optimizer. The encoder is a fully-connected neural network with an input layer of size equal to the data dimension, one hidden layer of size 256 with ReLU activations, and an output layer of size 128 for SINCERE-based encoder and 512 for CLAP-based encoder. The CLAP embeddings stay 512 so that we can compute cosine similarity with the text input. The output of the encoder is L2-normalized. We use linear probing to evaluate how well an encoder learns with each loss function and generalizes to downstream classification tasks. The classifier's output layer size is equal to the number of object classes in the test set. The learning rate of the classifier and encoder is 0.0001.

From a computational perspective, the main cost arises from passing audio through the frozen pretrained CLAP audio encoder to obtain embeddings. Text prompt embeddings can be computed and cached in advance, assuming the contexts and objects are known. After feature extraction, our transfer encoders are lightweight (a single hidden layer MLP), and inference requires only a forward pass followed by either linear classification (SINCERE) or cosine similarity computation (CLAP). Therefore, the additional cost introduced by the transfer framework is negligible compared to audio feature extraction, making the approach practical once embeddings are obtained.

C. Experiment Design

Transfer Learning Tasks: We consider two transfer tasks: ***cross-tool*** *transfer* and ***cross-behavior*** *transfer*. In the cross-tool setting, the source and target contexts differ only in the tool used (e.g., *scissor-stirring* as the source context and *spoon-stirring* as the target context). In the cross-behavior setting, the contexts differ only in the behavior performed (e.g., *scissor-stirring* as the source context and *scissor-poke* as the target context).

Transfer Settings: In both cross-tool and cross-behavior transfer, knowledge is transferred to a single target context (a specific tool–behavior pair). We consider two settings: ***1-to-1*** *transfer*, where data comes from a single alternative source context, and ***other-to-1*** *transfer*, where data comes from all other available contexts except the target. For example, in cross-tool transfer, the robot may learn from one source tool (*1-to-1*) or from all five tools other than the target (*other-to-1*), while keeping the behavior fixed. In cross-behavior transfer, the robot may learn from one source behavior (*1-to-1*) or from all four behaviors other than the target (*other-to-1*), while keeping the tool fixed. A detailed illustration is provided in Section V-A.

Transfer Methods: We introduce two methods to train the encoder in the transfer pipeline:

- *SINCERE-based Encoder*: Uses the SINCERE loss to learn a shared representation across contexts, trained on $\mathcal{X}_{\text{source}}^{\text{shared}}$, $\mathcal{X}_{\text{source}}^{\text{transfer}}$, and $\mathcal{X}_{\text{target}}^{\text{shared}}$.
- *CLAP-based Encoder*: Uses the CLAP framework to learn cross-context representations from $\mathcal{X}_{\text{source}}^{\text{shared}}$ and $\mathcal{X}_{\text{target}}^{\text{shared}}$ by conditioning on descriptive prompts for different contexts.

In the rest of the paper, we will use "SINCERE" and "CLAP" as shorthand to denote the corresponding transfer learning methods. When referring specifically to the objec-

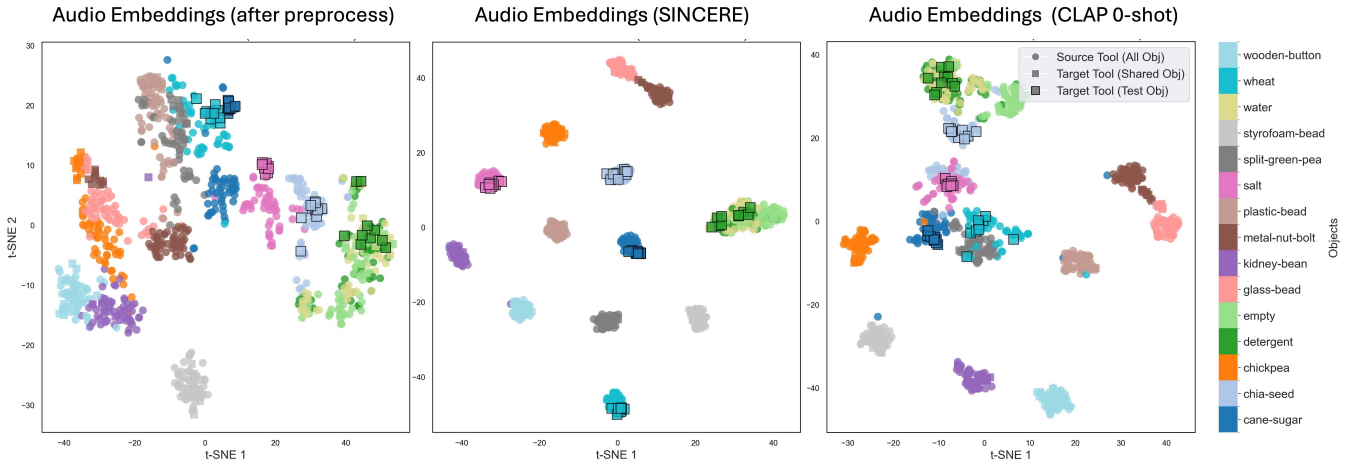

Fig. 4: Visualization of audio embeddings using t-SNE under a cross-tool transfer setting. Behavior: stirring-fast; Target tool (shown as squares): metal-scissor; Source tools (shown as circles): 5 other tools. Transfer (test) objects: {cane sugar, detergent, chia seed, salt, wheat}. Black-edged squares correspond to target-tool interactions with transfer objects, which were unseen with this tool during training. **Left**: embeddings obtained directly from the pretrained audio feature extractor, where some transfer-object samples under the target context are not close to their corresponding object clusters. **Middle**: embeddings projected by the SINCERE-style encoder, where transfer objects are clustered correctly in the latent space. **Right**: embeddings projected by the CLAP-style encoder. Even though the CLAP encoder has never seen the transfer objects during training, it still projects most test samples close to their corresponding clusters.

tive functions, we will explicitly write "SINCERE loss" or "CLAP loss".

Baselines: We compare our models with two baselines:

- *Baseline 1 (B1)*: Assumes no novel interaction in the target context by using $\mathcal{X}_{\text{target}}^{\text{transfer}}$ for encoder and classifier training. Data are split by trials instead of contexts. For each tool-behavior-object combination, we perform 5-fold cross-validation over the 10 trials: in each fold, 8 trials are used for training and 2 held-out trials are used for testing, so that every trial is evaluated once. Reported results are averaged across the 5 folds.
- *Baseline 2 (B2)*: Trained only on $\mathcal{X}_{\text{source}}^{\text{shared}}$ and $\mathcal{X}_{\text{source}}^{\text{transfer}}$ with no access to target context data (tool or behavior) during encoder training. It uses the same test and validation sets as our knowledge transfer models, ensuring a fair comparison under the assumption of no prior exposure to the target context.

We emphasize that B1 is not intended as a fully supervised upper bound, but rather as a practical data-scarce condition reflecting real-world robotic settings, where collecting interaction data for every tool–behavior combination is time-consuming and expensive. Our goal is not to outperform models trained with abundant target-context data, but to evaluate how effectively object knowledge can be transferred when target data are limited. In this sense, B2 provides a strict no-target baseline, while B1 represents a minimal-target-data scenario.

To ensure fair comparison, all models in our experiment share the same pipeline, except for the input data context based on the experiment. Fig. 2 demonstrates the transfer pipelines. Even though baselines do not perform knowledge transfer, they use the same encoder architecture. In this case, the encoder functions as a supervised representation learner that enhances linear probing recognition performance.

Sampling Procedure: We evaluate robustness and generalization by repeatedly sampling 5 transfer objects (without replacement) in 10 rounds. Each transfer round includes at most one of the classes *empty*, *water*, or *detergent*, since these objects produce little or no sound during interaction and are therefore difficult to distinguish from audio alone. The sampling procedure is balanced, leading to similar appearing frequencies for all objects. Final results are reported as average recognition accuracy among all sampling rounds.

Evaluation Metrics: We use *object recognition accuracy* as our primary performance metric, defined as:

$$A = \frac{\text{correct predictions}}{\text{total predictions}}. \tag{5}$$

Each object class in each 5-class target-transfer set $\mathcal{X}_{\text{target}}^{\text{transfer}}$ is balanced, with random guess accuracy of 20%.

V. Results

A. Illustrative Example

Cross-Tool and Cross-behavior: During cross-tool transfer, a robot performs a target behavior *stirring-fast* using a target tool *plastic-spoon* on 10 out of 15 objects. In the source context, the robot still performs *stirring-fast* but uses a *wooden-chopstick* on all 15 objects. In the 5-to-1 cross-tool transfer setting, the robot in the source context uses all 5 tools that are not *plastic-spoon* while still performing *stirring-fast*. In a cross-behavior transfer setting, in a target context, a robot performs a behavior *stirring-slow* using a

TABLE I: Cross-tool and cross-behavior transfer results. The values are reported as percentages ($\pm$ std). Results are averaged across behaviors for tool transfer or across tools for behavior transfer. Highlighted values indicate performance better than baseline B1, which was trained and tested within the target context. Random guess accuracy is 20% for 5-class classification.

	t_{target} or b_{target}	B1	B2 (1-to-1)	SINCERE (1-to-1)	B2 (other-to-1)	SINCERE (other-to-1)	CLAP (zero-shot) (other-to-1)
Cross Tool	metal-scissor	89.72 ± 7.80	80.62 ± 4.13	82.90 ± 5.00	89.40 ± 3.55	89.52 ± 3.08	69.32 ± 5.64
	metal-whisk	88.76 ± 7.96	85.98 ± 2.38	86.94 ± 4.69	**94.04** ± 3.74	**94.32** ± 3.77	71.72 ± 3.66
	plastic-knife	88.20 ± 7.11	84.86 ± 2.63	86.75 ± 6.42	**94.48** ± 3.83	**94.80** ± 3.76	67.88 ± 3.43
	plastic-spoon	89.52 ± 4.86	84.66 ± 5.75	86.23 ± 3.91	**93.92** ± 2.86	**94.20** ± 3.09	67.52 ± 2.54
	wooden-chopstick	86.96 ± 9.86	81.66 ± 4.12	85.71 ± 5.51	**92.20** ± 4.87	**92.80** ± 5.23	66.44 ± 3.02
	wooden-fork	91.04 ± 5.26	83.86 ± 4.09	86.84 ± 4.14	**93.84** ± 3.43	**94.80** ± 2.09	67.16 ± 4.65
	Average	89.03 ± 1.40	83.61 ± 2.05	85.90 ± 1.54	**92.98** ± 3.36	**93.40** ± 3.18	68.34 ± 3.11
Cross Behavior	stirring-slow	90.20 ± 1.40	62.43 ± 5.69	77.84 ± 3.98	83.36 ± 4.38	88.17 ± 1.91	66.77 ± 1.33
	stirring-fast	94.73 ± 1.65	65.88 ± 4.31	79.57 ± 2.83	80.33 ± 3.25	88.97 ± 1.43	53.90 ± 4.50
	stirring-twist	84.40 ± 3.22	54.83 ± 6.30	75.74 ± 4.10	75.93 ± 6.07	**86.73** ± 1.88	43.16 ± 5.50
	whisk	95.90 ± 1.09	65.98 ± 6.27	80.03 ± 3.47	84.06 ± 4.64	90.10 ± 4.34	53.46 ± 3.93
	poke	79.93 ± 5.45	49.31 ± 7.01	64.91 ± 5.30	57.46 ± 7.59	73.46 ± 5.40	49.10 ± 5.39
	Average	89.03 ± 6.81	59.68 ± 7.36	75.62 ± 6.22	76.23 ± 2.63	85.49 ± 1.91	53.28 ± 1.89

plastic-spoon on 10 out of 15 objects. In the source context, the robot still uses *plastic-spoon* but performs behavior *poke* on all 15 objects. In the 4-to-1 cross-behavior transfer setting, it performs all 4 behaviors that are not *stirring-slow* with the *plastic-spoon*.

Zero-shot CLAP: In this experiment, the robot does not interact with the transfer objects during data collection, regardless of the source or target tool–behavior contexts. For example, the robot may have interacted with only 10 objects using the behavior *stirring-fast* with tools such as a *wooden-chopstick* and a *plastic-spoon*. During inference, it is asked to recognize 5 previously unseen objects under target context *stirring-fast* with the *plastic-spoon*. Recognition of these novel objects relies on aligning audio observations with text descriptions of interaction contexts. This design enables zero-shot recognition of unseen objects in the target context.

B. Visualization of representations

In Figure 4, we use T-SNE [22] to project the learned high-dimensional embeddings into 2 dimensions for visualization. The left panel visualizes the preprocessed audio data. There are clear clusters by object, reflecting the generalization capability of the large pre-trained audio model. However, certain objects, such as *detergent*, *water*, and *empty*, remain challenging to distinguish, as tool-mediated interactions with these objects create highly similar sounds even to human ears.

The middle panel of Fig. 4 shows the shared latent space projected by our SINCERE encoder, where most objects occupy different areas with clear linear boundaries. The transfer objects are accurately projected to their own clusters. However, the acoustically similar objects that overlap in the left panel remain mixed, indicating that representation learning alone cannot fully separate objects whose auditory signals are intrinsically similar. The right panel shows the projection of the CLAP encoder. Even though the CLAP encoder has never seen the test (transfer) objects, it's still able to project most of them to their corresponding clusters.

C. Quantitative Results

Due to the large number of transfer experiments, results are averaged across behaviors for cross-tool transfer and across tools for cross-behavior transfer, as reported in Table I. B1, trained directly on the target context, achieves accuracies around 89%. While B1 is treated as a baseline for comparison, it cannot be considered a true upper bound. The training data available to B1 is highly limited – only 8 trials from each tool-behavior-object combination, with 2 additional trials reserved for cross-validation. Although preprocessed embeddings already cluster strongly by object, there remain challenging cases where subtle auditory cues make discrimination difficult. For example, the auditory data from *salt* and *cane sugar* during *poke* with a pointy *wooden chopstick* are weak signals hard to separate. When transfer is limited to a single source context (B2 and SINCERE 1-to-1), performance drops substantially, showing that transfer learning from a single source context is insufficient to achieve high performance.

In contrast, aggregating multiple sources (other-to-1) closes much of the gap to B1, with several cases achieving comparable performance, particularly in the cross-tool setting. As a result, other-to-1 B2 and SINCERE can sometimes achieve performance comparable to or even exceeding that of B1, despite B1 having direct access to the target context. For example, SINCERE outperforms B1 by around 10% when the target context is *poke* and *wooden-chopstick* by transferring the knowledge from all other tools using this behavior. CLAP faces a more challenging condition: the target objects are entirely novel. CLAP achieves 68% (cross-tool) and 53% (cross-behavior), far above random chance (20%), demonstrating that the learned representation captures object-related structure that generalizes to novel objects across diverse interaction contexts.

We also observe a clear difference between cross-tool and cross-behavior transfer. Tools tend to produce more similar auditory distributions when performing the same behavior on the same object (e.g., a *plastic spoon* and a *wooden fork* both generate comparable sound patterns when

stirring *wheat*). As a result, B2 transfer across tools achieves performance close to or better than B1, since source tool data partially represents the target tool's distribution. In contrast, behaviors create much larger shifts in the audio domain: for example, *poking* is quiet and infrequent, while *fast stirring* is loud and continuous. These larger distributional differences make cross-behavior transfer substantially harder. In this setting, SINCERE outperforms B2 by incorporating target-context data from non-test objects during training. SINCERE provides anchor points linking source and target contexts in the latent space, supervised by object identity. As a result, the transfer performance improves, particularly when behavior-related auditory distributions differ substantially. On the other hand, we hypothesize that the large performance gap between cross-tool and cross-behavior settings in CLAP stems from the method's reliance on instance-level cross-modal alignment rather than object-level supervision. Large behavior-dependent changes in the audio signal may make it more difficult to preserve object-level consistency across contexts, especially for unseen objects.

These results highlight three key insights: (1) transfer learning is most effective when multiple related contexts are leveraged; (2) CLAP enables meaningful recognition under the challenging setting of zero-shot object recognition, although a performance gap remains compared to supervised baseline B1; and (3) transfer learning can reduce the performance drop caused by domain shift, but large domain shifts—such as those induced by different behaviors—remain difficult to overcome.

VI. Conclusion And Future Work

We present two approaches for auditory knowledge transfer in tool-mediated robot interaction with granular objects. The shared-object transfer method (SINCERE-based) exploits limited target-context data to improve representations, often matching or surpassing a supervised baseline, sometimes in the challenging cross-behavior setting. The zero-shot transfer method (CLAP-based) extends this framework by aligning audio with language descriptions of interaction contexts, enabling recognition of novel objects. Despite the difficulty of this setting, CLAP achieves notably above-chance performance, demonstrating the viability of zero-shot auditory recognition. Our findings also reveal that cross-tool transfer is generally easier than cross-behavior transfer, reflecting the greater acoustic similarity across tools compared to across behaviors. Overall, this work shows that contrastive learning provides a powerful foundation for transferring auditory knowledge in tool-mediated robot interaction, and points toward scalable approaches for robust robot perception in diverse contexts.

One limitation of this work is that we only used audio data for object recognition. This may pose challenges especially when the sounds produced during interactions with different objects are similar or barely audible. Tool-dependent effects, such as quieter audio from a flexible tool versus louder audio from a rigid tool, may also lead to inconsistencies in object perception. Future work will focus on incorporating other modalities, such as tactile and force feedback, to learn object properties more comprehensively. Additionally, in this work, text descriptions of interaction contexts are manually defined using a small set of templates. Prompt learning offers a scalable alternative by automatically optimizing text embeddings, which helps capture more discriminative features beyond domain-invariant constraints. Future work will explore integrating prompt learning with domain adaptation to enhance cross-context knowledge transfer. Finally, the current evaluation is conducted on a single existing dataset consisting of offline audio recordings. Future work will explore deployment in small-scale real-robot online experiments to evaluate real-time inference performance. Collecting new datasets across different robots and environmental conditions would further strengthen the generalization of context-agnostic auditory representations.

References

[1] Meiying Qin, Jake Brawer, and Brian Scassellati. Robot tool use: A survey. Frontiers in Robotics and AI, 9:1009488, 2023.

[2] Jivko Sinapov, Taylor Bergquist, Connor Schenck, Ugonna Ohiri, Shane Griffith, and Alexander Stoytchev. Interactive object recognition using proprioceptive and auditory feedback. The International Journal of Robotics Research, 30(10):1250–1262, 2011.

[3] Huaping Liu, Yupei Wu, Fuchun Sun, and Di Guo. Recent progress on tactile object recognition. International Journal of Advanced Robotic Systems, 14(4):1729881417717056, 2017.

[4] Nicolas Gorges, Stefan Escaida Navarro, Dirk Göger, and Heinz Wörn. Haptic object recognition using passive joints and haptic key features. In 2010 IEEE International Conference on Robotics and Automation, pages 2349–2355. IEEE, 2010.

[5] Loic Lacheze, Yan Guo, Ryad Benosman, Bruno Gas, and Charlie Couverture. Audio/video fusion for objects recognition. In 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, pages 652–657. IEEE, 2009.

[6] Hua Zhang, Xiaochun Cao, and Rui Wang. Audio visual attribute discovery for fine-grained object recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 32, 2018.

[7] Gyan Tatiya and Jivko Sinapov. Deep multi-sensory object category recognition using interactive behavioral exploration. In 2019 International Conference on Robotics and Automation (ICRA), pages 7872–7878. IEEE, 2019.

[8] Xiaohui Chen, Ramtin Hosseini, Karen Panetta, and Jivko Sinapov. A framework for multisensory foresight for embodied agents. In 2021 IEEE International Conference on Robotics and Automation (ICRA), pages 10927–10933. IEEE, 2021.

[9] Aaron Lohner, Francesco Compagno, Jonathan Francis, and Alessandro Oltramari. Enhancing vision-language models with scene graphs for traffic accident understanding. In 2024 IEEE International Automated Vehicle Validation Conference (IAVVC), pages 1–7, 2024.

[10] Weihua Wang, Xiaofei Li, Yanzhi Dong, Jun Xie, Di Guo, and Huaping Liu. Natural language instruction understanding for robotic manipulation: a multisensory perception approach. In 2023 IEEE International Conference on Robotics and Automation (ICRA), pages 9800–9806, 2023.

[11] Mevlana C Gemici and Ashutosh Saxena. Learning haptic representation for manipulating deformable food objects. In 2014 IEEE/RSJ international conference on intelligent robots and systems, pages 638–645. IEEE, 2014.

[12] Gyan Tatiya, Jonathan Francis, and Jivko Sinapov. Cross-tool and cross-behavior perceptual knowledge transfer for grounded object recognition. arXiv preprint arXiv:2303.04023, 2023.

[13] Si Liu and Jivko Sinapov. Tool-mediated robot perception of granular substances using multiple sensory modalities. In 2025 IEEE International Conference on Robotics and Automation (ICRA), pages 3773–3779. IEEE, 2025.

[14] Gyan Tatiya, Yash Shukla, Michael Edegware, and Jivko Sinapov. Haptic knowledge transfer between heterogeneous robots using kernel manifold alignment. In 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 5358–5363. IEEE, 2020.

[15] 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.

[16] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1–5. IEEE, 2023.

[17] Gyan Tatiya, Jonathan Francis, Ho-Hsiang Wu, Yonatan Bisk, and Jivko Sinapov. Mosaic: Learning unified multi-sensory object property representations for robot learning via interactive perception. In 2024 IEEE International Conference on Robotics and Automation (ICRA), pages 15381–15387. IEEE, 2024.

[18] Cynthia Feeney and Michael C Hughes. Sincere: Supervised information noise-contrastive estimation revisited. arXiv preprint arXiv:2309.14277, 2024.

[19] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. Advances in neural information processing systems, 33:18661–18673, 2020.

[20] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1–5. IEEE, 2023.

[21] I Loshchilov. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101, 2017.

[22] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. Journal of machine learning research, 9(11), 2008.