MOSAIC: Learning Unified Multi-Sensory Object Property Representations for Robot Learning via Interactive Perception

Gyan Tatiya[1*]   Jonathan Francis[2]   Ho-Hsiang Wu[2]   Yonatan Bisk[3]   Jivko Sinapov[1]

Abstract— A holistic understanding of object properties across diverse sensory modalities (e.g., visual, audio, and haptic) is essential for tasks ranging from object categorization to complex manipulation. Drawing inspiration from cognitive science studies that emphasize the significance of multi-sensory integration in human perception, we introduce MOSAIC (Multimodal Object property learning with Self-Attention and Interactive Comprehension), a novel framework designed to facilitate the learning of unified multi-sensory object property representations. While it is undeniable that visual information plays a prominent role, we acknowledge that many fundamental object properties extend beyond the visual domain to encompass attributes like texture, mass distribution, or sounds, which significantly influence how we interact with objects. In MOSAIC, we leverage this profound insight by distilling knowledge from multimodal foundation models and aligning these representations not only across vision but also haptic and auditory sensory modalities. Through extensive experiments on a dataset where a humanoid robot interacts with 100 objects across 10 exploratory behaviors, we demonstrate the versatility of MOSAIC in two task families: object categorization and object-fetching tasks. Our results underscore the efficacy of MOSAIC's unified representations, showing competitive performance in category recognition through a simple linear probe setup and excelling in the fetch object task under zero-shot transfer conditions. This work pioneers the application of sensory grounding in foundation models for robotics, promising a significant leap in multi-sensory perception capabilities for autonomous systems. We have released the code, datasets, and additional results: `https://github.com/gtatiya/MOSAIC`.

I. INTRODUCTION

Humans first acquire knowledge about object properties through physical interaction—a process that involves the integration of multiple sensory inputs, including visual, auditory, and tactile cues [1]–[7]. For instance, we rely on vision to discern an object's color, sense of touch when we lift an object to gauge its weight, and hearing when we shake a container to determine if it is full or empty. The fusion of such multi-sensory information is pivotal in shaping our perception and guiding our decision-making processes concerning objects [8]–[13]. Similarly, robots can effectively engage with objects by simultaneously perceiving and processing multi-sensory signals, to tackle tasks such as object categorization [14], [15], material recognition [16], and even complex actions like packing and pouring [17].

Within vision and text, large-scale Vision-Language Models (VLMs) have demonstrated their ability to provide state-of-the-art representations for both visual and textual modalities, making them exceptionally valuable for a wide range of AI applications [18]–[20]. One such model, Contrastive Language-Image Pre-training (CLIP) [18], is trained from scratch on an extensive dataset comprising 400 million (image, text) pairs. CLIP's representations can seamlessly transfer to many downstream tasks without fine-tuning. While prior research has primarily focused on integrating audio modalities into CLIP's embedding space [21], including a robot's haptic data into this versatile space has yet to be explored. We address this gap by distilling the domain-general language grounding within CLIP and infusing it into a robot's sensory data from object interactions. This method effectively mitigates the often prohibitive costs of collecting interactive data by robots through extensive object exploration. The primary objective of this study is to expose VLMs to object property representations derived from robot interactions, highlighting how these representations can significantly improve the performance on interactive tasks by enhancing the robot's multimodal perceptual capabilities. This enhancement arises from interactive object exploration to understand the fundamentals of object properties, a perspective disembodied representations often lack.

We introduce our method and framework for learning **M**ultimodal **O**bject properties with **S**elf-**A**ttention and **I**nteractive **C**omprehension (MOSAIC)—an approach for acquiring versatile representations that are adaptable to various interactive perception tasks within robotics. MOSAIC is designed to extract unified multi-sensory object property representations, enabling understanding of object properties by leveraging diverse sensory modalities. This approach rests on the premise that natural language provides a versatile embedding space whose knowledge we can distill and align to different sensory modalities. We evaluate our approach on a publicly available dataset where a humanoid robot explored 100 objects, using 10 exploratory behaviors while recording sensory data, including vision, audio, and haptic. We evaluate on both object category recognition and the fetch object task, finding MOSAIC to be robust and adaptible. MOSAIC's performance in the object category recognition is notably competitive compared to state-of-the-art methods, showing the effectiveness of unified representations even within a straightforward linear probe setup. Furthermore, MOSAIC demonstrates exceptional capabilities in executing natural language instructions in the fetch object task under a zero-shot condition. In summary, MOSAIC offers a versatile framework for multimodal object property learning, bridging the gap between different sensory inputs and facilitating a wide range of downstream robot tasks.

*Work done during internship at Bosch Center for Artificial Intelligence.
[1]Department of Computer Science, Tufts University, Email: {Gyan.Tatiya, Jivko.Sinapov}@tufts.edu. [2]Bosch Center for AI, Email: {Jon.Francis, Ho-Hsiang.Wu}@us.bosch.com. [3]Carnegie Mellon University, Email: ybisk@andrew.cmu.edu.
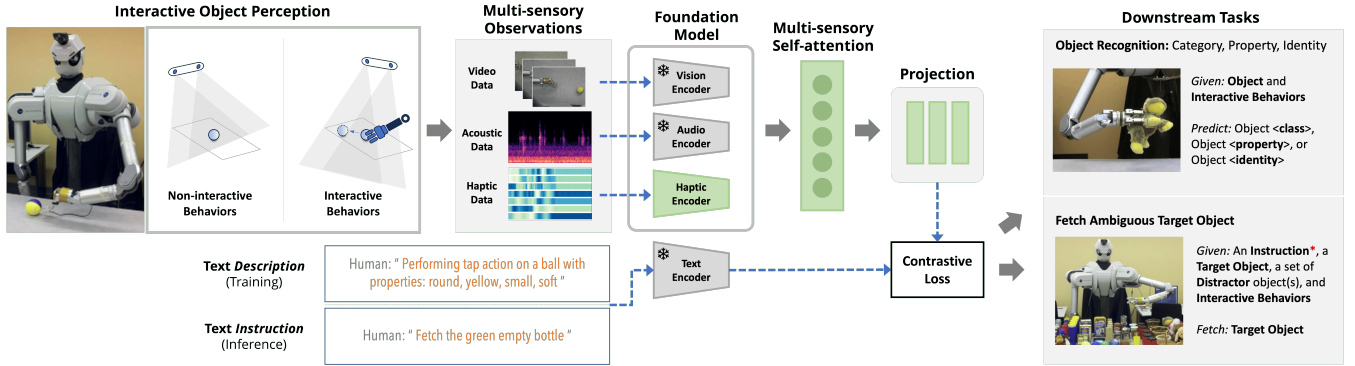
Fig. 1: **Overview of the MOSAIC Framework**: Initially, the robot collects sensory data through object exploration, which is then used to train models for distilling unified multimodal representations guided by a pre-trained text encoder. These acquired representations are subsequently applied to a variety of downstream tasks.

II. RELATED WORK

Multi-sensory Learning in Cognitive Science. Humans acquire knowledge about object properties through physical interactions, integrating multiple sensory signals [22]–[24]. Multi-sensory integration and attention processes occur at various stages in the human brain, crucially influencing our perception of objects and task performance [25]. Moreover, human perception involves the dynamic interplay between sensory inputs and existing cognitive knowledge rather than processing sensory inputs in isolation [26]. Our research extends these principles to robotics, extracting knowledge from pre-trained text encoders to align representations across diverse sensory modalities—mirroring how humans fuse sensory information with their established knowledge to perceive their environment holistically.

Interactive Perception. Robotics research has showcased the remarkable capabilities of robots in interacting with objects and leveraging sensory signals for an array of tasks, encompassing object categorization [14], [15], [27]–[29], material recognition [16], and intricate manipulation actions like packing and pouring [17]. Most successful prior work relies on handcrafted auditory, haptic, and visual features [14], or specialized architectures for processing raw multi-sensory data to predict object categories [15]. Recently, Li *et al.* [17], introduced a self-attention model to fuse information from visual, auditory, and tactile sensors, significantly enhancing the robot's capability to tackle complex manipulation tasks. Our research introduces a versatile framework for learning unified multi-sensory representations *from raw sensory data* acquired during robot-object interactions, offering adaptability across diverse downstream tasks. The generality of our network architecture has been is demonstrated across various applications with strong performance, and the inclusion of self-attention mechanisms further bolsters its performance.

Unified Multi-Sensory Representations with CLIP. Recent advances have revealed the potential of contrastive objectives to yield generalized representations for both text and images [18], [19]. Contrastive Language-Image Pre-training (CLIP) [18] has delivered state-of-the-art representations that excel in diverse tasks, including zero-shot image classification, image retrieval via text, and guiding generative models [30]. While CLIP's knowledge has been distilled for audio

[21], our MOSAIC approach is the first to ground sensory data obtained through robotic object exploration. MOSAIC accomplishes this by distilling knowledge from the extensive pre-trained CLIP text model. To test our learned unified representations, we rely on a dataset where a robot engages with 100 objects, executing 10 exploratory behaviors while recording multiple sensory signals. The robot tackles two tasks reliant on perceiving object properties: object categorization and the fetch object task. The results highlight the efficiency of our unified representations, clearly demonstrated in competitive performance in category recognition only by using a simple linear probe setup and in fetch object task using a zero-shot transfer approach.

III. LEARNING METHODOLOGY

Notation and Problem Formulation. Let a robot perform a set of exploratory behaviors $\mathcal{B}$ (e.g., *grasp*, *pick*) on a set of household objects $\mathcal{O}$ (e.g., *bottle*, *cup*), while recording a set of sensory modalities, $m = \{x^v, x^a, x^h\}$, which correspond to *vision*, *audio*, *haptics*, respectively. The robot performs each behavior n times on each object. During the ith exploratory trial, the robot collects sensory data m_i containing:

$$\mathrm{x}_i^v \in \mathbb{R}^{w \times h \times 3 \times t_i^v}, \mathrm{x}_i^a \in \mathbb{R}^{f \times t_i^a}, \mathrm{x}_i^h \in \mathbb{R}^{d \times t_i^h} \qquad (1)$$

where w and h are the width and height of each image, f is the number of frequency bins in the sound spectrogram, d is the number of robot joint-torque sensors, and t_i^v, t_i^a, and t_i^h are the number of time frames (e.g., number of images) produced during interaction for vision, audio, and haptics, respectively. Additionally, the robot has access to textual descriptions of each interaction, x_i^s, provided by human experts, complementing the sensory data.

Our primary objective is to learn a unified multimodal representation derived from the robot's observations across all modalities during an exploratory trial. To be more precise, we aim to learn the function $F_{m \rightarrow \mathcal{Z}} : x_i^v, x_i^a, x_i^h \rightarrow z_i$, where $z_i \in \mathbb{R}^{D_{\mathcal{Z}}}$ represents the unified multimodal embedding of dimension $D_{\mathcal{Z}}$. This unified representation is intended to encompass diverse object properties encountered during interactions, making it applicable to various downstream tasks that require understanding these object properties. By achieving this unified representation, the robot can rapidly adapt to different tasks by learning linear models or per-

Algorithm 1: `Training MOSAIC Framework`

V, A, H, S: Minibatch of aligned data (vision, audio, haptic, text)
n: Size of minibatch
$MOSAIC_\theta$: Learnable parameters of MOSAIC framework
// Extract feature vector for each modality

1 V_f = vision_encoder(V)   // Vision Transformer
2 A_f = audio_encoder(A)   // Wav2CLIP model
3 H_f = haptic_encoder(H)   // ResNet18 model
4 S_f = text_encoder(S)   // Text Transformer
   // Compute unified representation
5 U_f = concatenation(V_f, A_f, H_f)
6 U_f = multihead_attention(U_f)
7 U_f = MLP_encoder(U_f)   // MLP model
   // Scaled pairwise cosine similarities
8 $logits = U_f \cdot S_f^\top$
   // Symmetric loss function
9 $labels$ = range(n)   // returns 1, 2, ..., n
10 $loss_u$ = cross_entropy_loss($labels, logits$)
11 $loss_s$ = cross_entropy_loss($labels, logits^\top$)
12 $loss = (loss_u + loss_s)/2$
13 Update $MOSAIC_\theta$ to minimize $loss$

---

forming zero-shot transfers, thereby circumventing the need to train complex models dedicated to individual tasks.

Learning Unified Multimodal Object Properties. Our approach, MOSAIC (**M**ultimodal **O**bject property learning with **S**elf-**A**ttention and **I**nteractive **C**omprehension), involves a two-stage process, illustrated in Fig. 1. Initially, we aim to distill unified object property representations from diverse sensory modalities, guided by text embeddings from a pre-trained text encoder. Subsequently, we leverage these unified representations to solve downstream tasks that require understanding object properties. In the following sections, we introduce various modules integrated within our framework.

1) Encoders and Feature Extraction: For the `Vision Encoder`, we use the CLIP's Vision Transformer (ViT-B/32) [18], which is jointly trained with a text encoder to maximize the similarity of {image, text} pairs using a contrastive loss. For each interaction's video, the image encoder extracts image embeddings, and these embeddings are then aggregated using adaptive average pooling to generate a feature vector of size $D_{\mathcal{Z}}$. For the `Audio Encoder`, we leverage the Wav2CLIP model [21], which is trained to project audio data into the shared vision-language embedding space of CLIP; this approach enables the extraction of audio embeddings of size $D_{\mathcal{Z}}$. For the `Haptics Encoder`, we use a ResNet-18 [31] model, pre-trained on the ImageNet dataset, as the foundation. The input channels of the first convolutional layer are modified to one channel, and the output of the last fully-connected layer is adapted to match the desired embedding size of $D_{\mathcal{Z}}$; a sample haptic image is shown in Fig. 1. `Text Encoder`: For each exploratory trial, a corresponding natural language description is available. Leveraging CLIP's text encoder (ViT-B/32) [18], we extract embeddings of size $D_{\mathcal{Z}}$ from these text descriptions.

2) Multimodal Fusion: We employ a self-attention mechanism to integrate the feature sets from the three modalities. Beginning with the concatenation of feature vectors from each modality, we apply a two-step process: first, conventional multi-head self-attention [32] is applied to the concatenated features; subsequently, the resulting output is directed through a Multi-Layer Perceptron (MLP) to yield

Fig. 2: (A) 100 objects, grouped in 20 object categories. (B) The interactive behaviors that the robot performed on the objects.

the unified multi-sensory feature of size $D_{\mathcal{Z}}$.

3) Training: During training, we maintain the vision, audio, and text encoders in their frozen states since they were already tuned to project into a shared embedding space. We train the haptic, self-attention, and MLP networks. Our primary aim is to create unified multimodal representations within the same embedding space as CLIP's text embeddings [18]. To accomplish this, we employ a distillation method guided by CLIP's text embeddings. We follow the approach outlined in the original CLIP paper, using a contrastive loss mechanism. This involves employing positive examples from different modalities within the same data sample while considering negative examples from the remaining batch. The fundamental implementation of this training process is shown in Algorithm 1. This strategy is predicated on the concept that natural language offers a versatile grounding basis [33], facilitating the creation of generalized representations with effective transferability across diverse downstream tasks.

IV. EXPERIMENTAL DESIGN

Sensory Dataset. We used the publicly accessible dataset collected by Sinapov *et al.* [14]. In this experiment, a humanoid robot (depicted in Fig. 1) explored 100 household objects from 20 different categories (shown in Fig. 2A), using 10 exploratory behaviors. These behaviors included *Look*, *Press*, *Grasp*, *Hold*, *Lift*, *Drop*, *Poke*, *Push*, *Shake*, and *Tap* (shown in Fig. 2B). *Look* is a non-interactive behavior, only capturing visual data. For every interactive behavior, the robot collected sensory data including visual, audio, and haptic, acquired through three sensors: (1) A Logitech webcam capturing 320 x 240 RGB images at 10 frames per second; (2) An Audio-Technica U853AW cardioid microphone capturing audio sampled at 44.1 KHz; (3) Joint-torque sensors capturing torques from all 7 joints at 500 Hz. The robot repeated each behavior 5 times for each of the 100 objects, resulting in a total of 5,000 interactions (10 behaviors x 5 trials x 100 objects).

Text Dataset. The objects in our dataset were annotated with properties, shown in Table I, each with corresponding values. While not all properties were applicable to every object (e.g., the *baseball* object lacked a weight property), we leveraged these properties to generate text descriptions for each interaction. To ensure diversity, we randomly selected a subset of properties for each object and used them in the descriptions. For each object's text description, we ensured

that it included at least one property, and the maximum number of properties included was determined by the number of properties with values for that object. Moreover, we included the behavior's name being executed (e.g., *tap*), the object's category (e.g., *ball*), and the category of different object properties (e.g., *material*), all chosen randomly. Further variety was introduced by selecting synonyms for words within the description from a curated set of synonyms corresponding to the dataset's labels. We generated 100 unique text descriptions using this random selection process for each combination of object and behavior. For instance, an example text description might read: *"Performing tap action on a ball with properties: round, yellow, small, soft, toy"*.

Data Pre-processing. To ensure synchronization and consistency across all sensory modalities for each behavior $b \in \mathcal{B}$, we calculated the behavior's duration by dividing the average number of images recorded during behavior b by 10 (camera's frame rate). With the duration of each behavior now fixed, we compute the average number of time frames for each modality by multiplying this duration and the frame rate specific to that modality. These calculated averages were used for interpolation, ensuring uniform time frames for each modality during the interaction recording. For images and text, we employed the pre-processing provided by CLIP [18]. Audio data was transformed from raw waveforms (1D) to spectrograms (2D) using the audio preprocessor from Wav2CLIP [21]. For haptic signals, we applied dimensionality reduction by interpolating the original 500Hz sampling rate down to 50Hz, drawing inspiration from a similar technique used in a prior study [15] conducted with the same dataset we used in our experiments.

Model Implementation. We standardized the size of the embeddings at $D_{\mathcal{Z}} = 512$. Our framework was implemented in PyTorch [34], which includes the multi-sensory self-attention model and MLP encoder.

Validation Procedure. Each of the 20 object categories consists of 5 unique objects. To train our framework, we selected 4 objects from each category for the training set while reserving one object for testing, resulting in a training set with 80 objects and a testing set with 20 objects. We employed a 5-fold object-based cross-validation strategy to ensure that each object appeared four times in the training set and once in the test set. Given that the robot interacted with each object 5 times, our training set contained 400 examples (80 objects × 5 trials); the test set comprised 100 examples (20 objects × 5 trials) for each exploratory behavior. During training, we randomly select a text description for the given object and behavior from the corresponding pool of 100 text descriptions, ensuring variability in the training process.

Evaluation Tasks. After training our framework, we extracted the unified representations by freezing learned weights for all downstream tasks. We evaluated the acquired representations through two distinct tasks. The following subsections elaborate on these tasks, outlining our approach to tackling them with unified representations and discussing our performance metrics. Additionally, we discuss the baseline methods we employed for comparison with our method.

Algorithm 2: `Fetch_object(c, O, B, θ)`

$MOSAIC_\theta$: Learned parameters in Algorithm 1
1 t_c = text_encoder(c):   // Command to fetch target
2 **for** $o \in O$: Set of objects (target and distractor(s)) **do**
3     $similarity = 0$
4   **for** $b \in B$: Set of Behaviors **do**
5        $sensory\_data$ = perform_behavior(o, b)
6   u_b = get_unified_repr($sensory\_data$, $MOSAIC_\theta$)
7        $similarity$ += cosine_similarity(t_c, u_b)
8   **end**
9   Save $similarity$ for o
10 **end**
11 **return** Target Object o with highest $similarity$

TABLE I: Property categories and associated descriptive words.

Properties	Values
Color	brown, blue, pink, red, white, orange, yellow, green, purple, multicolored
Deform.	deformable, rigid, brittle
Hardness	soft, squishy, hard
Material	plastic, wicker, aluminum, foam, metal, rubber, paper, styrofoam, wood
State	closed, full, empty, open
Reflection	shiny, dull
Shape	cylindrical, wide, rectangular, block, box, cone, round
Size	small, short, big, large, tall
Transp.	transparent, opaque, translucent, see-through
Usage	container, toy
Weight	light, heavy

1) Object Category Recognition: In this task, the robot interacts with a given object to identify its category from a set of 20 categories. We use a standard multi-class linear classifier for supervised classification. Specifically, we use a Multi-Layer Perceptron (MLP) architecture that takes the unified representation as input, passes it through a hidden layer and a ReLU activation function, and produces 20 logits for 20 categories. We train this classifier using the cross-entropy loss function for 50 epochs, using the Adam optimizer [35] with a learning rate of 10^{-4}. The trained classifier is then used to recognize the category of test objects, and we compute accuracy as a performance metric, defined as $A = \frac{\text{correct predictions}}{\text{total predictions}}$ (%). We report the mean accuracy over 5 cross-validation folds, as mentioned earlier.

2) Fetch Object: In this task, the robot receives a natural language instruction to fetch an object, specifying its properties (e.g., *"fetch an object that is cylindrical and short"*). The robot is then presented with a group of objects, among which one matches the specified properties (i.e., target object), while the remaining distractor object(s) differ from the target object in at least one property. To illustrate, if the robot is instructed to fetch an object that is both *cylindrical* and *short*, the distractor objects might be *cylindrical* or *short*, but not both. The robot's objective is to interact with these presented objects and correctly identify one with the requested properties. This task presents a challenge as the robot needs to detect the target object's properties given in natural language and distinguish it from the distractors by interaction. We evaluate the robot's performance on the fetch task across different levels of complexity. In this task, we refer to the given instruction as a "command" and the objects presented to the robot are carefully chosen from the previously mentioned test set, ensuring that they are entirely new to the robot. In difficulty *Level 1*, the command specifies the category name of the target object (e.g., *"fetch a ball"*); a distractor object is chosen from a different category. In *Level 2*, the command describes a specific property of

the target object (e.g., *"bring an object that is hard"*). A distractor object is selected with a different property. In the *Level 3* scenario, the command includes two distinct properties of the target object (e.g., *"bring an object that is small and hard"*). The distractor object, on the other hand, possesses different properties. For *Level 4*, like Level 3, the command includes two target object properties. However, this time, two distractor objects are introduced, each with differing properties. *Level 5* represents a variation of Level 2, where the commands only contain a property from a specific category, as illustrated in Table I. For instance, in the *"Material"* category, the command might read, *"get an object that is plastic."* Level 5 was introduced to assess the robot's performance across various property categories. For each level, we created 20 commands for target objects and carefully selected corresponding distractor objects for each of the 5 previously explained folds. For each object (target and distractor(s)), we calculated its selection percentage, defined as $S = \frac{\text{number of times the object is selected}}{\text{total number of commands}}$ (%). Our results are reported as the mean selection percentage across the 5 folds.

We employ the approach outlined in Algorithm 2 for this task. Initially, we convert the natural language instruction into a text embedding, denoted as t_c, using CLIP's text encoder (step 1). Subsequently, the robot interacts with the presented objects, including the target object and distractors, using various available behaviors while simultaneously recording sensory signals (step 5). To simulate this step, we randomly select a trial from our dataset among 5 trials of each object. Leveraging our trained framework, we generate unified representations, denoted as u_b, by processing the sensory inputs for each behavior (step 6). Next, we calculate the cosine similarity between the command embedding (t_c) and the unified representation (u_b) for each behavior, maintaining a cumulative similarity score (step 7). Finally, once all behaviors are considered, the object with the highest cumulative similarity score is identified as the target object, concluding the task (step 11).

Baselines, Ablations, and Comparisons. We ablate our full framework (MOSAIC), featuring the multi-sensory self-attention module, against a framework that omits this component (MOSAIC-*w/o-SA*). We conduct these evaluations under two conditions: a non-interactive condition, where the robot solely performs the *Look* behavior, and an interactive condition, where the robot engages in all 9 aforementioned interactive behaviors. Notably, in the *Look* behavior, only visual embeddings are employed as the unified representations after passing through the self-attention layer in MOSAIC, while, in MOSAIC-*w/o-SA*, the *Look* behavior utilizes only CLIP's vision encoder. Conversely, for interactive behaviors, all three modalities (i.e., visual, auditory, and haptic) are used to create unified representations. For the object category recognition task, we report recognition accuracy separately for the *Look* behavior, each of the 9 interactive behaviors individually, and the combination of all 9 interactive behaviors. The combined accuracy is calculated through a weighted combination of each behavior's performance on the training data. We also compare our recognition accuracy
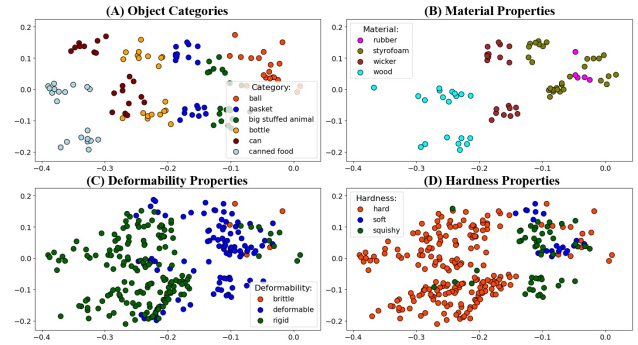

Fig. 3: 2D unified representations derived from autoencoder trained on *Push* behavior's data: (A) Object categories, (B) Material, (C) Deformability, and (D) Hardness properties.

with two baseline methods: Sinapov *et al.* [14], who trained a Support Vector Machine (SVM) classifier using handcrafted auditory, haptic features, and visual features, and Tatiya et al. [15], who applied a deep learning approach to raw multi-sensory data for object category classification. For the fetch object task (see Algorithm 2), the set B contains only the *Look* behavior for the non-interactive condition and all 9 interactive behaviors for the interactive condition.

V. RESULTS

An Illustrative Example. Let's consider a scenario where the robot performs the *Push* behavior on 80 objects (4 objects x 20 categories), recording visual, acoustic, and haptic data. With each object undergoing 5 trials, this yields a dataset of 400 examples (80 objects × 5 trials). Using our MOSAIC framework, we use this data to learn unified representations. For visualization, we subjected these representations to dimensionality reduction using a linear autoencoder, resulting in a concise 2-dimensional latent space (Fig. 3). This visualization encapsulates four object properties: object categories, material, deformability, and hardness. Distinct colors are used to differentiate objects based on different values of these properties. To maintain clarity, we selectively plot only specific categories or objects with particular properties. Due to the absence of certain properties in some objects, the observed inconsistencies in Fig. 3 for different properties arise, as not every object possesses all the properties.

These visualizations unveil meaningful insights. Objects within the same category or material composition form tight clusters in the 2D space, showing the efficiency of our unified representations in capturing object semantics and material characteristics. The deformability properties plot demonstrates a separation between *rigid* and *deformable* objects, with *brittle* ones inclining towards *deformable*. Similarly, in the hardness properties plot, *hard* objects cluster on one side, while *soft* and *squishy* objects gravitate towards the opposite side. Essentially, our unified representations effectively encode objects with similar properties, as evidenced by distinct clusters of similar objects, even when these objects belong to different categories or material groups across various property categories. This illustrates MOSAIC's capacity to capture nuanced object attributes and relationships, a pivotal aspect of its performance across diverse tasks.

TABLE II: Category recognition accuracy (%) for each behavior.

Behavior	Sinapov *et al.*[14]	Tatiya *et al.*[15]	MOSAIC-*w/o-SA*	MOSAIC (ours)
Look	67.7	—	86.4 ± 1.2	**87.4 ± 2.0**
Grasp	65.2	71.4	72.2 ± 6.7	**74.0 ± 5.8**
Hold	67.0	**76.8**	68.0 ± 5.3	69.6 ± 5.2
Lift	**79.0**	77.8	72.8 ± 4.2	77.8 ± 5.7
Drop	71.0	**78.0**	73.2 ± 3.8	77.2 ± 5.9
Poke	85.4	73.8	81.6 ± 2.2	**86.4 ± 1.0**
Push	88.8	67.4	85.6 ± 3.5	**89.4 ± 4.4**
Shake	76.8	83.6	81.2 ± 6.2	**84.0 ± 5.6**
Tap	82.4	81.6	81.2 ± 5.7	**84.4 ± 1.8**
Press	77.4	58.8	71.6 ± 8.7	**77.8 ± 6.4**
All behaviors	—	—	95.2 ± 3.6	**95.6 ± 3.9**

TABLE III: MOSAIC's target object selection (%) in various levels of the fetch object task, with and without Self-Attention.

	Look (non-interactive)		Interactive	
	-*w/o-SA*	*MOSAIC*	-*w/o-SA*	*MOSAIC*
LEVEL 1	74	82	97	**99**
LEVEL 2	61	65	**84**	81
LEVEL 3	60	74	**86**	83
LEVEL 4	54	70	72	**77**
LEVEL 5				
DEFORMATION	45	48	71	**74**
SHAPE	85	80	**97**	95
SIZE	62	74	72	**75**
TRASPARENCY	62	62	51	**63**
WEIGHT	52	63	**85**	**85**

Object Category Recognition Results. Object category recognition results are presented in Table II. Note that the *Look* behavior only relies on visual modality, and the "All behaviors" row at the bottom refers to all 9 interactive behaviors combined. Our approach, using unified representations, exhibits a remarkable level of competitiveness compared to state-of-the-art results for this dataset, demonstrating higher recognition accuracy in seven out of ten behaviors. For the remaining three behaviors, we achieved comparable accuracy. We achieved this level of performance using a straightforward linear model on top of the unified representations, a contrast to previous methods. Notably, the prior work [15] employed a specialized neural network architecture tailored specifically for this task, while [14] relied on handcrafted features. Furthermore, our results consistently indicate that our full framework, including self-attention, outperforms the counterpart without self-attention. This underscores the utility of the multi-sensory unified representation and the effectiveness of the self-attention mechanism in enhancing the robot's adaptability to diverse tasks.

Fetch Object Results. The fetch object task, whose results are summarized in Table III, comprises five distinct levels designed to assess the robot's ability to execute instructions. In *L-1* (Level-1), the command specified the object category name. Our complete MOSAIC framework excelled in interactive behavior conditions, achieving an impressive target object selection rate of 99.0%, outperforming all baseline models. *L-2 to L-5*: These levels introduced object properties into the command instead of specifying the object category name. Generally, the interactive behaviors condition outperformed the non-interactive one, with our full MOSAIC model excelling in most cases. Interestingly, providing more object properties in the command led to better performance, exemplified by a higher target object selection rate in L-3 compared to L-2, across all conditions, except for *"Look"* without self-attention. This suggests that learning unified representations with self-attention prioritizes the most relevant object properties. *L-4* presented greater challenges due to the inclusion of two distractor objects resembling the target object. Nevertheless, our complete MOSAIC framework with self-attention consistently outperformed all baselines.

To evaluate the robot's ability to fetch objects based on specific property categories, we considered L-5, where the command included only descriptive words related to specific property categories. For simplicity, we discuss five property categories. *Deformability and Weight*: In scenarios involving non-visual properties like deformability and weight, the interactive behaviors condition significantly outperformed the non-interactive one. This aligns with intuition, as visual observation alone may not suffice to disambiguate these properties. *Transparency and Size*: For visual properties like transparency and size, the interactive behaviors condition performed comparably to the non-interactive condition, suggesting that interaction with objects may not yield significantly more information in these scenarios. *Shape*: Intriguingly, for the shape property category, the interactive behaviors condition significantly outperformed the non-interactive one. This implies that interacting with objects enables the robot to observe them from various angles, enhancing its ability to predict object shape compared to merely observing from a top angle. In summary, our full MOSAIC framework demonstrated robust performance in the fetch object task, relying solely on unified representations without additional learning. These results underscore the adaptability and applicability of unified representations across diverse tasks, including those involving natural language instructions.

VI. CONCLUSION AND FUTURE WORK

We introduced the MOSAIC framework to enable robots to generate versatile, multimodal representations through interactive object perception and to leverage these unified representations across various downstream robot learning tasks. Through extensive performance evaluation, we have showcased the effectiveness of these unified representations in tasks such as category recognition, using a simple linear probe setup, and the fetch object task under zero-shot conditions. Moving forward, there are several exciting directions for future research. Firstly, we plan to consider the *transfer* of unified representations across different robot morphologies, enabling a broader range of robots to benefit from this technology. Furthermore, we envision settings where interactive behaviors are learned and composed, alongside the tasks we considered in this paper, thereby further increasing the efficacy of object exploration. These future endeavors hold the potential to further enhance the utility of unified representations in robotics and expand their applications across a multitude of scenarios and environments. One limitation in our current study is that, for the fetch object task, we evaluated using a zero-shot transfer condition rather than a learning-based approach to find the target object. For future work, it would be important to explore learning-based policies for solving the fetch object task, potentially increasing the versatility and adaptability of our framework.

REFERENCES

[1] T. Thesen, J. F. Vibell, G. A. Calvert, and R. A. Österbauer, "Neuroimaging of multisensory processing in vision, audition, touch, and olfaction," *Cognitive Processing*, vol. 5, pp. 84–93, 2004.

[2] D. Alais, F. Newell, and P. Mamassian, "Multisensory processing in review: from physiology to behaviour," *Seeing and perceiving*, vol. 23, no. 1, pp. 3–38, 2010.

[3] D. A. Bulkin and J. M. Groh, "Seeing sounds: visual and auditory interactions in the brain," *Current opinion in neurobiology*, vol. 16, no. 4, pp. 415–419, 2006.

[4] S.-C. Kim and S. Ryu, "Robotic kinesthesia: Estimating object geometry and material with robot's haptic senses," *IEEE Transactions on Haptics*, 2023.

[5] X. Zhang, S. Amiri, J. Sinapov, J. Thomason, P. Stone, and S. Zhang, "Multimodal embodied attribute learning by robots for object-centric action policies," *Autonomous Robots*, pp. 1–24, 2023.

[6] Q. Li, O. Kroemer, Z. Su, F. F. Veiga, M. Kaboli, and H. J. Ritter, "A review of tactile information: Perception and action through touch," *IEEE Transactions on Robotics*, vol. 36, no. 6, pp. 1619–1634, 2020.

[7] G. Tatiya, J. Francis, and J. Sinapov, "Transferring implicit knowledge of non-visual object properties across heterogeneous robot morphologies," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 11 315–11 321.

[8] J. K. Bizley, G. P. Jones, and S. M. Town, "Where are multisensory signals combined for perceptual decision-making?" *Current opinion in neurobiology*, vol. 40, pp. 31–37, 2016.

[9] C. V. Parise, C. Spence, and M. O. Ernst, "When correlation implies causation in multisensory integration," *Current Biology*, vol. 22, no. 1, pp. 46–49, 2012.

[10] J. Francis, N. Kitamura, F. Labelle, X. Lu, I. Navarro, and J. Oh, "Core challenges in embodied vision-language planning," *Journal of Artificial Intelligence Research*, vol. 74, pp. 459–515, 2022.

[11] Y. Hu, Q. Xie, V. Jain, J. Francis, J. Patrikar, N. Keetha, S. Kim, Y. Xie, T. Zhang, Z. Zhao *et al.*, "Toward general-purpose robots via foundation models: A survey and meta-analysis," *arXiv preprint arXiv:2312.08782*, 2023.

[12] C. Chen, K. Z. Zhang, Z. Chu, and M. Lee, "Augmented reality in the metaverse market: the role of multimodal sensory interaction," *Internet Research*, 2023.

[13] S. Cai, K. Zhu, Y. Ban, and T. Narumi, "Visual-tactile cross-modal data generation using residue-fusion gan with feature-matching and perceptual losses," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 7525–7532, 2021.

[14] J. Sinapov, C. Schenck, K. Staley, V. Sukhoy, and A. Stoytchev, "Grounding semantic categories in behavioral interactions: Experiments with 100 objects," *Robotics and Autonomous Systems*, vol. 62, no. 5, pp. 632–645, may 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S092188901200190X

[15] G. Tatiya and J. Sinapov, "Deep multi-sensory object category recognition using interactive behavioral exploration," in *International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, May 20-24, 2019*. IEEE, 2019, pp. 7872–7878. [Online]. Available: https://doi.org/10.1109/ICRA.2019.8794095

[16] P. Xiong, J. Liao, M. Zhou, A. Song, and P. X. Liu, "Deeply supervised subspace learning for cross-modal material perception of known and unknown objects," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 2, pp. 2259–2268, 2022.

[17] H. Li, Y. Zhang, J. Zhu, S. Wang, M. A. Lee, H. Xu, E. H. Adelson, L. Fei-Fei, R. Gao, and J. Wu, "See, hear, and feel: Smart sensory fusion for robotic manipulation," in *Conference on Robot Learning (CoRL)*, vol. 205. Auckland, New Zealand: Proceedings of Machine Learning Research (PMLR), dec 2022, pp. 1368–1378. [Online]. Available: https://proceedings.mlr.press/v205/li23c.html

[18] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning (ICML)*, vol. 139. Proceedings of Machine Learning Research (PMLR), 2021, pp. 8748–8763. [Online]. Available: http://proceedings.mlr.press/v139/radford21a.html

[19] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz, "Contrastive learning of medical visual representations from paired images and text," in *Machine Learning for Healthcare Conference*. PMLR, 2022, pp. 2–25.

[20] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds *et al.*, "Flamingo: a visual language model for few-shot learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 23 716–23 736, 2022.

[21] H.-H. Wu, P. Seetharaman, K. Kumar, and J. P. Bello, "Wav2clip: Learning robust audio representations from clip," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Virtual and Singapore: IEEE, May 2022, pp. 4563–4567. [Online]. Available: https://ieeexplore.ieee.org/document/9747669

[22] A. Borghi, A. Di Ferdinando, and D. Parisi, "The role of perception and action in object categorisation," in *Connectionist models of cognition and perception*. World Scientific, 2002, pp. 40–50.

[23] S. Lacey and K. Sathian, "Visuo-haptic multisensory object recognition, categorization, and representation," *Frontiers in psychology*, vol. 5, p. 730, 2014.

[24] G. A. Calvert and T. Thesen, "Multisensory integration: methodological approaches and emerging principles in the human brain," *Journal of Physiology-Paris*, vol. 98, no. 1-3, pp. 191–205, 2004.

[25] T. Koelewijn, A. Bronkhorst, and J. Theeuwes, "Attention and the multiple stages of multisensory integration: A review of audiovisual studies," *Acta psychologica*, vol. 134, no. 3, pp. 372–384, 2010.

[26] D. Talsma, "Predictive coding and multisensory integration: an attentional account of the multisensory mind," *Frontiers in Integrative Neuroscience*, vol. 9, p. 19, 2015.

[27] G. Tatiya, R. Hosseini, M. C. Hughes, and J. Sinapov, "Sensorimotor cross-behavior knowledge transfer for grounded category recognition," in *Joint IEEE 9th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob), Oslo, Norway, August 19-22, 2019*. IEEE, 2019, pp. 1–6. [Online]. Available: https://doi.org/10.1109/DEVLRN.2019.8850715

[28] G. Tatiya, Y. Shukla, M. Edegware, and J. Sinapov, "Haptic knowledge transfer between heterogeneous robots using kernel manifold alignment," in *IEEE/RSJ International Conference on Intelligent Robots and Systems IROS, Virtual Event, Las Vegas, NV, USA, October 25-29, 2020*. IEEE, 2020, pp. 5358–5363. [Online]. Available: https://doi.org/10.1109/IROS45743.2020.9340770

[29] G. Tatiya, R. Hosseini, M. Hughes, and J. Sinapov, "A framework for sensorimotor cross-perception and cross-behavior knowledge transfer for object categorization," *Frontiers in Robotics and AI*, vol. 7, p. 137, 2020. [Online]. Available: https://www.frontiersin.org/article/10.3389/frobt.2020.522141

[30] R. Gal, O. Patashnik, H. Maron, A. H. Bermano, G. Chechik, and D. Cohen-Or, "Stylegan-nada: Clip-guided domain adaptation of image generators," *ACM Transactions on Graphics (TOG)*, vol. 41, no. 4, pp. 1–13, 2022.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[33] Y. Bisk, A. Holtzman, J. Thomason, J. Andreas, Y. Bengio, J. Chai, M. Lapata, A. Lazaridou, J. May, A. Nisnevich *et al.*, "Experience grounds language," *arXiv preprint arXiv:2004.10151*, 2020.

[34] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," *CoRR*, vol. abs/1912.01703, 2019. [Online]. Available: http://arxiv.org/abs/1912.01703

[35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, San Diego, CA, USA, may 2015. [Online]. Available: https://arxiv.org/abs/1412.6980